Smart Home Personal Assistants: A Security and Privacy Review

A Preprint

Jide S. Edu
Department of Informatics
King's College London
Faculty of Natural and Mathematical Science
Strand campus
jide.edu@kcl.ac.uk

Jose M. Such
Department of Informatics
King's College London
Faculty of Natural and Mathematical Science
Strand campus
jose.such@kcl.ac.uk

Guillermo Suarez-Tangil
Department of Informatics
King's College London
Faculty of Natural and Mathematical Science
Strand campus
guillermo.suarez-tangil@kcl.ac.uk

December 21, 2020

Abstract

Smart Home Personal Assistants (SPA) are an emerging innovation that is changing the means by which home users interact with technology. However, there are a number of elements that expose these systems to various risks: i) the open nature of the voice channel they use, ii) the complexity of their architecture, iii) the AI features they rely on, and iv) their use of a wide range of underlying technologies. This paper presents an in-depth review of the security and privacy issues in SPA, categorizing the most important attack vectors and their countermeasures. Based on this, we discuss open research challenges that can help steer the community to tackle and address current security and privacy issues in SPA. One of our key findings is that even though the attack surface of SPA is conspicuously broad and there has been a significant amount of recent research efforts in this area, research has so far focused on a small part of the attack surface, particularly on issues related to the interaction between the user and the SPA devices. We also point out that further research is needed to tackle issues related to authorization, speech recognition or profiling, to name a few. To the best of our knowledge, this is the first article to conduct such a comprehensive review and characterization of the security and privacy issues and countermeasures of SPA.

Keywords Smart Home Personal Assistants · Security and Privacy · Voice Assistants · Smart Home · Amazon Echo/Alexa · Google Home/Assistant · Apple Home Pod/Siri · Microsoft Home Speaker/Cortana

1 Introduction

Human-computer interaction (HCI) has traditionally been conducted in the form of different types of peripheral devices such as the keyboard, mouse and most recently tactile screens. This has been so because computing devices were not able to decode the meaning of our word, let alone understand our intent. Over the last few years, however, the paradigm has shifted, as we witnessed the rapid development of voice technology in many computing applications. Since voice is one of the most effective and expressive communication tools, voice technology is changing the way in which users interact with devices and the manner they consume services. Currently, one of the most significant innovations that uses voice technology are Smart Home Personal Assistants (SPA). SPA are intelligent assistants that take instructions from users, process them and perform the corresponding tasks. They offer hands-free and eye-free operations, thereby allowing users to perform diverse activities using voice commands while concentrating elsewhere on other tasks. Besides offering users the benefit of a quick interaction — humans speak faster than they type [1], using voice for human-computer interaction can be considered more natural [2] when compared to other interfaces like keyboard, and mouse. Not to mention the stronger social presence offered to users when they hear synthesized speeches very much like their own as responses from this technology [3].

SPA are rapidly becoming everyday features in homes and are increasingly becoming integrated with other smart devices [4]. It is believed that 10% of the world consumers own SPA devices [5]. According to a recent survey by Voicebot, over 50 million Alexa Echo devices have been sold to date in the US alone [6]. There are a number of features that contribute to the popularity of SPA. SPA are quite different from early voice activated technologies that could only work with small inbuilt commands and responses. Instead, SPA use Internet services and benefits from recent advances in Natural Language Processing (NLP), which allow them to handle a wide range of commands and questions. They enable a playful interaction, making their use more engaging [7]. They are assigned a name and a gender, which encourages users to personify them and therefore interact with them in a human-like manner [8]. They are used to maintain shopping and to-dos lists, purchase goods and food, ask knowledge questions, play audio-books, play games, stream music, radio and news, set timers, alarms and reminders [9], get recipe ideas, control large appliances [10], send messages, make calls [11] and many more depending on their usage context [12, 13]. With the continuous proliferation and the rapid growth of SPA, we are now approaching an era when SPA will not only be manoeuvring our devices at home but also replacing them in many cases. For instance, many SPA are now able to make phone calls, which positions them as a communicating device, and a likely alternative to landlines phones in the future, and some SPA are also equipped with display interface for watching videos/movies and smart home cameras directly in the SPA devices [14].

Given the increasing popularity of these devices, it is paramount to understand the underlying risks behind their use and fathom how to mitigate them. While most of these devices have incorporated some security and privacy mechanisms in their design, there is still a significant number of security and privacy challenges that need to be addressed. This is all the more important because SPA carry out distinct roles and perform various functions in single and multi-user environments, particularly in an intimate domain like homes. Since the users co-locate with this technology, it also has an impact on the changes in their neighbouring environment [15]. In fact, there have already been reported security and privacy incidents in the media involving SPA, such as the case of an Amazon Alexa recording an intimate conversation and sending it to an arbitrary contact [12]. There is also research evidence that users are concerned about the security and privacy of these devices [16, 17] and try to implement workaround non-technical countermeasures in the absence of better technical security and privacy controls (like turning off the SPA when they are not using it to avoid the device listening to what they say), with a study by Lau et al. [18] finding that these concerns were a major deterring factor for new users.

Despite the fast growing research on the security and privacy issues of SPA, the literature lacks a detailed characterization of these issues. This paper offers the first comprehensive review of existing security and privacy attacks and countermeasures in smart home personal assistants and presents a categorization for them. For this, we first provide an overview and background of the architectural elements of SPA, which is important to understand both potential weaknesses and countermeasures. In addition, and based on our analysis and categorization of risks, attacks and countermeasures, this paper presents a roadmap of future research directions in this area. We found out that while the attack surface of SPA is distinctly broad, the research community has focused only on a small part of it. In particular, recent works have focused largely on issues related to the direct interaction between a user and their SPA. While those problems are indeed very important and further research is needed for effective countermeasures, we also found that research is needed to address other issues related to authorization, speech recognition, profiling, and the technologies integrated with SPA (e.g. the cloud and other smart devices).

The rest of this paper is structured as follows: Section 2 offers an introduction to SPA and their architecture. In Section 3, we describe the different security and privacy issues in the SPA. Known attacks on SPA are discussed in Section 4. Section 5 describes existing countermeasures, and Section 6 provides a summary and some discussions on future research directions. Finally, Section 7 draws the conclusion.

2 Background

Smart Home Personal Assistants (SPA) have a complex architecture (see details in Section 2.1). As a general introduction, and despite the fact that different SPA across different vendors have a few distinctive characteristics, all SPA perform similar functions and share some common features. In particular, SPA's architectures usually include, together with other architectural elements such as cloud-based processing and interaction with other smart devices, the following: i) a *voice-based intelligent personal agent* such as Amazon's Alexa, Google's Assistant, Apple's Siri, and Microsoft's Cortana [19]; and ii) a *smart speaker* such as Amazon's Echo family, Microsoft's home speaker, Google's Home Speaker, and Apple's Home Pod. This review focuses on SPA because their architecture is complex and significantly different from other architectures that use voice-based intelligent personal agents, which make SPA worth studying separately. However, this review also provides insights that might be useful to other domains of application of voice-based personal assistants (e.g., smartphones), particularly those related to speech recognition.

SPA decode users' voice input using Natural Language processing to understand users' intent. Once the intent is identified, it delegates the requests to a set of *Skills*[1] from where it obtains answers and recommendations. Conceptually, skills are similar to mobile apps, which interface with other programs to provide functionality to the user. The entire skills ecosystem provides an environment that offers the user the ability to run more complex functions such as calendar management, shopping, music playback, and other home automation tasks. There are two types of skills, namely: *native skills* and *third-party skills*. The former are skills given by the SPA provider that perform basic functions and leverage providers' strengths in areas such as productivity (Microsoft Cortana), search (Google Assistant) and e-commerce (Amazon Alexa) [20]. The latter are skills built by third-party developers using Skill Kits [21, 22], which are development frameworks with a set of APIs offered by the SPA provider to perform basic operations. There are currently thousands of SPA skills hosted online, although the numbers keep growing daily. For example, Amazon's skill market currently has over 70,000 Alexa skills worldwide [23] and Google Assistant skill market has over 2,000 skills [24]. These skills are classified into different categories such as Home Control Skills, Business and Finance Skills, Health and Fitness Skills, Games and Trivia Skills, News Skills, Social Skills, Sports Skills, Utilities Skills, etc. As further support to the skills, SPA often have the ability to learn information about users' preferences such as individual language usages, words, searches, and services using Machine Learning (ML) techniques [25] to make them smarter over time.

2.1 Smart Home Personal Assistants Architecture

SPA are Internet-based systems with a regular iteration of updates. One benefit of this is that its capabilities are wide-ranging and dynamic — they will evolve along with the proliferation of new Internet services. Figure 1 shows the key components in the SPA system architecture. Each component is a potential attack point for an adversary. How some of them are exploited is discussed in Section 4. Point 1 represents the point of interaction between the users and the SPA devices. SPA devices such as Amazon Echo are equipped with powerful microphones and the device itself consists of a voice interpreter that records users' utterances. To make use of the SPA, the voice interpreter needs to be activated. Many of the voice interpreters are often pre-activated and run at the background. After the voice interpreter is activated, it then waits for the wake-up word to be triggered. Once it receives the wake-up keyword, it puts the SPA into recording mode. In recording mode, any user utterances are processed and sent through the home router (Point 2) to the SPA cloud (Point 3) for further analysis. Only the wake-up command is executed locally, while all other commands are sent to the cloud. Hence, the SPA must always be online.

In the SPA cloud, the captured utterances are decoded using NLP as we detail in Section 2.2 below. It must overcome the issue of background noise, echo, and accent variation in the process of extracting the intent. Once the intent is extracted it is used to determine which skill to invoke. There are two ways to invoke a skill. First, they can be explicitly invoked by using their activation name: for example, where a skill name is "Tutor Head," it can be triggered by saying the words: "talk to Tutor Head." Explicit invocation can be extended to use a deep link connection, as detailed here [26] for Google Assistant. For instance, "talk to Tutor Head to find the next course" where the next course is a predefined action under the "Tutor Head" skill. Second, skills can be implicitly invoked by an intent's query composition without the need to explicitly use their invocation name. In a case where a query does not directly match with a skill, the SPA will either inform the user or match the query to another similar skill when appropriate.

By default, the SPA provider will try to find a native skill to process the request invoked by the user. In this case, the SPA cloud service then sends the intent to its native skill, which processes the request in the cloud of the SPA (Point 5) and sends a response back to the SPA device. When there are no native skills available, the request is sent to a third-party skill (Point 6). These are typically hosted in a remote web service host controlled by the developer of the third-party skill. Once the request is processed, the third-party skill returns the answers to the SPA cloud service, which sometimes asks for more information before the request is finalized. In the case where the intent is meant to control other smart devices, the relevant information is forwarded to their respective cloud service (at Point 7), and from there, the instructions are relayed to the target smart device (at Point 8).

[1] Note that, for ease of exposition, we adopt Amazon's terminology of skills, but these may be called differently in other SPA platforms. For instance, in Google's Assistant and Google Home, skills are called *Actions* instead.

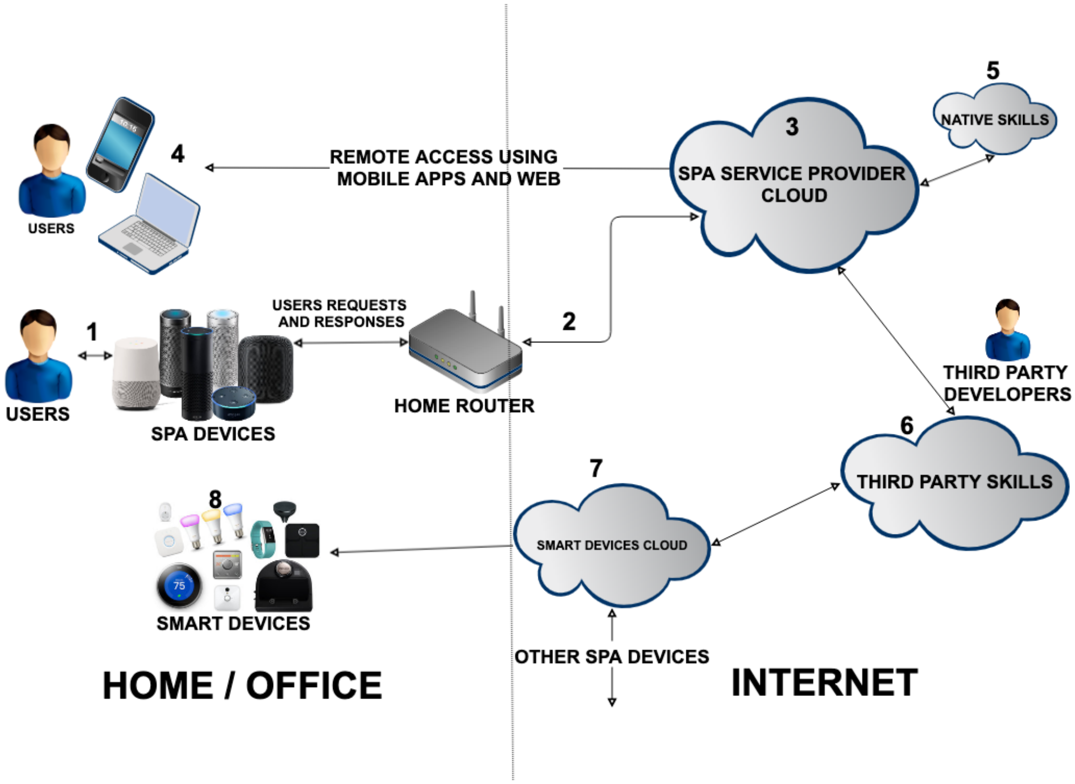

Figure 1: SPA architecture and its key components.

2.2 Natural Language Processing in SPA

SPA benefit from recent advances in Natural Language Processing (NLP), which allow them to handle a wide range of commands and questions. The NLP improvements are attributed to: i) a number of novel advances in ML, ii) a better knowledge of the construction and use of the human language, iii) an increase in the computing power, and iv) the availability of sizable labeled datasets for training speech engines [27]. Processing user speech includes a complex procedure that involves audio sampling, feature extraction and speech recognition to transcribe the requests into text. Since humans speak with idioms and acronyms, it takes an extensive analysis of natural language to get correct outputs. For instance, issuing a command to an SPA asking it to remind you about a meeting at a specific time can be done in several ways. While some parts of this command are more specific than others and can easily be understood, such as the day of the week, other words that support them can be dynamic. This implies that understanding an intention as simple as a meeting reminder might require non-trivial interactions. Figure 2 illustrates the process involved in understanding a user's intent and generating responses.

Intent recognition starts with signal processing, which offers the SPA a number of chances to make sense of the audio by cleaning the signal. The idea is to enhance the target signal, which implies recognizing the surrounding noise to reduce it. That is one of the reasons why most SPA devices are equipped with multiple microphones to roughly ascertain where the signal is coming from so that the device can concentrate on it. Once the original signal is identified, acoustic echo cancellation [28] is then used to subtract the noise from the received signal so that only the vital signal remains. Typically, most speech recognition systems work by converting the sound waves from the user's utterances into digital information [29]. This is further analyzed in order to extract features from user's speech, such as frequency and pitch. Primarily, Automatic Speech Recognition (ASR) comprises of two steps: features extraction and pattern classifiers using ML [30]. There are several feature extraction methods, with Mel frequency cepstral coefficient (MFCC) being one of the most popular, since it is believed to mimic the human auditory system [31]. These features are then fed into an acoustic model trained using ML techniques to match the input audio signal to the correct text. For instance, machine learning models based on Hidden Markov Model (HMM)[32] often compare each part of the waveform against what comes previously and what comes next, and against a dictionary of waveforms to discover what is being said.

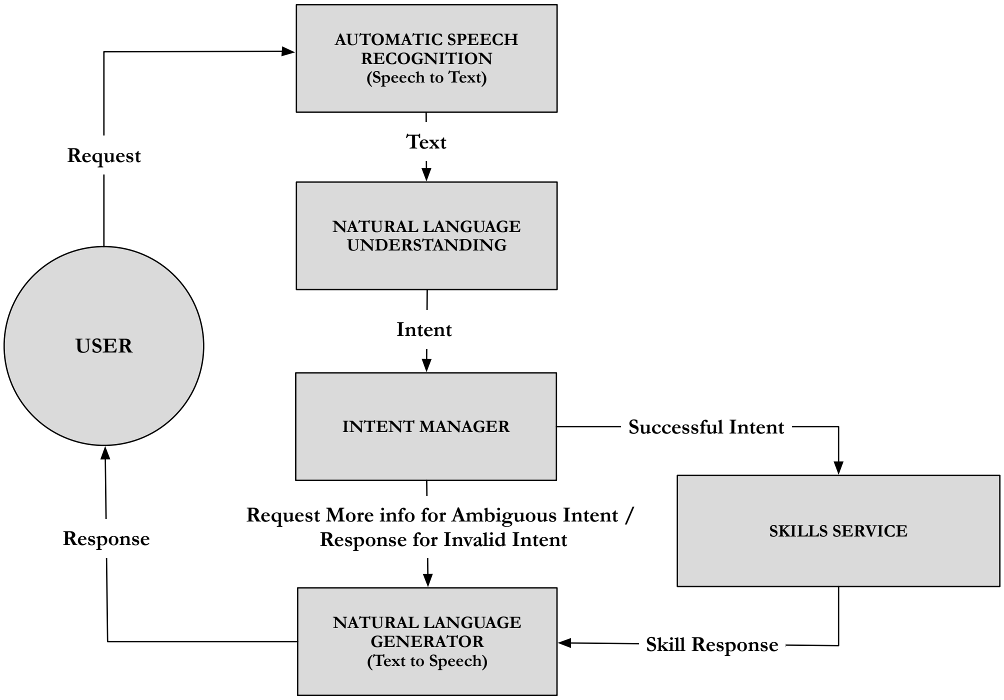
AUTOMATIC SPEECH RECOGNITION (Speech to Text)
Text
NATURAL LANGUAGE UNDERSTANDING
Intent
INTENT MANAGER
Successful Intent
SKILLS SERVICE
Request More info for Ambiguous Intent / Response for Invalid Intent
NATURAL LANGUAGE GENERATOR (Text to Speech)
Skill Response
USER
Request
Response

Figure 2: NLP speech system from *Speech To Text* (ASR) to *Text To Speech*
.

Once the SPA cloud has the text that transcribes what the user has said, it employs Natural Language Understanding (NLU) which is a key component of Natural Language Processing (NLP) to understand what the user intends to do. This is done using discreet to discreet mapping with some instances relying on statistical models or machine learning techniques like deep learning to make an assumption about the likely intent. The more data available to the NLP system from regular usage, the better the prediction of the user's intent. After the NLU extracts the intent, the intent manager then decides whether more information is needed to provide an accurate answer before forwarding the intent to the skill service for processing. After the intent is processed, the generated skill response is send to the Natural Language Generation (NLG) where it is converted into natural language representation. Then, it is communicated back to the user and it is typically (e.g., Amazon Echo) played by a smart speaker.

3 Security and Privacy Issues

In this section, we present a classification of the main security and privacy issues of SPA. We use this classification together with the points and components in the architecture described in Section 2 to later categorize current attacks and countermeasures in Sections 4 and 5.

3.1 Weak Authentication

Here, we discuss the issues related to how SPA verify users and how such process can be exploited by an adversary.

3.1.1 Wake-up Words

By design, SPA authentication is done using wake-up words that are recognized locally in the device. A user has the option to select a wake-up word from a set of predefined options, having one by default. It is therefore very easy for an attacker to infer the wake-up word of the user. In addition to the wake-up word, SPA have no additional ways of authenticating the user. The device will accept any command preceding the wake-up keyword. Hence, it is easy for anyone in proximity to issue commands to the SPA. Authors in [33, 34, 35, 36] have shown how this weak authentication can be used as a proxy to more elaborated attacks.

3.1.2 Always On, Always Listening

As mentioned, the voice command interpreter constantly listens to the user utterances while waits for the wake-up word. Having a device permanently on and always listening poses important security and privacy concerns. Accidentally saying the wake-up word or any other phonetically similar words will put the assistants to record. Consequently, any conversation that follows is uploaded to the Internet. This issue could have an implication on the users' privacy in a situation where private or confidential conversations are accidentally leaked, or where an attacker is able to retrieve sensitive information from these devices. Likewise, it could also affect the device security as an adversary can easily compromise such devices and use them to target other connected smart devices. Recently, due to this feature, a private conversation of a couple was accidentally recorded and sent to a random contact with the Echo device [37]. This example comes to show that the users are not in total control of their voice data.

3.1.3 Synthesized Speech

SPA devices lack protection against machine synthesized speech imitating a legitimate user. For instance, they are vulnerable to inaudible sound reproduced at ultrasonic frequencies [35, 38]. Since the SPA wake-up word can be readily guessed, there is little or no limit to which speech can be supplied to them and by whom, provided it is meaningful and can be matched with an intent. The lack of protection against these inaudible audio signal also offers a covert channel to an adversary. For example, an attacker can embed an inaudible signal into the audio of a TV or radio broadcast to attack multiple targets at once. Just recently, a Burger King TV advert prompted Google Home to read information to the user from Wikipedia about the Whopper hamburger [39].

3.1.4 Weak Payment Authentication

SPA system are increasingly supporting online ordering. Implementing proper security controls challenges usability. For instance, Amazon Alexa users have the option to set a 4-digit PIN code to confirm purchases. At the time of writing, this option is not enabled by default. Even when such an option is turned on, it is vulnerable due to weak lockout[2] implementation [40]. This is because Alexa allows two PIN tries before an ordering process lockout, after which the user has to restart the ordering process from the beginning. However, there is no restriction on how many times a user can try to order after every lockout [40]. Following this, vendors have tried to implement alternative countermeasures against misuse in the ordering process. We next show two cases of this. First, some vendors have prevented changes to the shipping address during ordering. Preventing any change to the shipping address during this process is not enough when dealing with "insiders" (i.e., unauthorized users that have access to the premises where the SPA are installed). The case described in [41] shows how a kid recently made an unauthorized order worth of about $300 using her mother's Amazon account [41]. Second, other vendors have tackled this weak authorization problem by providing prompt notification to the users about orders. This poses a problem to users who do not frequently check their phones or emails, or who may not understand what is happening.

3.2 Weak Authorization

In this part, we evaluate the issues regarding how the SPA manages the level of access to data and the mechanisms users have to control that.

3.2.1 Multi-user Environment

The absence of proper functional role separation prevents users from correctly defining what and how resources should be accessed. It is difficult to specify who has access to which resources and how such access should be granted. By default, in a multi-user environment — which many households are, any user can put the SPA into recording mode and issue out instructions to it. Even though the main user can specify certain access controls for secondary users, the level of granularity is generally coarse and not extensive. For instance, any member of an Amazon household (a feature that allows sharing of contents with family members) can modify the device set-up such as the network connection, sound, and many more without the primary user consent.

3.2.2 Skills

When users invoke a skill, they also take ownership of all the vulnerabilities in its services. Authorizing a vulnerable skill to access confidential information may result in leaking sensitive information to unwanted parties. Unlike apps on smartphones that can be tested for known security vulnerabilities and issues, skills used by SPA are not currently tested.

[2] Lockout is a security mechanism that locks an application for a period of time before a reattempt is allowed.

A user must rely on the SPA provider to ensure that such services are as secure as they need to be. However, even if the SPA provider would provide a vetting process, related works have shown how they could be evaded [36]. For instance, third-party skills can be updated after the vetting process (c.f., Point 6 in Figure 1). Besides, a malicious Alexa-skills developer could leverage skill connections pairing [42] (an important capability that allows the skill to skill integration) to route victims to malicious skills.

3.2.3 Third-Party Access

One important concern is how SPA providers, Skills developers, Developers of integrated Smart home devices, and those that have direct access to any of the points of the SPA architecture secure users from external parties that do not have access to any of these points. Like in every other cloud service, the question stays on how data gathered by those involved in the SPA system is shared with third parties, particularly in terms of what kind of controls and mechanisms can be implemented to provide more control to users. Informed decisions can sometimes be taken when third parties provide privacy policies and terms of use [43]. However, it is currently uncertain what the scope of those terms might imply and how they are enforced. For instance, the Chinese government in collaboration with iFlytek: a voice recognition service provider, have set a national voice biometric database to support its surveillance and social control efforts [43].

3.3 Profiling

Beyond authorization, i.e., deciding who has access to what data, there is also the problem of data inference — traditionally known as information processing [44]. Data inference has a particularly dangerous incarnation in SPA in the form of profiling. Profiling identifies, infers and derives relevant personal information from data collected from users. Profiled data can be related to: the interests, behaviors and preferences of the targeted users [45]. In this subsection, we look into how SPA data can be used to profile users.

3.3.1 En-route Profiling

A good instance of an en-route type of profiling is traffic analysis. Traffic analysis can be used to profile a user as shown in [46]. In particular, attackers can leverage en-route profiling to infer a user's presence. This can be further used to conduct more sophisticated attacks. En-route profiling attacks can be done even when the network traffic is encrypted. While there are obfuscation techniques that can be used to hinder these type of attacks, they have not been adopted in SPA. In this scenario, the most plausible adversary would be a dishonest or unethical Internet service provider. Governments or other global adversaries with access to the network traffic of a user can also exploit this weakness. The practicality of this threat on encrypted SPA traffic is shown in [46]. While authors in [46] perform traffic analysis without even needing a deep inspection of the network packages, MiTM techniques — such as SSL-stripping [47] — might be used to perform profiling over plain-text.

3.3.2 Profiling by the Third-Party Developers

As part of the measures to offer ubiquitous communication, SPA are integrated with third-party skills to boost its capabilities. This integration involves the sharing of valuable data between the SPA system and the third-party skills. Third-party skills require permissions to access users' information such as *location, mobile number, email address, name, device address, payment information and many more* — which users need to approve to use the skill. However, even when users are able to choose whether they share this information, they have no control over what the third party can do with the data, or what kind of inferences or aggregations they could make to derive other new personal information about the use, e.g., users' tastes. Even for those skills that do not ask for permissions, users do not know what can be learnt from interacting with the skill. Furthermore, malicious skills could collude to aggregate personal data from multiple skills similar to what we have seen in smartphone apps [48]. Here, skill connections pairing [42] may be leveraged to create colluding skills aiming at getting more elaborated profiling.

3.3.3 Profiling by the SPA Providers

SPA providers have a huge amount of data, which in most cases is actually needed to properly run the SPA architecture. For instance, SPA need to continuously learn from past computations to generate reliable, repeatable outcomes and decisions. To achieve this, SPA need a large training dataset of users' conversations. Upholding user's privacy is important, as dealing with data of this nature poses more challenging privacy connotations, including the sensitivity of such data, where the collected data is physically located, and any data retention periods [18, 49]. This is especially critical as advances in data analysis enable automated techniques to make sense of unstructured data at scale. Recent works such as [50] have shown that by using SPA data they could have the capability to profile the intimacy of a couple and infer how healthy their relationship is, through acoustic analysis of communication between them. Their analysis is able to understand the context of the conversation, including the semantics of everyday encounters and more complex interactions such as arguments.

3.4 Adversarial AI

As described in Section 2.2, for an SPA to understand what the users want, it needs to first understand what it is said. For this, the speech recognition system uses AI techniques like NLP and ML. However, these techniques can introduce the issues discussed below.

3.4.1 Adversarial ML

Conventionally, ML is designed based on the notion that the environment is safe and there is no interference during training and testing of the model [51]. However, such an assumption indirectly overlook cases where adversaries are actively meddling with the learning process [51]. ML is known to be vulnerable to specially-crafted input samples, described as adversarial examples, which are usually derived by slightly perturbing legitimate inputs [52]. These perturbations typically remain unknown to the person supervising the ML task. Most ML models that perform the same task tend to be affected by similar adversarial inputs even if they use different architectures and they are trained on different datasets [53]. This allows the attacker to easily craft adversarial inputs with little knowledge about the target ML model. As the SPA employs ML in decoding user's intent, specifically to match the user utterances to the correct text as discussed in Section 2.2, it is thus prone to adversarial attacks. Examples can target the machine learning models used by the SPA to poison the matching process done to transcribe the user utterances into text. An attacker can use this to generate a denial of service attack by causing misclassification of intents, to issue malicious commands, or to make the SPA invoke an incorrect skill [54]. For instance, an attacker can target a skill name “Boil” by registering a malicious skill as “Boyle” to mislead the SPA system. When any of these two words are uttered, it will be challenging for the SPA speech recognition system to properly differentiate among the two and transcribe in text, the correct and intended skill name.

3.4.2 Adversarial NLP

Unlike in adversarial ML where an attacker exploits the limitations of the underlying machine learning model used for speech recognition, here, the attacker targets other parts of the whole speech recognition framework. Following the example of skill invocation, the adversarial NLP problem comes once user utterances have already been transcribed into text and the system needs to decide which skill to invoke given the text (note the difference with the problem of translating into text two words with similar pronunciation). In particular, Amazon's Echo and Alexa seem to use the lengthiest string match when deciding which skill is called [36]. For example, the text “talk to Tutor Head for me please” will trigger the skill “Tutor Head For Me” rather than the skill “Tutor Head.” In a similar way to adversarial ML, such difficulty could be used by an attacker to intentionally trick users into invoking a malicious skill. This can be achieved by registering a skill with the same name (but longest possible string match) than a legitimate skill. Besides, there is currently no restriction on the number of skills that can be registered, hence, an adversary can register as many skills as possible to increase the possibility of getting their skills called.

3.5 Underlying and Integrated Technologies

Apart from the novel architectural elements unique to SPA, current frameworks also rely on other existing infrastructures like cloud services and smart devices. This means that they can potentially inherit or be subject to issues and vulnerabilities present in or arising from these technologies.

3.5.1 Cloud

The use of cloud storage and processing services redefines how and where SPA data is stored and accessed. While cloud technologies offer the advantage of having readily available virtually unlimited resources, they also present attackers new opportunities [55]. First, they are data-rich environments that are centrally located in a single point. If this element is breached, attackers may get access to highly valuable and sensitive information. Second, they usually offer multiple ways of accessing the data (e.g., web- or app-enabled access), widening the attack surface. Third, they can facilitate other issues mentioned above. For instance, as all data (even user utterances) are stored in the cloud together, they make it easier to conduct profiling.

3.5.2 Smart Home Devices

In the light of the numerous advantages offered by SPA, especially from the usability perspective, SPA are widely integrated with other smart home devices such as smart heating and cooling devices (e.g., Nest, or Ecobee 4), Smart security (e.g., Scout, or Abode), smart lighting devices (e.g., Philip Hue, or LIFX), Smart kitchen (e.g., GE+ Geneva) and surveillance cameras (e.g., Cloud Cam, Netgear Arlo Q). With such integration, a user can control his home temperature with voice by speaking the instruction to the SPA devices which is in turn instructed to the smart heating or the cooling device. As illustrated in Figure 1, the instruction is relayed to the smart device through the SPA provider cloud, the skill services and the smart device cloud. This integration brings the smart home into one verbally controlled system and offers the SPA the privilege to manage the services of other connected smart devices. Coincidentally, this integration also creates a single key point of interest to attackers. Attackers can take advantage of this in two ways. On the one hand, breaching the SPA can allow attackers to take control over a wide range of connected devices. More so, privacy issues could emerge from data accumulation, data acquisition, and integration as discussed in [56, 57], where the authors perform a comprehensive review of privacy threats of Information Linkage from data integration in IoT ecosystems. On the other hand, vulnerabilities in connected smart devices could be used as an intermediate step to attack the SPA [58, 59, 60]. Attacks in connected smart home devices have been investigated in numerous works, including snooping, privilege escalation, remote attacks, and insecure programming interfaces [59, 61]. For instance, in [58], authors describe a threat that allows IoT devices that are close to one another to spread a worm that propagates so rapidly, as long as the density of the vulnerable smart devices is more than a specific critical mass.

4 Attacks

In this section, we offer a review of known attacks on the SPA system and examine the vulnerabilities they exploit w.r.t. the issues described in Section 3 and the point they target in the architecture shown in Section 2. An overview of the most relevant attack papers together with the vulnerability they exploit and the point in the architecture are given in Table 1. We found that most of the attacks target the following elements of the architecture depicted in Figure 1:

1. **User to SPA device** (#1): There is a wide range of attacks targeting this point of the architecture. In particular, we identify related works i) exploiting weak authentication, and ii) attacking underlying and integrated technologies.
2. **SPA device to SPA service provider cloud** (#2): There is an attack reported in the literature that targets this point of the architecture and exploits en-route profiling.
3. **SPA service provider cloud** (#3): Several attacks are also found at this point of the architecture targeting the SPA cloud components. We identify works exploiting i) ML and NLP Vulnerabilities, and ii) underlying technologies.
4. **Remote access using mobile and Web** (#4): We identified related work remotely exploiting the data-rich environment offered by SPA cloud services.
5. **Third-party Web skills** (#6): Attacks targeting this point of the architecture exploit user misconceptions about the SPA system, and in particular about the skill. We show related works exploiting NLP subsystem vulnerabilities.

We could not find any attacks targeting architectural elements #5, #7, and #8. However, this does not mean that attacks targeting those architectural elements are not possible. In fact, some of the threats outlined in [59] and the attacks demonstrated by researchers in [62] could possibly exploit #8. Besides, some of the vulnerabilities that exist in #3 might also be found in #7 as they are both cloud technology. Likewise, attacks targeting #6, such as voice squatting and voice masquerading [36], might also be possible in #5 since both are skill services. Nevertheless, as far as we know, they have not been exploited yet. We discuss this more in detail later on in Section 6.

We next detail attacks in related works targeting the architectural elements mentioned above (#1-4 and #6), particularly looking at the vulnerabilities (described in Section 3) that they exploit and the assumptions they make on the environment.

Table 1: categorization of attacks found in previous studies based on vulnerabilities exploited and attack point.

Studies	Weak Authentication				Weak Authorization			Profiling			Adversarial AI		Integrated Techs.		Attack Point
	Wake-up Word	Always Listening	Synthesized Speech	Payment Auth.	Multiuser Environ.	Skills	Third-Party Access	En-route	Third-Party	SPA provider	Adversarial ML	Adversarial NLP	Cloud	Smart Devices	
Lei Xinyu et al. [33]	✓	✓		✓										✓	1
Zhang et al. [35]	✓	✓	✓												1
Chung & Lee [49]										✓			✓		4
Zhang et al. [36]						✓				✓		✓			3, 6
Kumar et al. [63]						✓				✓		✓			3, 6
Roy et al. [38]	✓	✓	✓												1
Gong & Poellabaeur [64]	✓	✓									✓				1, 3
Schönherr et al. [65]	✓	✓									✓				1, 3
Carlini and Wagner [66]	✓	✓									✓				1, 3
Vaidya et al. [54]	✓	✓	✓								✓				3
Carlini et al. [67]	✓	✓	✓								✓				3
Apthorpe et. al. [46]								✓							2

4.1 User to SPA Device (#1)

We observed that within the *weak authentication* category, *always on, always listening* and the *lack of arbitrary wake-up words* are the most exploited vulnerabilities at this point of the architecture. This is followed by a lack of protection against *synthesized speech.*

Lei Xinyu et al. [33] look at issues in single-factor authentication methods based on a wake-up word, and the lack of a mechanism that can be use to figure-out if a user is close-by or not. Using Amazon's Echo device as proof of concept, the authors perform a home burglary attack using Alexa to manipulate a connected door lock. Likewise, they successfully make a purchase using the compromised device.

The non-linearity in the Micro-Electro-Mechanical Systems (MEMS) microphone over ultrasound is exploited by Zhang et al. [35]. Non-linearity is describe as hardware features that cause signals with high-frequency trigger at high power to be shifted to low frequencies by microphones (and speakers) [38]. Even though microphones are designed to be a linear system, they exhibit non-linearity in higher frequencies. By synthesizing high-frequency sounds that are not within the human hearing range but are still intelligible to SPA devices, the authors are able to activate and control the voice of the SPA. This technique is called the Dolphin attack as it uses ultrasonic frequencies like what Dolphins use to communicate among themselves. This attack was confirmed on 7 popular voice intelligent assistants (Siri, Cortana, Huawei Hi Voice, Google Now, Samsung S Voice, and Alexa) over a range of different voice platform. On the downside, this attack cannot be conducted more than a distance of 5ft from the targeted device. Likewise, it requires special hardware to synthesize and play the ultrasonic signal, thereby making it unrealistic for a real-world attack.

In a different study, Roy et al. [38] develop a long-range version of Dolphin attack. They achieved a range of 25ft from their target. By exploiting the non-linearity inside the microphone, like in [35], they generated long range high-frequency signals that are inaudible to human but intelligible to SPA. As in the previous study, they control and issue commands to SPA devices with the assumption that the adversary can synthesize a legitimate voice signal. However, rather than using a single ultrasound speaker as done in [35] to play the synthesized signal, the authors used multiple speakers that are physically separated in space. They employ spectrum splicing to optimally slice voice command frequencies and play each slice on independent speakers in a way that the total speaker output is inaudible. Nevertheless, the attack is only feasible in an open environment. This is because high frequencies are more susceptible to interference which is a limiting factor to the distance [68]. Likewise, this attack requires multiple ultrasound speakers, making it more difficult to implement in a real-world attack.

4.2 SPA device to SPA service provider Cloud (#2)

At this point of the infrastructure, where an SPA device exchanges information with the SPA cloud provider, we found an attack that exploits *en-route* vulnerabilities within the *profiling* category. Authors in [46] identify privacy vulnerabilities with SPA by passively analyzing encrypted smart home traffic. Their study indicates that encryption alone does not offer all the necessary privacy protection requirements. The authors profile users interaction with Amazon Echo device by plotting send/receive rates of the stream even with encrypted traffic. This poses a serious privacy implication to smart home users as an attacker can use this to infer their lifestyle and the best time to conduct an attack undetected as discussed in Section 3.3.1. However, the method used in this study might not be applicable to a situation where different IoT devices communicate with the same domain because of the difficulty of labelling streams by device type.

4.3 SPA Service Provider Cloud (#3)

Here, we discuss attacks exploiting the SPA cloud subsystem where the speech recognition and the intent identification are performed. Specifically, attacks at point #3 of the architecture exploit *Adversarial ML* vulnerabilities described in Section 3.4.

Looking at where data-driven Machine Learning (ML) models operate, authors in [64] show a new end-to-end scheme that creates adversarial examples by perturbing the raw waveform of an audio recording. With their end-to-end perturbation scheme, the authors crafted adversarial inputs that mislead the ML model. Note that this is widely used in para-linguistic applications. Their adversarial perturbation has a negligible effect on the audio quality and leads to a vital drop in the efficiency of the state-of-the-art deep neural network approach. On the downside, such attack needs to be embedded in a legitimate audio signal to make them truly obscure.

More recently, Schönherr et al. [65] have proposed an adversarial example based on psychoacoustic hiding to exploit the characteristics of Deep Neural Network (DNN) based ASR systems. The attack extended the initial DNN analysis process by adding a back-propagation step to study the level of freedom of an adversarial perturbation in the input signal.

It uses forced alignment to identify the best temporal fitting alignment between the maliciously intended transcription and the benign audio sample. It is also used to reduce the perceptibility of the perturbations. The attack is performed against Kaldi[3], where it obtained up to 98% success rate with a computational effort for a 10-secs sound file in less than 2-mins. However, like in [64], this attack also needs to be embedded in another audio file which greatly influences the quality of the adversarial example.

Another important study conducted by Carlini and Wagner in [66] proposes an attack on speech recognition systems using Connectionist Temporal Classification (CTC) loss. They demonstrated how a carefully designed loss function can be used to generate a better lower-distortion adversarial input. This attack works with a gradient-descent based optimization [69] and replaces the loss function with the CTC-loss, which is optimized for time sequences. However, the audio adversarial examples generated when played over-the-air cease to be adversarial, and thus, make it unrealistic for a real-world attack.

Similarly, Vaidya et al.[54] perform an attack on speech recognition systems using unintelligible sound. This is done by modifying the Mel-Frequency Cepstral Coefficients (MFCC) — feature of the voice command. The attack is done in two steps: first, altering the input voice signal through feature extraction with adjusted MFCC parameters, and then regenerating an audio signal by applying a reverse MFCC to the extracted features. When put together, this attack is able to craft a well designed adversarial input. The MFCC values are selected in a way that they can create a distorted audio output with least sufficient acoustic information. This audio output can still achieve the desire classification outcome and is correctly interpreted by the SPA while being unintelligible to human listeners. Although this attack successfully exploits the differences between how computers and humans decode speech, it could, however, be detected if a user is in proximity — provided that they hear unsolicited SPA responses. The attack presented by Vaidya et al. [54] is extended in the work of Carlini et al. [67], where the authors test the attack effectiveness under a more realistic scenario and craft an adversarial example completely imperceptible to humans by leveraging the knowledge of the target speech recognition system.

4.4 Remote Access Using Mobile and Web (#4)

An Attack can also take advantage of the remote access offered by the SPA cloud services. By exploiting *cloud* vulnerabilities within the *underlying and integrated technologies* category, an attacker can conduct profiling attacks on the users and infer their behavioral characteristics. In the work of Chung & Lee [49], the authors demonstrate how personal information can be inferred from SPA cloud data by using a forensic toolkit that extracts valuable artifacts from Amazon Alexa by taking advantage of the Alexa unofficial Application Programming Interfaces. By aggregating the data collected from mobile Apps, Alexa cloud service and web browsers [70] using their toolkit, they were able to uncover information such as: user interests, usage patterns and sleeping patterns. This shows the possibility of profiling with cloud data and highlights the privacy implications of the SPA to the users and vendors. However, one limitation of this study is that the authors were only able to access the cloud-native database on the assumption that there is access to valid user's credentials. Without a valid user's credential, it is not possible to collect cloud-native data, thereby making such attack less likely except if conducted by a malicious SPA service provider, who does obviously not require credentials before accessing such data, which is in their infrastructure.

4.5 Third-party Skills (#6)

Other related works exploit how SPA skills are invoked and the way they interact with each other. In fact, these attacks normally exploit one way or other *Adversarial NLP* vulnerabilities described in Section 3.4.

Authors in [36] target the interaction between third-party skills and the SPA service. Specifically, they analyze two basic threats in Amazon's Alexa and Google's Assistant SPA services: voice squatting and voice masquerading. Voice squatting allows an attacker to use a malicious skill with longest matching skill name, similar phonemes or paraphrased name to hijack the voice command of another skill as described in section 3.4.2.

In five randomly sampled vulnerable target skills, the authors successfully "hijacked" the skill name of over 50% of them. The feasibility of this type of attack is high, particularly in SPA such as Alexa that allows multiple skills with the same invocation name. This attack can be used to damage the reputation of a legitimate skill as any poor services of the malicious skill will be blamed on it.

Equally, in voice masquerading attack, a malicious skill pretend to invoke another skill or fake a skill termination. Then, the malicious skill keeps recording the user's utterances. This attack could be used to snoop on the conversations of the user. While voice squatting attack exploit the weaknesses in the skill's invocation method, voice masquerading targets

[3] A widely adopted open-source toolkit written in C++ which offer a wide range of modern algorithms for Automatic Speech Recognition.

user's misconceptions about how SPA skill-switch services work. With some skills requesting for private information, an adversary could use these attacks to obtain sensitive information and cause crucial information leakage to unwanted parties. Voice squatting attack is also shown in the work of Kumar et al. [63]. But unlike what was done in [36], Kumar et al. use the intrinsic errors in NLP algorithms and words that are often misinterpreted to craft malicious skills and exploit the ambiguity in the invocation name method.

5 Countermeasures

In mitigating the identified risks and attacks, there have been a number of studies proposing various countermeasures. In this section, we summarise research on countermeasures, highlighting limitations and deficiencies. We give a summary of these in Table 2. We categorized the proposed countermeasures by using the issues discussed earlier. It should be noted that we do not use their effectiveness to justify the categorization but rather, the possibility of the solution mitigating these issues. The current mitigation level in the table (last row of Table 2) aims to provide a quick indication of the extent of the issues identified having been resolved by the countermeasures and methods proposed by the existing publications analyzed to date. We also map these countermeasures to the elements of the architecture depicted in Figure 1 to describe the mitigating point. Most countermeasures map to:

1. **User to SPA device** (#1): There is a wide range of countermeasures proposed to mitigate attacks at this point of the architecture. In particular, we found many related works mitigating *weak authentication* vulnerabilities.
2. **SPA device to SPA service provider cloud** (#2): At this point of the infrastructure, we found studies proposing different mitigation techniques to obfuscating traffic between the SPA device and the SPA service provider cloud, with the aim to mitigate *en-route* vulnerabilities within the *profiling* category.
3. **SPA service provider cloud** (#3): Few of the existing countermeasures also focused on the *Adversarial AI* vulnerabilities that are found at this point of the architecture and recommended measures aim to mitigate the risks associated with them.
4. **Others**: Countermeasures in this category modify to some extent the existing SPA architecture as part of the mitigation and/or mitigate vulnerabilities that cut across multiple points of the infrastructure. We mapped these countermeasures to multiple elements of the architecture to signal the points where the mitigations apply or the points that would change as part of an architecture modification.

5.1 User to SPA Device (#1)

This section discusses countermeasures intended to mitigate attacks targeting the voice interaction between users and SPA devices. Weak authentication is the most commonly exploited vulnerability at this point of the architecture as discussed in Section 4, and we can see in Table 2 that it is also the vulnerability that has been considered the most in terms of countermeasures. In particular, we found *synthesized speech* to be the one receiving most attention from the research community.

One of the first defence that has been put in place against weak authentication is voice authentication. With this defense, the SPA can tell apart individual users when they speak. For instance, some SPA such as Google and Amazon perform speaker verification through voice authentication, known as Voice Match [71] and Voice Profiles [72] respectively. However, none of these mechanisms is enabled by default and it is left to the users to first realize about their existence and then decide whether they would like to activate them or not. Even when these mechanisms are activated, they are still open to attack as an attacker can still trick the system with a collected or synthesized voice sample of the legitimate user [73]. Collecting voice samples is an easy task since the human voice is open to the public. Unlike passwords that can easily be changed if compromised, a human voice is a feature that is difficult to replace. Another important voice authentication method is proposed in [74]. In this study, the authors present a continuous authentication VAuth system that aims to ensure that the SPA works only on commands from legitimate users. The solution consists of a wearable security token that repeatedly correlates the utterances received by the spa with the body-surface vibrations it acquires from the legitimate user. The solution was said to achieve close to 0.1% false positive and 97% detection accuracy and works regardless of differences in accents, languages, and mobility. Even though this system achieves a high detection accuracy, the need to wear devices such as eyeglasses, headset, and necklaces would introduce a potentially unbearable burden and inconvenience on the users.

Kepuska and Bohouta [75] also proposed a multi-modal dialogue system that combines more than one of voice, video, manual gestures, touch, graphics, gaze, and head and body movement for secure SPA authentication. Even though this system might be able to solve the authentication and voice impersonation challenges earlier discussed, the authors have only been able to test the individual components of the system and not the entire system as a whole.

Another important measure implemented against weak authentication is user presence-based access control system. This system allows an SPA to verify if a user is truly nearby before accepting any voice commands. Lei Xinyu et al. [33] propose a solution that uses the channel states information of the router Wi-Fi technology to detect human motions. Interestingly, it eliminates the need for some wearable devices and introduces no added development cost as it uses the existing home Wi-Fi infrastructure. The solution has an advantage over the traditional voice biometrics recognition, i.e.: that becomes ineffective as users age, become tired, or ill. However, the effectiveness of the system depends on selecting the best location for the Wi-Fi devices and setting the right parameters for the detection. Besides, it only supports commands that come from the same room where the SPA device is deployed: in their case, an Amazon Echo. Likewise, the system is situational as it works best if there is no structural change to the location where the devices are deployed.

We observe another category of countermeasures, at this point of the architecture, aiming at protecting the SPA against synthesized speech attacks. In the work of Roy et al. [38], the authors propose a system nicknamed Lip Read that is based on the assumption that some of the features of voice signals–basic frequencies and pitch–is preserved when it passes through non-linearity. It was reported that this system obtains a precision rate of 98% and a recall rate of 99% in a situation where the adversary does not influence the attack command. However, there is no formal guarantee of this countermeasure as they are unable to model the frequency and phase responses for general voice commands. Likewise, their defense only considers inaudible voice attack ignoring finding the true trace of non-linearity. Similarly, Zhang et al. [35] propose another set of countermeasures against synthesized speech attacks. The authors recommend 2 hardware-based mitigating measures—the first one aim to enhance the microphones use by the SPA devices while the latter hardware-based defense is intended to cancel any unwanted baseband signal. Enhancing the microphones approach entails designing an improved microphone similar to the one found in Phone 6 plus that can subdue any ultrasonic sound. On the other hand, cancelling the unwanted baseband signal of the inaudible voice command solution entails the introduction of a module before the low pass filter in the subsystem used for voice capturing to identify and cancel the inaudible voice commands baseband signal. Likewise, the software-based countermeasure relies on the principle that a demodulated attack signal can be distinguished from legitimate ones using a machine-based learning classifier.

In another important countermeasure against synthesized speeches, Chen et al. [73] propose a software-only impersonation defensive system. The system is developed base on the notion that most synthesized speech needs a loudspeaker to play the sound to an SPA device. As conventional loudspeakers generate a magnetic field when broadcasting a sound, the system monitors the magnetometer reading which is used to distinguish between voice commands from a human speaker and a loudspeaker. In a situation where the magnetic field emitted is too small to be detected, the system uses the channel size of the source of the sound to develop a means of authenticating the sound source. However, the effectiveness of the system depends heavily on the environmental magnetic interference. Likewise, the source of the sound needs to be at a distance of more than 2.3in (6cm) to their system to prevent the magnetic field from interfering with the magnetometer's reading. In addition, the system has a high false acceptance rate when the sound source distance to their system is greater than 4in (10cm) in a situation where the loudspeaker magnetic field is un-shielded and less than about 3in (8cm) when shielded.

Table 2: Categorization of countermeasures found in related studies.

Studies / Techniques	Weak Authentication				Weak Authorization			Profiling			Adversarial AI		Integrated Techs.		Mitigating Point
	Wake-up Word	Always Listening	Synthesized Speech	Payment Auth.	Multiuser Environ.	Skills	Third-Party Access	En-route	Third-Party	SPA provider	Adversarial ML	Adversarial NLP	Cloud	Smart Devices	
Voice Match / Profiles [71, 72]			✓	✓											1
Kepuska and Bohouta [75]	✓	✓	✓	✓											1
Lei Xinyu et al. [33]	✓	✓													1
Huan et al. [74]	✓	✓	✓	✓											1
Roy et al. [38]			✓												1
Zhang et al. [35]			✓												1
Chen et al. [73]			✓												1
Lavrentyeva et al. [76]			✓												3
Zhang et al. [36]												✓			3
Kumar et al. [63]												✓			3
Gong & Poellabaeur [64]			✓								✓				3
Liu et al. [77]								✓							2
Coucke et al. [78]	✓					✓	✓	✓	✓	✓			✓		others
Current Mitigation Level	◐	◐	●	◐	○	◐	◐	◐	◐	◐	◐	◐	◐	○	

○-NOT YET ADDRESSED, ◐- NOT FULLY ADDRESSED, ●- WELL ADDRESSED

Finally, Lavrentyeva et al. [76] explore different countermeasures to defend against voice replay attacks. Even though the countermeasure is implemented at #3 of the architecture because it needs extensive computational power, it aims at securing #1. The researchers use a reduced version of Light Convolutional Neural Network architecture (LCNN) based on the method of the Max-Feature-Map activation (MFM). Their LCNN approach with Fast Fourier Transform (FFT) based features obtained an equal error rate of 7.34% on ASVspoof 2017 dataset compared with the spoofing detection method in [79] with an error rate of 30.74%. They further utilized Support Vector Machine (SVM) classifier to offer valuable input into the efficiency of their system. Consequently, their primary system based on systems scores fusion of LCNN (with FFT based features), SVM (i-vector approach), recurrent neural network and convolutional neural network (with FFT based features) shows a better equal error rate of 6.73% on their evaluation dataset.

5.2 SPA device to SPA service provider Cloud (#2)

At this point of the infrastructure where SPA devices exchange information with the SPA provider cloud, Liu et al. [77] propose a countermeasure to mitigate *en-route* vulnerabilities (part of the *profiling* category). The authors present a countermeasure that protects smart home against traffic analysis—a community-based "differential privacy framework". The framework route traffic between different gateway routers of multiple cooperating smart homes before sending it to the Internet. This mask the source of the traffic with little bandwidth overhead. Nevertheless, this approach requires cooperation from multiple homes, which makes it difficult to implement. In addition, it could result in long network latency if the homes are not geographically close.

5.3 SPA service provider Cloud (#3)

Here, we discuss the countermeasures for securing point #3 of the architecture. In particular, the NLP and the ML vulnerabilities as part of the *Adversarial AI category*.

Zang et al. [36] present a system that examines the skill's response and the user's utterance to detect voice masquerading attacks. The system relies on a User Intention Classifier (UIC) and a Skills Response Checker (SRC). The SRC semantically analyze the response from the skill and compares it against utterances from a black-list of malicious skill responses to flag off any malicious response. While the user UIC, on the other hand, protects the user by checking their utterances to correctly determine their intents of context switches.[4] This is done by matching the meaning of what the user says to the context of the skill the user is presently interacting with and also that of the system commands. They also consider the link between what the user says and the skill that they are currently using. UIC complement the SRC and their system reports an overall detection precision rate of 95.60%. Nevertheless, one key shortcoming of this system is the difficulty in implementing a generic UIC due to variation in Natural language-based command and how to distinguish legitimate commands.

In a similar study, Kumar et al. [63] suggests performing phonetic and text analysis for every new skill's invocation name to mitigate voice squatting attacks. They check whether the new skill's invocation name can be mistaken with an existing one, vetting then the creation of the clashing skill. Their solution is similar to what is currently been implemented during domain registration where registrars do not allow registration of domains that resemble that of popular domains.

5.4 Others

In this section, we discuss countermeasures that cut across different points of the architecture. In particular, we discuss the work proposed by Coucke et al. [78], which proposes changes to the architecture, particularly in terms of the speech recognition functionality. Coucke et al. [78] present a *privacy by design spoken Language Understanding platform* that does not send user queries to the cloud for processing. The speech recognition and the intent extraction are done locally on the SPA devices themselves using a model that is partially trained with *crowd-sourced* data and using *semi-supervised* learning. Many use cases do not need Internet access. However, when the use case requires internet access, such as when data needs to be retrieved or transmitted to an Internet service, then the system processes the data within the SPA device where it was generated rather than in the cloud. This makes it hard for an adversary to perform a mass attack as they can only target a single user or device at once. With such an infrastructure, issues related to *always on always listening*, *cloud*, and *third-party access*, have limited impact since the data is processed locally. Besides, it allows personalizing the wake-up word, mitigating the wake-up word vulnerability introduced in Section 3.1.1. However, the platform requires a user to specify the skills on which their assistant will be trained on. Hence, such assistant can only work within predefined scopes of the selected skills on which their model was trained, thereby restricting their capabilities to only those skills used for their training. It is important to also note that,

[4] This is, examining the intents of changing from one task to the other.

although this infrastructure modifies the existing SPA architecture so that speech recognition and intent identification is conducted locally, it does not completely eliminate data transmission to other devices or cloud services. The SPA still communicates with other connected devices or cloud services depending on the context of use. This means that attacks like the one described in [58] may still be possible.

6 Discussion and Open Challenges

Building on the analysis and categorization of the related literature studied in the previous sections, we then offer a synthesis and summary of this review and suggests areas for future research.

One can easily observe in Table 1 that vulnerabilities related to weak authentication are the most exploited flaws. The *wake-up word* and the *always listening features* are typically combined and can be described as the gateway of synthesized speech attacks. No related works currently exploit the multiuser environment and third-party access. We also observed that the majority of the attacks target point #1 of the architecture: the point of interaction between the users and the SPA devices as it requires an attacker with lower capabilities. Although few attacks exploit more than one point of the architecture — e.g. [36, 63, 64], none is observed at point #5, point #7 and #8 even though attacks targeting those architectural elements seem possible as discussed in Section 4. Similarly, Table 2 shows that countermeasures for *weak authentication* vulnerabilities, and in particular countermeasures towards mitigating synthesized speech, have received wide attention in the literature. Taking both Table 1 and Table 2, we can see a concentration of research efforts towards one very specific part of the whole SPA architecture, the direct interaction between the user and the smart speaker — or point #1 of the architecture. While indeed this is an important part of the architecture, SPA should consider security in a holistic manner. This shows that despite the growing research efforts in security and privacy in SPA, we, as a community, need to also recognise and tackle SPA problems that go beyond that point of the architecture. Based on our findings, we suggest a number of open challenges in SPA. These include: i) a practical evaluation of existing attacks and countermeasures, ii) making authentication and authorization stronger as well as smarter, building secure and privacy-aware speech recognition, iii) conducting systematic security and privacy assessments to better understand the SPA eco-system and associated risks, iv) increasing user awareness and the usability of security and privacy mechanisms in SPA, and v) understanding better profiling risks and potential countermeasures. All of which are discussed below in the following subsections.

6.1 Practical Evaluation of Existing Attacks and Countermeasures

We found that many of the attacks target the underlying hardware of the voice infrastructure. For instance, [38] and [35] use high frequencies signal to attack the non-linearity in SPA devices microphones. While some of these attacks synthesize speech in a way that may be intelligible to human and easily noticed by users in proximity [33], there are other attacks that synthesize speech in a way that is unintelligible to the users [35, 38]. Thus, one could argue that the second type of attacks is more likely to be successful in practice than the first type. In addition, our study revealed that many of the attacks require different domain-specific knowledge to be successful, which might not always be available. For example, attacks conducted in [38, 54, 64] need knowledge of the machine classifiers, while the one demonstrated in [36] needs the understanding of the SPA skills invocation model. In some cases, this knowledge is available or can be reverse-engineered from interactions with the SPA and their architecture. However, beyond these observations that we can derive from a literature review, research efforts to evaluate and compare the severity, feasibility, cost, and likelihood of success in practice across existing attacks and countermeasures are currently missing.

6.2 Making Authentication Stronger

Despite receiving most of the attention in terms of countermeasures, with some of the issues and attacks having a counterpart countermeasure, weak authentication issues have not been completely addressed yet. As discussed earlier, many of the attacks targeting the SPA system exploit its weak authentication especially the *always on, always listening features*. This attack is usually combined with other vulnerabilities. Although one could say that the *always on, always listening features* improve the responsiveness of the devices by making resources available to the user before they start uttering commands, the security and privacy risks may outweigh the benefit. Several independent input variables such as voice, video, manual gestures, touch, graphics, gaze, and others like the solution proposed in [75] could be combined to make authentication stronger. However, most SPA are designed without environmental sensors. The lack of environmental sensors makes it difficult to implement context-aware authentication systems that could sense the physical environment, and leverage such information to adjust the security parameters accordingly. Also, there may be privacy issues and concerns when using even more personal information (e.g., video). Likewise, current authentication mechanisms in integrated technologies like other smart home devices are decentralized. Each integrated technology has its own authentication mechanism. By implementing a centralized mechanism, potentially in an SPA, a user could gain access to multiple integrated technologies by authenticating only once. This would not only enhance usability by lessening the authentication burden on users, but also improve security as it would ensure consistent authentication across smart home devices.

Future research can also consider how communication protocols may improve current authentication mechanisms in SPA. There are examples of how these mechanisms can be used in other systems such as contactless smart cards, where they are becoming an effective way to verify users' presence [80]. Popular among them are the distance-bounding protocols which can be used to authenticate the user and access their location. These protocols have proven to be practicable especially in a system that is susceptible to distance-based frauds. Distance-bounding protocols are based on timing the delay between when a verifier sends a challenge to the moment the response is received. This allows the verifier to detect a third party interference as any sudden delay in the proper response, which is considered to be the result of a delay due to a long distance transmissions [81, 80]. Nevertheless, the effectiveness of this protocol depends on getting the correct propagation time.

6.3 Enhanced Authorization Models and Mechanisms

More flexible access control and authorization models and mechanisms are needed. This mechanism should be able to dynamically authorize and adapt permissions to users based on the current context and their preferences. According to a recent study, users preferred authorization policies in smart homes are affected by some distinct factors [82]: i) the capabilities within a single device, ii) who is trying to use that capability, iii) and the context of use. Hence, designing authorization models that consider SPA capabilities in addition to the context of use may help to create authorization rules that adequately balance security, privacy, and functionality. In fact, similar models have already been implemented successfully in other domains like smartphones [83]. Furthermore, we have observed that SPA system require more fine-grained authorization mechanisms. This not only applies to the voice of the user itself, but also to the data that can be obtained from the way in which users interact with the devices. In particular, these interactions can be used to infer, for instance, the sleeping patterns of a user as discussed earlier.

Novel authorization models and mechanism for SPA should not only consider single users, but also multiple users. However, there are no security and privacy mechanisms for SPA that considers *multi-user environment* issues. This is important, as even if SPA would support multiple accounts, it is a common practice to share accounts between multiple users [84] (especially if one of the accounts has more privileges). The lack of proper authorization can prompt insider misuse, e.g.: members of the household spying on their partners [85], which can be particularly problematic in the case of intimate partner abuse [86]. Moreover, smart home data is relational and it usually refer to a group of people collectively [87], e.g., if there is a way to infer whether there is someone at home or not, this already gives information that can be sensitive to everyone living there. Some general-purpose smart home privacy-enhancing IoT infrastructures like the Databox [87] recognize the multiuser problem but no solution has been proposed yet in general for smart homes or in particular for multiuser sharing management in SPA. A great deal of research on methods and tools to help users manage data sharing in multiuser and multiparty scenarios have been proposed for social media (see [88] for a survey), and particular methods for detecting and resolving multiuser data sharing conflicts, such as [89], could be adapted from there or used to inspire multiuser solutions for the SPA case.

6.4 Secure and Privacy-aware Speech Recognition

NLP and ML models are used in conjunction for speech recognition. Protecting these models against manipulation, e.g.: through well-crafted adversarial inputs as pointed out in Section 3.4, becomes paramount. It is apparent from Table 1 and Table 2 above that there are many attacks exploiting adversarial ML and NLP issues, and there are substantially more attacks than defences studied in the related literature. SPA providers need to consider adversarial examples when developing their speech recognition models. However, that is not an easy task and more research is required in this direction. Some existing countermeasures used in other domains such as adversarial training and distillation could help to develop robust ML models for speech recognition in SPA, but they can be defeated using black-box attacks or attacks that are constructed on iterative optimization [90]. Also, validating the input and reprocessing it to eliminate possible adversarial manipulations before it is feed to the model is a countermeasure that greatly depends on the domain, and is subjected to environmental factors [51]. Likewise, testing is inadequate to secure ML, as an adversary can use a different input from those used for the testing process.

Furthermore, the performance of the current speech recognition system still deserves improvement as shown earlier — recall that these systems often find it difficult to i) understand words with similar phonemes [63], ii) understand different but similar words, and iii) resolve variation in natural language-based command words [35]. Since the word error rate (WER) is the common metric used for evaluating the performance of automatic speech recognition systems [91], it may be easy for an adversary to craft an adversarial input that could maximize the WER of the speech recognition system by exploiting the NLP framework and the ML techniques. This is shown in [35], where the speech recognition system is exploited to manipulate the intent the system understands from the user's command.

Beyond security, obtaining valuable information from big data while still protecting user's privacy has become interesting research in data analysis. While SPA providers let users review and delete their voice recordings, a recent study shows that users are unaware (or do not use) those privacy controls [18]. Thus, it is also unclear how effective these controls actually are, e.g.: these controls allow the user to delete particular raw utterances but they can not delete what it could be inferred from them (i.e., the model). In light of this, SPA vendors need to understand the privacy challenges of machine learning. For instance, although most existing SPA providers aim to ensure privacy while processing users' voice in the cloud, that is a difficult endeavour with current SPA architectures. With edge computing gradually coming into the limelight, data can now be processed locally, where it is generated, rather than being transmitted to a centralized data processing centre [92]. This helps to reduce the current dependency on the Internet and eliminate the necessity of putting sensitive data into the cloud. While related work [78] address this direction with a decentralized voice processing platform, it is difficult to build a general purpose SPA using such platforms. This is because SPA developed with such platforms can only work within predefined scopes of the selected skills on which their model was trained. There is, therefore, the need for future efforts on how to effectively make voice processing privacy-preserving without hindering the capabilities of SPA.

6.5 AI-based Security and Privacy

In addition to using AI techniques for SPA functionality, e.g., speech recognition, they could also be used to make SPA more secure and aid users to manage their privacy as they see fit. AI techniques would include not only data-driven techniques like Machine Learning but also knowledge-based techniques such as normative systems and argumentation, which have been successfully used to develop intelligent security and privacy methods in other domains [93, 94]. Examples of the use of ML techniques include methods to detect malicious commands being spoken to the SPA devices (i.e., to make authentication stronger and more resilient to attacks), and to help users configure the permissions they grant to third-parties skills to have better authorization mechanisms. Similar research has already been shown to detect intrusions [95] and to help users in other domains like mobile App permission management [96] and Social Media privacy settings and data sharing [97]. As for the speech recognition, these ML-based methods need to be engineered considering adversarial cases [69].

Examples of the use of knowledge-based AI techniques include the use of norms, which have been widely explored in recent years, especially as a means of reducing the autonomy of autonomous and intelligent systems to conform to decent behaviours [98]. Norms are usually delineated formally using deontic logic to state what is permissible, obligatory and prohibited, providing a rich framework to express context-dependent policies, e.g., based on Contextual Integrity [99], and they can be defined and verified for socio-technical systems like SPA [100]. For instance, norms would be very useful to avoid issues like the case discuss in [12] where a private conversation is recorded by an Alexa and forwarded to an arbitrary contact [12], as a norm could specify the type of conversations that may or may not be shared with particular contacts. Another example is norms that govern multiuser interactions with the SPA as discussed in Section 6.3. Norms for SPA could be elicited automatically as in [101] or by crowd-sourcing the acceptable flows of information as in [102]. Other AI techniques that could be used include negotiation [103, 104] to help SPA users navigate the trade-offs and negotiate consent in the very complex SPA ecosystem, including third-party skills and smart devices; and computational trust [105] to choose and only share data with third-party skills and smart devices that are privacy-respecting and trustworthy.

6.6 Systematic Security and Privacy Assessments

SPA are a type of cyber-physical system. Previous research looked at how the assurance techniques and testing methodologies most commonly used in regular IT systems [106], such as penetration testing, static & dynamic analysis, fuzzing, formal verification, etc., apply to cyber-physical systems. Assurance techniques are known to have different cost-effectiveness in practice [107], and that cost-effectiveness for one very same assurance technique has been shown to vary across different cyber-physical systems [108], such as Industrial Control Systems [109]. Therefore, a direction for future research is to study and evaluate how these assurance techniques will perform for the case of SPA and whether or not SPA's unique features like voice recognition and its integration with other technologies like the cloud and other smart devices require novel techniques or methodologies. For instance, the potential to have composite vulnerabilities that exploit both the physical and the IT part of cyber-physical systems [110] has already been shown to also apply to SPA, e.g., [38]. Additionally, authors in [35] show that physical properties can be used to compromise the SPA by using high frequencies signals to attack the non-linearity in SPA devices microphones as detailed above in section 4.1.

Future work should also look at the best and most systematic way to conduct privacy assessments in SPA [111]. Of particular interest might be to study the (extent of) traceability between the actions of the data specified in privacy policies, such as those in the privacy policies of the third-party skills developers in SPA, and the related data operations obvious to users via SPA and/or associated smartphone interfaces, which will also be crucial to help tackle the current *weak authorization* and *profiling* issues of SPA. Methodologies for this could be adapted from the social media [112] and smartphone apps [113], which already showed the extent of traceability in these domains, together with methods to help developers automatically map traceability between policies and operationalized controls and maintain it through the development cycle [114]. As real breaches like those already mentioned before, e.g. [12], happen methods to study whether there are gaps in security and privacy policies, such as [115], applied to SPA would also be helpful. Likewise, a longitudinal study is required to comprehend the SPA skill's ecosystem in order to understand the type of skills available, the capabilities they have, how they are been used, with who is behind them (number of third-party developers, etc.). This will further ensure a better understanding of the different kinds of risk that the ecosystem present and aid in formulating appropriate security and privacy policies for the users.

6.7 User Awareness and Usability

Although implementing a technical defensive measure might go a long way in mitigating some of the identified risks, effective countermeasures will be difficult without better user awareness and usability. In fact, most users still believe they do not have any valuable information that will make them a valuable target for attackers, while some are uninformed, or unconcerned when it comes to security and privacy issues in smart home devices [116]. It is worth mentioning that some non-technical recommendations have also been suggested to improve SPA security. This includes muting the microphone or switching off the SPA devices when they are not in use [117], with research actually showing some users do in practice with their SPA [18]. Nonetheless, a primary concern of these recommendations is that they affect the SPA devices usability as users might see them as an additional burden thereby refusing to implement them. Since convenience and connectedness are the main concern for smart homeowners and dictate privacy opinions as well as direct owners' attitude towards external entities that design and regulate these devices [118], vendors and developers must consider usability in any control measures to be implemented.

Nevertheless, while factoring usability in any control measures will help towards having a secure and privacy-aware SPA, it is also important to note that users are often not good at implementing appropriate control measure. As mentioned earlier, many users are unconcerned when it comes to security and privacy issues in smart home [116]. With this in mind, there is a need for improved design policies and practices that guarantee secure and privacy-aware SPA without extensive user involvement. For instance, AI-based techniques as discussed in Section 6.5 could be adopted to predict user's preferences and to automatically improve the usability, palliating the limitations that these devices pose to HCI. In fact, automated techniques (e.g., using ML) approaches to factor in (predict) the user's preferences could be an interesting open direction as mentioned in the previous section. Another important area for future research direction to help users manage their data in SPA is to design usable novel notice and control mechanisms for SPA. Notice and control mechanisms must be relevant, actionable and understandable as discussed in [119]. In addition, four main dimensions should be considered when designing those mechanisms: timing, when should a notice be presented; channel, how should the notice be delivered; modality, how the information should be conveyed; and control, how choice options are integrated into the notice.

6.8 Profiling Attacks and Defences

Regarding profiling, we can clearly see in Table 1 that few attacks have been reported on this. Some of these attacks make some hard assumptions, like having access to all cloud data about a user through their user account. We believe that further research is needed in order to assess whether other types of more sophisticated profiling could be conducted with access to less information. Furthermore, the community needs to understand whether tracking, which is pervasive across the web [120], could also apply and be feasible across the SPA ecosystem. In terms of defences, we can also see in Table 2 the lack of work in this area. Some of the challenges we mentioned before would indeed help alleviate profiling such as user awareness and usable controls (Section 6.7), systematic privacy assessments (Section 6.6), and knowledge-based AI techniques to express/verify norms about how data are collected and use of data across the SPA ecosystem (Section 6.5). However, other open challenges would remain and profiling-specific countermeasures are also needed. For instance, SPA traffic needs to be properly obfuscated and masked to encode user's interaction with the devices in addition to the existing encryption mechanisms already in place. Note that current encryption mechanisms are not sufficient to avoid traffic profiling as shown in [46]. Beyond differential-private approaches like the countermeasure introduced earlier [77], one possible avenue would be to adapt existing mechanisms to the case of SPA, such as traffic morphing techniques [121] to prevent statistical traffic analysis. This can be done by altering one category of traffic to look like another one. However, this and most other existing traffic analysis countermeasures are vulnerable as they only obfuscate exposed features of the traffic by muffling this features and adding dummy packets. Thus, they are unable to prevent the leakage of many identifying information [122]. Another avenue could be based on mix networks [123] and/or onion routing [124]. However, both of them may also be vulnerable to attack. For instance, mixing is susceptible to long term correlation and sleeper attacks [123]. Finally, onion routing is susceptible to an adversary correlating the traffic [125] and to misconfigured and malicious relays [126].

7 Conclusions

This paper analyzes and classifies the security and privacy issues associated with SPA and how a range of malicious actors can exploit them to harm the security and privacy of end users. We have shown that the attack surface of this increasingly popular technology is very broad. We have noted that the interaction between the users and the SPA devices is currently the weakest link. However, we have identified a wide range of attacks that can put users at stake.

In as much as there is no single panacea solution for all security issues, the proper understanding of security pitfalls will go a long way in enabling manufacturers, researchers, and developers to design and implement the robust security control measures. Although there is already very active research on securing intelligent assistants, few of the approaches consider the whole picture of the complex architecture SPA have. We particularly highlighted open challenges for future research that we deem of critical importance, including making authentication stronger, enhancing authorization models and mechanisms, building secure and privacy-aware speech recognition, conducting systematic security and privacy assessments, developing AI-based security and privacy countermeasures, improving user awareness and usability, and studying further profiling attacks and defences.

References

[1] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–23, 2018.

[2] CANDACE KAMM. User interfaces for voice applications. *Colloquium Paper*, 92:10031–10037, 1995.

[3] Kwan-Min Lee and Clifford Nass. Social-psychological origins of feelings of presence: Creating social presence with machine generated voices. *Media Psychology*, 7(1):31–45, 2005.

[4] Toni Reid. Everything Alexa learned in 2018. `https://blog.aboutamazon.com/devices/everything-alexa-learned-in-2018`, 2018. [Online; last accessed 4-January-2019].

[5] OVUM. Virtual digital assistants to overtake world population by 2021. 2017.

[6] Bret Kinsella. The Information Says Alexa Struggles with Voice Commerce But Has 50 Million Devices Sold. `https://voicebot.ai/2018/08/06/the-information-says-alexa-struggles-with-voice-commerce-but-pass`, 2018. [Online; last accessed 7-January-2018].

[7] Ewa Luger and Abigail Sellen. "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 5286–5297, New York, NY, USA, 2016. ACM.

[8] Clifford Nass, Youngme Moon, and Paul Carney. Are people polite to computers? responses to computer-based interviewing systems1. *Journal of Applied Social Psychology*, 29(5):1093–1109, 1999.

[9] Micah Singleton. Alexa can now set reminders for you. `https://www.theverge.com/circuitbreaker/2017/6/1/15724474/alexa-echo-amazon-reminders-named-timers`, 2017. [Online; last accessed 21-December-2018].

[10] Taylor Martin. 12 reasons to use Alexa in the kitchen. `https://www.cnet.com/how-to/how-to-use-alexa-in-the-kitchen/`, 2018. [Online; last accessed 17-December-2018].

[11] Heather Kelly. Apple's HomePod is coming. Here's what you need to know about smart speakers. `http://money.cnn.com/2017/06/08/technology/gadgets/apple-homepod-smart-speaker-faq/index.html`, 2017. [Online; last accessed 21-December-2018].

[12] Sam Wolfson. Amazon's Alexa recorded private conversation and sent it to random contact. `https://www.theguardian.com/technology/2018/may/24/amazon-alexa-recorded-conversation`, 2018. [Online; last accessed 17-December-2018].

[13] Matthew B. Hoy. Alexa, siri, cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, 37(1):81–88, 2018.

[14] Amazon. All-new Echo Show (2nd Gen). `https://www.amazon.com/All-new-Echo-Show-2nd-Gen/dp/B077SXWSRP`, 2019. [Online; last accessed 7-January-2019].

[15] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. "alexa is my new bff": Social roles, user satisfaction, and personification of the amazon echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, pages 2853–2859, New York, NY, USA, 2017. ACM.

[16] Nathaniel Fruchter and Ilaria Liccardi. Consumer attitudes towards privacy and security in home assistants. CHI EA '18, pages LBW050:1–LBW050:6, New York, NY, USA, 2018. ACM.

[17] Aarthi Easwara Moorthy and L Vu. Privacy concerns for use of voice activated personal assistant in the public space. *International Journal of Human Computer Interaction*, 31(4):307 to 335, April 2015.

[18] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):102:1–102:31, November 2018.

[19] Statista. Worldwide intelligent/digital assistant market share in 2017 and 2020, by product. `https://www.statista.com/statistics/789633/worldwide-digital-assistant-market-share/`, 2018. [Online; last accessed 21-December-2018].

[20] Ryen W. White. Skill discovery in virtual assistants. *Commun. ACM*, 61(11):106–113, October 2018.

[21] Amazon. The Alexa Skill Store for France is a Fast Growing Land of Opportunity. `https://developer.amazon.com/docs/ask-overviews/understanding-the-different-types-of-skills.html`, 2018. [Online; last accessed 29-December-2018].

[22] Google. Actions on Google. `https://developers.google.com/actions/samples/`, 2018. [Online; last accessed 29-December-2018].

[23] Bret Kinsella. The Alexa Skill Store for France is a Fast Growing Land of Opportunity. `https://voicebot.ai/2018/11/03/the-alexa-skill-store-for-france-is-a-fast-growing-land-of-opportunity/`, 2018. [Online; last accessed 22-December-2018].

[24] Ava Mutchler. Google assistant app total reaches nearly 2400. but that is not the real number. it is really 1719. `https://voicebot.ai/2018/01/24/google-assistant-app-total-reaches-nearly-2400-thats-not-real-number-really-1719/`, 2018. [Online; last accessed 22-December-2018].

[25] Chetan Naik, Arpit Gupta, Hancheng Ge, Mathias Lambert, and Ruhi Sarikaya. Contextual slot carryover for disparate schemas. In *Proc. Interspeech 2018*, pages 596–600, 2018.

[26] Google. Invocation and Discovery. `https://developers.google.com/actions/sdk/invocation-and-discovery`, 2018. [Online; last accessed 17-December-2018].

[27] J. Hirschberg and C. D. Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

[28] Zhu-yi WANG, Jian-po YANG, Yong-chao YIN, and Zhen-chao WANG. Echo cancellation based on blind source separation. *Journal of Computer Applications*, 32(10):2707–2710, 2013.

[29] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art, current trends and challenges. *CoRR*, abs/1708.05148, 2017.

[30] N. D. Londhe, M. K. Ahirwal, and P. Lodha. Machine learning paradigms for speech recognition of an indian dialect. *2016 International Conference on Communication and Signal Processing (ICCSP)*, 2016.

[31] András Zolnay, Daniil Kocharov, Ralf Schlüter, and Hermann Ney. Using multiple acoustic feature sets for speech recognition. *Speech Communication*, 49(6):514–525, 2007.

[32] Frederick Jelinek. *Statistical methods for speech recognition*. MIT Press, 1998.

[33] Lei Xinyu, Tu Guan Hua, Alex X.and Liu, Li Chi Yu, and Tian Xie. The insecurity of home digital voice assistants: Amazon alexa as a case study. 2017.

[34] Efthimios Alepis and Constantinos Patsakis. Monkey says, monkey does: Security and privacy on voice assistants. *IEEE Access*, 5:17841–17851, 2017.

[35] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security - CCS '17*, 2017.

[36] Nan Zhang, Xianghang Mi, Xuan Feng, XiaoFeng Wang, Yuan Tian, and Feng Qian. Understanding and mitigating the security risks of voice-controlled third-party skills on amazon alexa and google home. *CoRR*, abs/1805.01525, 2018.

[37] Helena Horton. Amazon Alexa recorded owner's conversation and sent to 'random' contact, couple complains. `https://www.telegraph.co.uk/news/2018/05/25/amazon-alexa-recorded-owners-conversation-sent-random-contact/`, 2018. [Online; last accessed 17-December-2018].

[38] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. Inaudible voice commands: The long-range attack and defense. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*, pages 547–560, Renton, WA, 2018. USENIX Association.

[39] Venessa Wong. Burger King's New Ad Will Hijack Your Google Home. `https://www.cnbc.com/2017/04/12/burger-kings-new-ad-will-hijack-your-google-home.html`, 2017. [Online; last accessed 25-December-2018].

[40] William Haack, Madeleine Severance, Michael Wallace, and Jeremy Wohlwend. Security analysis of amazon echo. 2017.

[41] Angelica Lai. Sneaky Kid Orders $350 Worth of Toys on Her Mom's Amazon Account. `https://mom.me/news/271144-sneaky-kid-orders-350-worth-toys-her-moms-amazon-account/`, 2018. [Online; last accessed 17-December-2018].

[42] Bret Kinsella. Amazon Introduces Skill Connections so Alexa Skills Can Work Together. `https://voicebot.ai/2018/10/04/amazon-introduces-skill-connections-so-alexa-skills-can-work-together/`, 2018. [Online; last accessed 24-December-2018].

[43] Human Right Watch. China: Voice biometric collection threatens privacy. 2017.

[44] D.J. Solove. A taxonomy of privacy. *University of Pennsylvania Law Review*, 154(3):477–560, 2006.

[45] Ayse Cufoglu. User profiling-a short review. *International Journal of Computer Applications (0975 8887)*, 108(3), 2014.

[46] Noah Apthorpe, Dillon Reisman, and Nick Feamster. A smart home is no castle: Privacy vulnerabilities of encrypted iot traffic. *CoRR*, abs/1705.06805, 2017.

[47] Sendong Zhao, Wu Yang, Ding Wang, and Wenzhen Qiu. A new scheme with secure cookie against sslstrip attack. In Fu Lee Wang, Jingsheng Lei, Zhiguo Gong, and Xiangfeng Luo, editors, *Web Information Systems and Mining*, pages 214–221, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[48] Atif M Memon and Ali Anwar. Colluding apps: Tomorrow's mobile malware threat. *IEEE Security & Privacy*, 13(6):77–81, 2015.

[49] Hyunji Chung and Sangjin Lee. Intelligent virtual assistant knows your life. *CoRR*, abs/1803.00466, 2018.

[50] eharmony. The future of dating report 2018: smart devices will predict if your relationship is on the rocks. `https://www.eharmony.co.uk/dating-advice/dating/the-future-of-dating-report-2018-smart-devices-to-predict-if-your-relationship-is-on-the-rocks?`, 2018. [Online; last accessed 21-December-2018].

[51] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. In *3rd IEEE European Symposium on Security and Privacy*, 2018.

[52] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[53] Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1605.07277, 2016.

[54] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. Cocaine noodles: Exploiting the gap between human and machine speech recognition. In *9th USENIX Workshop on Offensive Technologies (WOOT 15)*, Washington, D.C., 2015. USENIX Association.

[55] Chirag Modi, Dhiren Patel, Bhavesh Borisaniya, Avi Patel, and Muttukrishnan Rajarajan. A survey on security issues and solutions at different layers of cloud computing. *The Journal of Supercomputing*, 63(2):561–592, 2012.

[56] Nishtha Madaan, Mohd Abdul Ahad, and Sunil M. Sastry. Data integration in iot ecosystem: Information linkage as a privacy threat. *Computer Law and Security Review*, 34(1):125–133, 2018.

[57] Rodrigo Roman, Javier Lopez, and Stefanos Gritzalis. Evolution and trends in the security of the internet of things. *IEEE Computer*, 51:16–25, 07/2018 2018.

[58] E. Ronen, A. Shamir, A. Weingarten, and C. O Flynn. Iot goes nuclear: Creating a zigbee chain reaction. *IEEE Security Privacy*, 16(1):54 to 62, January 2018.

[59] Tamara Denning, Tadayoshi Kohno, and Henry M. Levy. Computer security and the modern home. *Commun. ACM*, 56(1):94–103, January 2013.

[60] Guillermo Suarez-Tangil, Juan E Tapiador, Pedro Peris-Lopez, and Arturo Ribagorda. Evolution, detection and analysis of malware in smart devices. *IEEE Communications Surveys & Tutorials*, 16(2):961–987, 2014.

[61] E. Fernandes, J. Jung, and A. Prakash. Security analysis of emerging smart home applications. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 636–654, May 2016.

[62] Chouhan Priyanka and Singh Rajendra. Security attacks on cloud computing with possible solution. 6(1), January 2016.

[63] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. Skill squatting attacks on amazon alexa. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 33–47, Baltimore, MD, 2018. USENIX Association.

[64] Yuan Gong and Christian Poellabauer. Crafting adversarial examples for speech paralinguistics applications. *CoRR*, abs/1711.03280, 2017.

[65] Lea Schonherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *CoRR*, abs/1808.05665, 2018.

[66] Nicholas Carlini and David A. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pages 1–7, 2018.

[67] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 513–530, Austin, TX, 2016. USENIX Association.

[68] Texas Instruments. AN-1973 Benefits and Challenges of High-Frequency Regulators. `http://www.ti.com/lit/an/snva399a/snva399a.pdf`, 2013. [Online; last accessed 17-December-2018].

[69] Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56–66, 2018.

[70] Hyunji Chung, Jungheum Park, and Sangjin Lee. Digital forensic approaches for amazon alexa ecosystem. *Digital Investigation*, 22:S15 to S25, 2017.

[71] Google. Set up multiple users for your speaker or smart display. `https://support.google.com/assistant/answer/9071681`, 2017. [Online; last accessed 20-February-2018].

[72] Amazon. About Alexa Voice Profiles. `https://www.amazon.com/gp/help/customer/display.html?nodeId=202199440`, 2017. [Online; last accessed 20-February-2019].

[73] Si Chen, Kui Ren, Sixu Piao, Cong Wang, Qian Wang, Jian Weng, Lu Su, and Aziz Mohaisen. You can hear but you cannot steal:defending against voice impersonation attacks on smartphones. *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, 2017.

[74] Huan Feng, Kassem Fawaz, and Kang G. Shin. Continuous authentication for voice assistants. *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking - MobiCom '17*, 2017.

[75] Veton Kepuska and Gamal Bohouta. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 99–103, 2018.

[76] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim Shchemelinin. Audio-replay attack detection countermeasures. *CoRR*, abs/1705.08858, 2017.

[77] J. Liu, C. Zhang, and Y. Fang. Epic: A differential privacy framework to defend smart homes against internet traffic analysis. *IEEE Internet of Things Journal*, 5(2):1206–1217, April 2018.

[78] Alice Coucke, Alaa Saade, Adrien Ball, Theodore Bluche, Alexandre Caulier, David Leroy, Clement Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Mael Primet, and Joseph Dureau. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190, 2018.

[79] Massimiliano Todisco, Hctor Delgado, and Nicholas Evans. Constant q cepstral coefficients. *Comput. Speech Lang.*, 45(C):516–535, September 2017.

[80] Gildas Avoine, Muhammed Ali Bingol, Ioana Boureanu, Srdjan capkun, Gerhard Hancke, Suleyman Kardas, Chong Hee Kim, Cedric Lauradoux, Benjamin Martin, Jorge Munilla, Alberto Peinado, Kasper Bonne Rasmussen, Dave Singelee, Aslan Tchamkerten, Rolando Trujillo-Rasua, and Serge Vaudenay. Security of distance-bounding: A survey. *ACM Comput. Surv.*, 51(5):94:1–94:33, September 2018.

[81] Xueou Wang, Xiaolu Hou, Ruben Rios, Per Hallgren, Nils Ole Tippenhauer, and Martin Ochoa. Location proximity attacks against mobile targets: Analytical bounds and attacker strategies. In *23rd European Symposium on Research in Computer Security (ESORICS 2018)*, volume 11099 of *LNCS*, pages 373–392, Barcelona, 2018. Springer, Springer.

[82] Weijia He, Maximilian Golla, Roshni Padhi, Jordan Ofek, Markus Durmuth, Earlence Fernandes, and Blase Ur. Rethinking access control and authentication for the home internet of things (iot). In *Proceedings of the 27th USENIX Conference on Security Symposium*, SEC'18, pages 255–272, Berkeley, CA, USA, 2018. USENIX Association.

[83] Xudong Ni, Zhimin Yang, Xiaole Bai, A. C. Champion, and D. Xuan. Diffuser: Differentiated user access control on smartphones. In *2009 IEEE 6th International Conference on Mobile Adhoc and Sensor Systems*, pages 1012–1017, Oct 2009.

[84] Tara Matthews, Kerwell Liao, Anna Turner, Marianne Berkovich, Robert Reeder, and Sunny Consolvo. She'll just grab any device that's closer: A study of everyday device & account sharing in households. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5921–5932. ACM, 2016.

[85] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. A stalker's paradise: How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 667. ACM, 2018.

[86] Tara Matthews, Kathleen OLeary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F Churchill, and Sunny Consolvo. Security and privacy experiences and practices of survivors of intimate partner abuse. *IEEE Security & Privacy*, (5):76–81.

[87] Charith Perera, Susan YL Wakenshaw, Tim Baarslag, Hamed Haddadi, Arosha K Bandara, Richard Mortier, Andy Crabtree, Irene CL Ng, Derek McAuley, and Jon Crowcroft. Valorising the iot databox: creating value for everyone. *Transactions on Emerging Telecommunications Technologies*, 28(1):e3125, 2016.

[88] J. M. Such and N. Criado. Multiparty privacy in social media. *Communications of the ACM*, 61(8):74–81, 2018.

[89] J. M. Such and N. Criado. Resolving multi-party privacy conflicts in social media. *IEEE TKDE*, 28(7):1851–1863, 2016.

[90] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *CoRR*, abs/1608.04644, 2016.

[91] Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6980–6990, 2017.

[92] Rodrigo Roman, Ruben Rios, Jose A. Onieva, and Javier Lopez. Immune system for the internet of things using edge technologies. *IEEE Internet of Things Journal*, In Press.

[93] Jose M Such, Natalia Criado, Laurent Vercouter, and Martin Rehak. Intelligent cybersecurity agents. *IEEE Intelligent Systems*, 31(5):3–7, 2016.

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

[95] Emilio Corchado and Álvaro Herrero. Neural visualization of network traffic data for intrusion detection. *Applied Soft Computing*, 11(2):2042–2056, 2011.

[96] Katarzyna Olejnik, Italo Dacosta, Joana Soares Machado, Kevin Huguenin, Mohammad Emtiyaz Khan, and Jean-Pierre Hubaux. Smarper: Context-aware and automatic runtime-permissions for mobile devices. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 1058–1076. IEEE, 2017.

[97] Gaurav Misra and Jose M Such. Pacman: Personal agent for access control in social media. *IEEE Internet Computing*, 21(6):18–26, 2017.

[98] Natalia Criado, Estefania Argente, and V Botti. Open issues for normative multi-agent systems. *AI communications*, 24(3):233–264, 2011.

[99] Helen Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79:119, 2004.

[100] Özgür Kafali, Nirav Ajmeri, and Munindar P Singh. Revani: Revising and verifying normative specifications for privacy. *IEEE Intelligent Systems*, 31(5):8–15, 2016.

[101] Natalia Criado and Jose M Such. Implicit contextual integrity in online social networks. *Information Sciences*, 325:48–69, 2015.

[102] R. Fogues, P. K. Murukannaiah, J. M. Such, and M. P. Singh. Sharing policies in multiuser privacy scenarios: Incorporating context, preferences, and arguments in decision making. *ACM TOCHI*, 24(1):5, 2017.

[103] T. Baarslag, A. T. Alan, R. C. Gomer, I. Liccardi, H. Marreiros, E. Gerding, et al. Negotiation as an interaction mechanism for deciding app permissions. In *Proc. of CHI Extended Abstracts*, pages 2012–2019, 2016.

[104] J. M. Such and M. Rovatsos. Privacy policy negotiation in social media. *ACM Trans. on Autonomous and Adaptive Systems*, 11(1):4, 2016.

[105] I. Pinyol and J. Sabater-Mir. Computational trust and reputation models for open multi-agent systems: a review. *Artif Intell Rev*, 40(1):1–25, 2013.

[106] Marco Prandini and Marco Ramilli. Towards a practical and effective security testing methodology. In *Computers and Communications (ISCC), 2010 IEEE Symposium on*, pages 320–325. IEEE, 2010.

[107] Jose M. Such, Antonios Gouglidis, William Knowles, Misra Gaurav, and Rashid Awais. Information assurance techniques: Perceived cost effectiveness. *Computers and Security*, 60:117–133, 2016.

[108] Sara Abbaspour Asadollah, Rafia Inam, and Hans Hansson. A survey on testing for cyber physical system. In *IFIP International Conference on Testing Software and Systems*, pages 194–207. Springer, 2015.

[109] William Knowles, Jose M Such, Antonios Gouglidis, Gaurav Misra, and Awais Rashid. Assurance techniques for industrial control systems (ics). In *Proceedings of the First ACM Workshop on Cyber-Physical Systems-Security and/or PrivaCy*, pages 101–112. ACM, 2015.

[110] Pierre Ciholas and Jose M Such. Composite vulnerabilities in cyber physical systems. *Security and Resilience of Cyber–Physical Infrastructures*, page 4, 2016.

[111] David Wright and Paul De Hert. Introduction to privacy impact assessment. In *Privacy Impact Assessment*, pages 3–32. Springer, 2012.

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

[113] Gaurav Misra, Jose M Such, and Lauren Gill. A privacy assessment of social media aggregators. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 561–568. ACM, 2017.

[114] Pauline Anthonysamy, Matthew Edwards, Chris Weichel, and Awais Rashid. Inferring semantic mapping between policies and code: the clue is in the language. In *International Symposium on Engineering Secure Software and Systems*, pages 233–250. Springer, 2016.

[115] Özgür Kafali, Jasmine Jones, Megan Petruso, Laurie Williams, and Munindar P Singh. How good is a security policy against real breaches? a hipaa case study. In *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, pages 530–540. IEEE, 2017.

[116] Eric Zeng, Shrirang Mare, and Franziska Roesner. End user security and privacy concerns with smart homes. pages 65–80, 2017.

[117] Bill Brenner. Know the risks of Amazon Alexa and Google Home. `https://nakedsecurity.sophos.com/2017/01/27/data-privacy-day-know-the-risks-of-amazon-alexa-and-google-home/`, 2017. [Online; last accessed 29-December-2018].

[118] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. User perceptions of smart home iot privacy. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):200:1–200:20, November 2018.

[119] F. Schaub, R. Balebako, and L. F. Cranor. Designing effective privacy notices and controls. *IEEE Internet Computing*, 21(3):70–77, 2017.

[120] Jonathan R Mayer and John C Mitchell. Third-party web tracking: Policy and technology. In *2012 IEEE Symposium on Security and Privacy*, pages 413–427. IEEE, 2012.

[121] Charles Wright, Scott Coull, and Fabian Monrose. Traffic morphing: An efficient defense against statistical traffic analysis. In *Proceedings of the Network and Distributed Security Symposium*. IEEE, February 2009.

[122] K. P. Dyer, S. E. Coull, T. Ristenpart, and T. Shrimpton. Peek-a-boo, i still see you: Why efficient traffic analysis countermeasures fail. In *2012 IEEE Symposium on Security and Privacy*, pages 332–346, May 2012.

[123] Paul Syverson. Sleeping dogs lie on a bed of onions but wake when mixed. In *Proceedings of HotPETS 2011*, 2011.

[124] Steven J. Murdoch and Piotr Zielinski. Sampled traffic analysis by internet-exchange-level adversaries. In *Proceedings of the 7th International Conference on Privacy Enhancing Technologies*, PET'07, pages 167–183, Berlin, Heidelberg, 2007. Springer-Verlag.

[125] Paul Syverson. Why i'm not an entropist. In *In the Proceedings of Security Protocols XVII: 17th International Workshop*, 2009.

[126] George Kadianakis, Claudia V. Roberts, Laura M. Roberts, and Philipp Winter. Anomalous keys in tor relays. *CoRR*, abs/1704.00792, 2017.