TU WIEN Informatics

Towards Secure and Usable Authentication for Voice-Controlled Smart Home Assistants

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Software Engineering & Internet Computing

eingereicht von

Alexander Ponticello
Matrikelnummer 01226441

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Thomas Grechenig
Mitwirkung: Katharina Krombholz
Florian Fankhauser

Wien, 14. Oktober 2020

Unterschrift Verfasser

Unterschrift Betreuung

TU WIEN Informatics

Towards Secure and Usable Authentication for Voice-Controlled Smart Home Assistants

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Software Engineering & Internet Computing

by

Alexander Ponticello
Registration Number 01226441

to the Faculty of Informatics

at the TU Wien

Advisor: Thomas Grechenig
Assistance: Katharina Krombholz
Florian Fankhauser

Vienna, 14th October, 2020

Signature Author

Signature Advisor

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Towards Secure and Usable Authentication for Voice-Controlled Smart Home Assistants

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Software Engineering & Internet Computing

eingereicht von

Alexander Ponticello
Matrikelnummer 01226441

ausgeführt am
Institut für Information Systems Engineering
Forschungsbereich Business Informatics
Forschungsgruppe Industrielle Software
der Fakultät für Informatik der Technischen Universität Wien

Betreuung: Thomas Grechenig

Wien, 14. Oktober 2020

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Alexander Ponticello

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 14. Oktober 2020

Alexander Ponticello

Danksagung

Ich bedanke mich bei meinem Betreuer Thomas Grechenig, der es mir ermöglicht hat an diesem Thema zu arbeiten. Dank ihm hatte ich während meinem Studium eine Vielzahl an Gelegenheiten mich weiterzubilden, fachlich wie auch beruflich. Ein großes Dankeschön dafür.

Ein herzliches Dankeschön geht an Katharina Krombholz für die Betreuung während meines Forschungspraktikums und darüber hinaus beim gesamten Projekt dieser Arbeit. Ihre unkomplizierte und herzliche Art hat mir das Arbeiten nicht nur erleichtert, sondern war auch eine Motivation in schweren Momenten. Danke Katharina, dass du mir erklärt hast wie Wissenschaft funktioniert und dafür, dass du in mir das Feuer für sie entfacht hast.

Ein weiterer Dank geht an Florian Fankhauser für das wertvolle Feedback und die Unterstützung, nicht nur im Rahmen der Masterarbeit, sondern während meines gesamten Masterstudiums. Vielen Dank für den Einsatz, den du ständig zeigst. Außerdem habe ich mich dank deiner Vorlesungen überhaupt erst für Security und später Usable Security interessiert. Ein großer Glücksfall, wie ich meine. Auch den übrigen Kollegen der Forschungsgruppe ESSE spreche ich auf diesem Wege meinen Dank aus für die Mithilfe und Kameradschaft.

Ein großer Dank geht außerdem an meine Kollegen in der Usable Security Gruppe, besonders Simon Anell für die Hilfe beim Transkribieren. Nicht zuletzt möchte ich mich bei Matthias Fassl bedanken. Ohne deine Hilfe wäre die Arbeit wohl nie fertig geworden. Danke für deine ständige Verfügbarkeit für alle möglichen Fragen und danke für deine Freundschaft während dieser Zeit, sowie hoffentlich auch in Zukunft.

Ich drücke weiters meine innige Dankbarkeit aus gegenüber meinen Eltern Ulli und Rudi und meinen Großeltern, für die ständige Unterstützung, durch mein gesamtes Studium hindurch, sei es in finanzieller, aber vor allem in seelischer Natur. Es tut gut einen Ort zu haben, wohin ich zurückkehren kann und wo ich bedingungslosen Beistand erfahre.

Zu guter Letzt, von Herzen Danke an Simone für den Rückhalt den ich ständig erfahren darf und für die Ermutigung, die oft genug nötig war.

Kurzfassung

Smart-Home-Assistenten wie Amazon Alexa und Google Home werden derzeit üblicherweise für nicht-vertrauliche oder nicht-sicherheitsrelevante Tätigkeiten wie Wetterberichtabfragen oder die Steuerung verknüpfter Geräte verwendet.

In ausgewählten Märkten sind jedoch bereits sicherheitskritische Anwendungsfälle wie Online-Banking und sprachgesteuerte Türschlösser verfügbar, welche mit zunehmender Beliebtheit in Zukunft weitere Verbreitung finden werden. Dabei sind grundlegende Sicherheitsfunktionen erforderlich, um das Vertrauen der AnwenderInnen zu stärken.

Derzeit besteht der Stand der Technik der Authentifizierung für Smart-Home-Assistenten aus unsicheren Sprachcodes, welche von BenutzerInnen laut ausgesprochen werden müssen, sowie biometrischer Authentifizierung mittels Stimmerkennung. Für diese Authentifizierungsmethoden wurde in der Vergangenheit eine Reihe von Angriffen demonstriert. Frühere Arbeiten legen es außerdem nahe, dass die Übernahme neuer, sicherheitskritischer Funktionen von fehlendem Vertrauen der BenutzerInnen in das System behindert werden könnte. Deshalb untersuchen wir den Design-Space für zukünftige Authentifizierungsmechanismen. Diese haben das Potenzial stärkere Sicherheitseigenschaften aufgrund gesteigerter Benutzerfreundlichkeit zu erzielen, sowie den BenutzerInnen ein gestärktes Sicherheitsgefühl während der Ausführung von Anwendungen geben zu können.

Wir führen semi-strukturierte Interviews mit BenutzerInnen von Amazon Alexa durch, in welchen vier sicherheitskritische Szenarien behandelt werden. In jedem Szenario befragen wir die TeilnehmerInnen nach ihrer Wahrnehmung bezüglich Bedrohungen, Schutzmaßnahmen sowie für ihr Sicherheitsempfinden wesentliche Designaspekte.

Unsere Erkenntnisse, welche wir aus einer thematischen Analyse der Daten gewonnen haben, zeigen u. a., dass BenutzerInnen (1) primär besorgt sind über umstehende Zuhörer, (2) wenig Vertrauen in ausgesprochene Codes haben und stattdessen vertrauenswürdige Stimmerkennung bevorzugen würden, und (3) kontextabhängige Anforderungen an Authentifizierungssysteme haben, speziell in Abhängigkeit von Ort und Anwesenheit anderer Personen. Die von uns bereitgestellten Designempfehlungen können die Basis für hochentwickelte zukünftige Authentifizierungsmechanismen bilden.

Schlüsselwörter *Smart Home Assistenten, Authentifizierung, Benutzerwahrnehmung, Bedrohungen, Schutzmaßnahmen, Designaspekte, Qualitative Methoden*

Abstract

Smart home assistants such as Amazon Alexa and Google Home are currently usually used for non-confidential or non-security-sensitive activities such as checking the weather report or for controlling linked devices.

In selected markets, security-critical applications such as online banking and voice-controlled door locks are already available and will become more widespread in the future as their popularity increases. Basic security functions are required to strengthen users' confidence.

At the moment, state-of-the-art authentication for smart home assistants consists of insecure voice codes that users have to utter out loud, as well as biometric authentication using voice recognition. A number of attacks have been demonstrated for these authentication methods in the past. Previous work also indicates that the adoption of new, security-critical functions could be hindered by a lack of user confidence in the system. We are, therefore, examining the design space for future authentication mechanisms. These have the potential to achieve stronger security properties due to increased usability and give users a stronger feeling of security during the execution of applications.

We carry out semi-structured interviews with Amazon Alexa users, in which four security-critical scenarios are dealt with. In each scenario, we ask the participants about their perception of threats, mitigation strategies, and design aspects essential for their perception of security.

Our findings, which we have gained from a thematic analysis of the data, show amongst others that users (1) are primarily concerned about bystanders, (2) have little trust in spoken codes and would instead prefer trustworthy voice recognition, and (3) have context-dependent requirements for authentication systems, especially depending on the location and presence of other people. The design recommendations we provide can form the basis for highly developed future authentication mechanisms.

Keywords *Smart Home Assistants, Authentication, User Perceptions, Threats, Mitigation Strategies, Design Aspects, Qualitative Methods*

Contents

Kurzfassung **xi**

Abstract **xiii**

Contents **xv**

1 Introduction **1**
1.1 Problem Description and Motivation 2
1.2 Aim of the Work 2
1.3 Structure of the Thesis 3

2 Background **5**
2.1 Voice-Controlled Smart Home Assistants 5
2.2 Authentication 8
2.3 Usability and Security 12

3 Related Work **15**
3.1 Exploring User Mental Models 16
3.2 Security and Privacy of Smart Home Assistants 19
3.3 Authentication Schemes for Smart Home Assistants 20
3.4 Attacks on Smart Home Assistants 23

4 Attacks on Voice-Controlled Authentication Systems **27**
4.1 Examples of Attacks 27
4.1.1 Faking a User's Voice 28
4.1.2 Hidden and Inaudible Command Injection 30
4.1.3 Malicious Skills 34
4.2 Executing Selected Attacks 34
4.2.1 Experimental Setup for Execution of Attacks 35
4.2.2 Selected Attacks to be Executed 36
4.2.3 Results from Executing Selected Attacks 37
4.3 Implications for Smart Home Assistant Authentication 37

5 Conducting a User Study to Investigate Users' Perceptions of Smart Home Assistant Authentication **41**
5.1 Study Design . . . 42
5.1.1 Scenarios and Vignettes . . . 42
5.1.2 Expectations for User Mental Models . . . 43
5.1.3 Scenario Design . . . 44
5.1.4 Interview Guideline . . . 50
5.1.5 Interview Procedure . . . 51
5.1.6 Pilot Testing . . . 53
5.2 Recruitment . . . 54
5.3 Data Analysis . . . 56
5.4 Ethical Considerations . . . 57

6 Exploring the Design Space of Secure and Usable Smart Home Assistant Authentication **59**
6.1 Perception of Threats . . . 60
6.1.1 Amazon as a Threat . . . 60
6.1.2 Bystanders as Threat . . . 62
6.1.3 Insiders as Threat . . . 63
6.1.4 Pranks as Threat . . . 64
6.1.5 Criminals as Threat . . . 64
6.1.6 Cyberattacks as Threat . . . 65
6.1.7 Malicious Skills as Threat . . . 66
6.1.8 Accidents as Threat . . . 67
6.1.9 Summary . . . 68
6.2 Mitigation Strategies . . . 69
6.2.1 Refrain from Using the System due to Security Reasons . . . 69
6.2.2 Move to Another Room to Use Alexa . . . 70
6.2.3 Voice Interaction Inappropriate in Specific Social Situations . . . 71
6.2.4 Take Time for Important Actions . . . 72
6.2.5 Change Voice Code Regularly . . . 72
6.2.6 Whispering the Voice Code Protects Against Eavesdropping . . . 73
6.2.7 Users Notice Acoustic Attacks on Their Alexa if They Are Present 73
6.2.8 Building Up Trust by Trial-and-Error of the Security Mechanism 74
6.2.9 Voice Code Protects Against Unauthorized Access . . . 74
6.2.10 Summary . . . 75
6.3 Important Properties of Secure and Usable Smart Home Assistant Authentication Systems . . . 76
6.3.1 Building Trust . . . 76
6.3.2 Transparency and Agency . . . 78
6.3.3 Risk Assessment of Authentication . . . 79
6.3.4 Perception of Authentication Methods . . . 80

7 Discussion **85**

7.1 Relating Actual and Perceived Threats . . . 85
7.1.1 Eavesdropping . . . 85
7.1.2 Insiders as Threat Actors . . . 86
7.1.3 Security Breaches Due to Misinterpretations . . . 86
7.1.4 Attacks on Voice Recognition . . . 86
7.1.5 Injecting Commands without the User Noticing . . . 87
7.1.6 Cyberattacks . . . 88
7.1.7 Non-Targeted Attacks . . . 88
7.2 Design Implications . . . 88
7.2.1 Rely on Voice Recognition as an Intuitive and Trustworthy Authentication Method . . . 88
7.2.2 Include Demo Mode to Let Users Experience the Effectiveness of the Authentication Method . . . 89
7.2.3 Provide Unobtrusive Authentication for Social Situations . . . 89
7.2.4 Maintain Low-Effort Interaction . . . 90
7.2.5 Keep the Authentication Process Transparent . . . 90
7.2.6 Account for Varying Requirements . . . 90
7.3 Limitations . . . 91

8 Conclusion and Future Work **93**
8.1 Conclusion . . . 93
8.2 Future Work . . . 94

Glossary **101**

Acronyms **103**

Bibliography **105**
Online References . . . 113

A Appendix **117**
A.1 Interview Guideline . . . 117
A.2 Codebook . . . 119
A.3 Participant Drawings . . . 122

CHAPTER 1

Introduction

Voice-controlled assistants are one of the main entry-points for end-users into the Internet of Things (IoT)-driven smart home technology. Popular examples include Amazon Alexa and Google Home. Such devices allow for a broad spectrum of use cases. Some of the most popular ones are related to entertainment (e.g., playing music, games), information gathering (e.g., weather, cooking recipes), or personal planning (e.g., calendar, task list). Furthermore, it is possible to control other IoT devices, such as smart light bulbs or heating. Smart home assistants provide the possibility for "hands-free" interaction. In contrast to traditional user interfaces, such as personal computers or smartphones, smart home assistants generally do not have any tangible input in the form of a keyboard or a touch screen but rely on the audio channel for user input and output generated by the device. This input method enables users with certain physical limitations to interact with and use the system. Also, younger children not yet capable of reading can interact with a smart home assistant through language commands, which are arguably one of the most natural forms of human communication.

In that regard, it differs from human-computer interaction via touchable devices, which has been studied for some time now. Another distinction is that smart home assistants, as studied here, in a smart home environment, are commonly multi-user devices. In such a scenario, there might be several users interacting with the same system regularly, possibly also at the same time, reacting to one another's actions. Besides, there might be several infrequent users using the system on specific occasions. This usage pattern is enabled by the open nature of the voice input channel and the setting of most smart homes, only requiring physical presence to interact with the system. This characteristic stands in sharp contrast with smartphones or personal computers as they have been projected as single-user devices and mostly studied under that perspective.

1.1 Problem Description and Motivation

As smart home assistants get adopted by more and more end-users, vendors work towards new use cases. Online shopping has already been deployed to several different devices, allowing users to order goods using only their voice, without the need to interrupt their current activity and, e.g., sit down in front of a screen. Compatible locking systems permit users to open doors via voice commands[Aug]. Capital One, one of the first banks coming to the smart home assistant platform, enables customers to retrieve account information, such as their current balance, or pay credit card bills via Amazon Alexa[Cap]. These use cases are highly critical from a security and privacy point of view and, therefore, call for more robust protection mechanisms.

Authentication is one mechanism users get in touch with frequently (e.g., on every log-in). It enables a variety of further security measures to preserve confidentiality and integrity, with access control being a prominent example. Although authentication for smart home assistants is still in an early stage of development, several attacks already emerged for current state-of-the-art authentication mechanisms. Hence, there is a need for improvements in existing and potential new authentication systems. He et al. [He+18] discuss the importance of access control mechanisms as well as authentication for smart home scenarios. Current devices do not implement sufficient user authentication, including error-prone voice-based biometrics, as noted by Kwak et al. [Kwa+19]. Headlines were made by television (TV) commercials resulting in smart home assistants trying to order goods[Lip17] and attackers injecting commands via a tilted window[Til16]. Sugawara et al. [Sug+20] presented an attack using lasers to inject voice commands into smart speakers over a considerable distance.

Several papers discuss the special needs and challenges to be taken into account in a smart home setting. Well studied systems like smartphones are conceived as single-user devices, in contrast to IoT and smart home scenarios. Blue et al. [Blu+18] and Feng, Fawaz, and Shin [FFS17] have proposed authentication schemes for voice-controlled smart home assistants. However, little research has been conducted regarding such systems' design space in terms of users' attacker and threat models. Several papers point out the importance of understanding user mental models in order to design secure, usable systems, e.g., Krombholz et al. [Kro+19] and Kang et al. [Kan+15].

1.2 Aim of the Work

This work's aim is to close the gap in the literature, which has been discussed in the previous section, and provide valuable insights for the design of authentication schemes for smart home assistants. To understand the design space, we explore users' perception of threats and mitigation strategies in this context. We focus on e-banking applications and physical access control as examples of high-risk tasks relevant to many end-users. The threat model we consider in this work includes motivated attackers, possibly acquaintances or children and outside attackers, capable of eavesdropping on a user's regular interactions

with the smart home device, and injecting voice commands over the audio channel without the user noticing.

As the first outcome of this work, an overview of current threats and possible attacks on smart home assistants relevant under the described attacker model is given and discussed from a technical point of view. Such an overview can help to motivate a need for more sophisticated and better-adapted authentication mechanisms for smart home assistants. It is not in the scope of this thesis to discuss technical threats on smart home assistants on a firmware level, e.g., attacks on the network stack, nor physical attacks targeting hardware such as circuit modifications.

Furthermore, the goal of this work is to gain a better understanding of the design space of authentication for smart home assistants. Little research has been done, and several key aspects, such as attacker models of users, are yet to be described. Therefore, a significant result of this work is to explore users' perceptions of threats and protective strategies employed by them. Based on the gained insights, we offer design recommendations for future authentication methods to help build systems with enhanced security and usability.

In this work, we want to contribute to current research by providing answers to the following research questions (RQs):

RQ1 What are technical threats for current authentication methods considering the voice input channel?

RQ2 Which kind of attackers and threats are users concerned about when performing high-risk tasks via voice-controlled assistants in a smart home environment?

RQ3 Which potential mitigation strategies do users apply to protect themselves?

RQ4 Which properties does an authentication system for voice assistants need in order to be perceived as secure by users?

1.3 Structure of the Thesis

This work is structured as follows: **Chapter 2 (Background)** describes the background relevant to the design of secure and usable authentication schemes for smart home assistants.

In **Chapter 3 (Related Work)** we highlight previous work conducted in the domain of smart home assistant security and privacy. We describe how users' perceptions have been studied in the past and insights gained by different authors. Furthermore, this chapter includes several proposed authentication schemes for voice-controlled systems and state-of-the-art attacks on smart home assistants with a focus on authentication.

In **Chapter 4 (Attacks on Voice-Controlled Authentication Systems)**, we answer RQ1 by giving a detailed description of real-world attacks relevant for the design of

authentication systems. We describe how we executed selected attacks on the Amazon Alexa platform and report the results. Finally, we illustrate implications of presented attacks for smart home assistant authentication.

Chapter 5 (Conducting a User Study to Investigate Users' Perceptions of Smart Home Assistant Authentication) gives a detailed report about the user study we conducted to answer our research questions. We describe our study design based on previous work from Section 3.1 and justify relevant design decisions taken in the process. This chapter also includes details about recruitment and our qualitative approach to analyse the collected data, while finally stating some ethical considerations relevant in this context.

Chapter 6 (Exploring the Design Space of Secure and Usable Smart Home Assistant Authentication) presents the results of the user study outlined in Chapter 5. We report our findings classified by our research questions: Section 6.1 answers RQ2 by describing the perceptions of threats our study's participants had. RQ3 is addressed by Section 6.2, where we report mitigation strategies described by users. Finally, results relevant for RQ4 can be found in Section 6.3.

We discuss our findings from Chapter 6 in **Chapter 7 (Discussion)**. First, we relate the attacks from Chapter 4 to threats we found users to be concerned about, as highlighted in Section 6.1. Next, we derive some design implications for future smart home assistant authentication schemes, providing answers to RQ4. Finally, this chapter includes some limitations of this work.

Chapter 8 (Conclusion and Future Work) summarizes the work and draws a conclusion before wrapping up this thesis by proposing interesting topics for future work built upon our findings.

CHAPTER 2

Background

We present some of the necessary background of our work. In order to investigate users' perceptions regarding authentication of smart home assistants, we first describe voice user interfaces and how they are used in a smart home environment. Next, we give a detailed overview of authentication processes and different methods used to match a subject to a system-internal entity. Finally, we give some insights into why usability and security need to be studied conjointly in the context of authentication. We explain why it is crucial for the design and development process of security- and privacy-related systems to take its users into account, as their decisions, if based on wrong perceptions or incomplete mental models, can compromise even a theoretically secure system.

2.1 Voice-Controlled Smart Home Assistants

The term *smart home assistants* describes a software, usually coupled with a physical device such as a smart speaker or a smartphone, that is designed to aid the users with every-day tasks. Two of the most prominent examples are Amazon Alexa [Amab] and Google Assistant[Gooa]. Figure 2.1 depicts some examples of smart speaker hardware used to run smart home assistants. Typical use cases for smart home assistants include retrieving information from online service (e.g., weather forecast, news), entertainment (e.g., play games, replay audiobooks), and controlling other IoT devices, as described by Lei et al. [Lei+18]. In the latter role, smart home assistants can be a primary entry point for users into the smart home. The authors state that smart home assistants are designed for hassle-free and convenient interaction. As such, these systems use voice as the main input and output mechanisms. Hence, hardware running smart home assistant software is usually equipped with a microphone and a speaker. Since smart home assistants are designed to be used without physical interaction (i.e., *hands-free*), no display or keyboard is required.

Figure 2.1: Three examples of smart speakers hosting smart home assistants: (Left) Google Home running Google Assistant [Gooc] (Middle) Amazon Echo Dot running Alexa [Amaf] (Right) Amazon Echo running Alexa [Amaf]

Speaking is arguably one of the most natural forms of human conversation. Smart home assistants rely on conversational agents, as described by McTear, Callejas, and Griol [MCG18], to hold a conversation with the users, allowing them to ask questions and give orders, similar to a personal assistant or a butler. As Porcheron et al. [Por+18] describe, this concept was first brought up by science fiction works and has since made its way into the every-day life of millions of people. Conversational agents employ a combination of machine learning and natural language processing to decode utterances first into text, from which actionable commands or questions get extracted. After executing subroutines according to the intended instructions, the output is translated from text to speech again and uttered back to the user. Extensive research has been conducted into making synthetic voices sound natural, as McTear, Callejas, and Griol [MCG18] further explain.

Smart home assistants have a complex architecture. The device located at the user's home acts as a mere entry-point into the system. Most processing is not performed locally but on back end servers, i.e., the cloud, as Abdi, Ramokapane, and Such [ARS19] describe. Hence, a smart home assistant usually needs to be connected to the Internet to function properly. Smart home assistants continuously record audio and analyze it in order to detect an activation command referred to as *wake word*, see Schönherr et al. [Sch+20]. For Amazon Alexa, this is typically "Alexa"[Amab], while Google Assistant uses "Hey Google"[Gooa]. After this activation word is processed locally, a connection to the cloud is established. All further conversation is sent to this back end for decoding and interpretation. Most smart home platforms allow third-party developers to deploy skills in addition to built-in features. Users can install those programs from a central market place. If a skill is invoked, the back end server will still perform the speech-to-text (STT) conversion before handing the control flow over to the third-party software. The possible output is sent back to the cloud where a text-to-speech (TTS) engine prepares

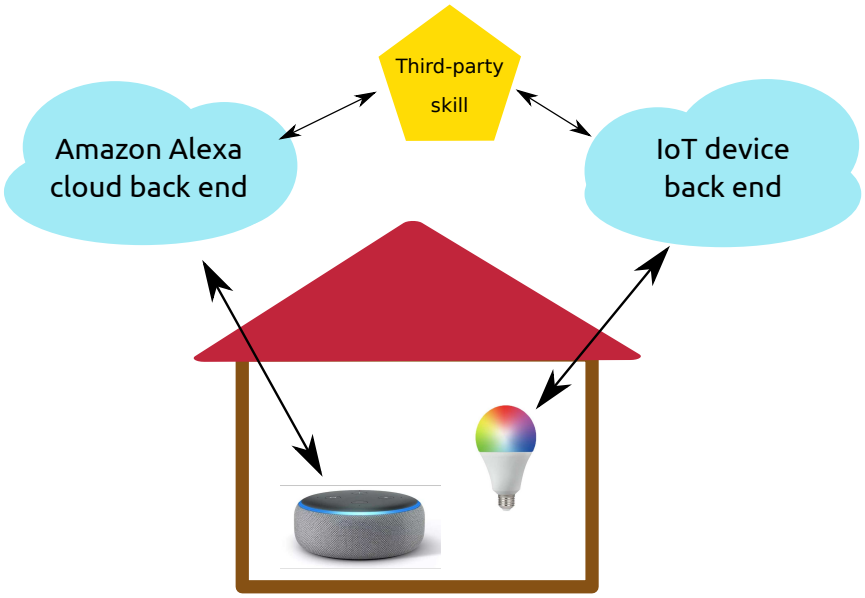

Figure 2.2: A high-level data flow overview of a smart home assistant (Amazon Alexa) controlling an IoT device (smart light bulb) via a third-party skill. Commands received by the Alexa device are processed by the Amazon cloud back end, which invokes the third-party skill. The skill sends a request to the IoT device's back end, which in turn sends instructions to the light bulb. (Inspired by Abdi, Ramokapane, and Such [ARS19])

the audio files replayed to the user by the speaker. Skills can also be used to control IoT devices in a similar way, with the skill software sending commands via its back end to the device. Figure 2.2 depicts how the ecosystem of a smart home assistant might look like in practice.

Smart speakers are typically not equipped with a power button and are by default always turned on if connected to a power source. The device also needs to actively listen to its environment continuously to detect a wake word, as Lau, Zimmerman, and Schaub [LZS18] further explain. This always-on mechanic has caused significant privacy concerns among users, as the authors report. Smart home assistants have the potential to monitor every conversation happening within audio reach. Although vendors claim that no data is sent to the cloud unless the wake word is received, users are sceptical about this. Privacy concerns were also the main reason non-users gave when asked why they did not adopt this technology. Most smart home assistants ship with a non-listening mode were the device stays turned on while the microphone gets disconnected, impeding the activation of the device. Amazon Alexa uses a red circle around the top of their smart speakers to indicate that the device is neither listening nor giving feedback to the user[Amac].

2.2 Authentication

According to Bishop, Sullivan, and Ruppel [BSR19], the three pillars of computer security are *confidentiality, integrity* and *availability.* Confidentiality is about the concealment of information. Only authorized entities should be able to access data according to pre-established rules, while unauthorized access should be prevented. Integrity refers to the integrity of data, meaning data should be accurate and reliable. A secure system has to prevent unauthorized modifications of its data. Finally, availability is the concept of providing access to authorized entities at pre-established times. A system fails with regard to availability if a user is unable to interact with the system by the agreed means. For instance, a server that does not display a website due to high load fails the security target of availability. What all three concepts have in common is the need to differentiate authorized from unauthorized access.

A typical process to decide whether an operation is authorized or not involves three steps, as described by Renaud [Ren05]:

- Identification
- Authentication
- Authorization

During the identification step, a user accessing a system provides some form of identifier. This identifier can have the form of a physical (e.g., a passport) or a digital token (e.g., a username). Typically, these identifiers are chosen during a registration procedure and provide a one-to-one mapping to a system-internal entity. Hence, identifiers need to be unique, as the author further states. After the user claims an identity via an identifier, the system needs to validate if the claim holds by requiring proof. This process of providing proof for a claimed identity followed by system-side verification is referred to as authentication, as described by Renaud [Ren05]. Bishop, Sullivan, and Ruppel [BSR19] define authentication as follows: *"Authentication is the binding of an identity to a subject."* Meaning, after the check succeeds, the system has tied the real-life subject (i.e., the user) to a system-internal entity (e.g., a profile). The final step in the described procedure includes a check whether the internal entity holds the requested permissions according to a predefined set of rules and if so, the system grants access. This step is termed authorization.

We focus on the authentication step. There exist different ways for a subject to prove its identity. Renaud [Ren05] differentiates between three types of authentication:

- Knowledge-based authentication
- Possession-based authentication

- Authentication based on biometric data

Knowledge-based authentication systems require a subject to prove its identity via a pre-shared secret. Renaud [Ren05] describes it as "Something the user knows or recognizes". Classic examples of such an authentication mechanism are passwords. A user chooses a password during registration and enters it along with an identifier (e.g., e-mail address) when logging in to a system. If only numbers are used for a password, it is generally referred to as personal identification number (PIN). Both passwords and PINs fall into the category of *memometrics*, as the author describes. The second category of knowledge-based authentication is *cognometrics*. An example system of this category is a graphical password were a user picks a subset of pictures (i.e., the shared secret) from a larger set during each authentication procedure. In contrast to a password, the users do not need to remember the pictures of their shared secret, but rather recognize them among several others, as described by Jermyn et al. [Jer+99]. Graphical passwords are a distinct variation of knowledge-based authentication techniques that use images or drawings to compose the shared secret.

While knowledge-based passwords can be relatively simple to use, they come with certain drawbacks, as Yan et al. [Yan+05] describe, passwords can be hard to remember. Bishop, Sullivan, and Ruppel [BSR19] describe that, in general, the security of a password is measured in terms of complexity. Choosing a larger set of characters (e.g., a combination of letters, numbers and special characters) and taking more characters from this set to form a password leads to higher entropy provided the characters are chosen at random. However, Yan et al. [Yan+05] argue that such random, complex passwords are hard to remember for humans. Therefore, users often adopt coping strategies which can weaken the security of the system. Instead of choosing random passwords, many users select meaningful words and phrases as they are easier to recall. Shay et al. [Sha+10] describe how many systems require users to select passwords that follow a password policy. An example of such a policy could be: must be at least twelve characters long, and include an upper-case letter and a number. This procedure should prevent users from selecting weak, low-entropy passwords. However, as the authors note, it can lead to other coping mechanisms, such as replacing letters in a word with numbers (e.g., "4" for "A") or users writing down passwords. Furthermore, users might reuse the same password across multiple systems.

There exist several attacks on knowledge-based authentication systems, as Bishop, Sullivan, and Ruppel [BSR19] describe. Most threats revolve around guessing or stealing the user's secret. Low-entropy passwords can be susceptible to a *brute-force* attack. An attacker can try every possible password and will eventually succeed unless they run out of time or resources. An attacker might also be able to get a hold of the secret by observing a user while they enter their password (*shoulder surfing*) or tricking a user into revealing their secret (*phishing*). Attackers might use specially crafted websites or e-mails to pose as a legitimate party and ask the user for their credentials (i.e., username and password), as described by Miller and Wu [MW05].

Another way of authenticating a user is via something they hold, e.g., a key to a door lock. Such authentication items are referred to as tokens. Piazzalunga, Salvaneschi, and Coffetti [PSC05] describe different kinds of tokens. Passive tokens, such as a door key, have a unique characteristic, which an authentication system can check. This characteristic is set during the registration process and usually does not change after that. As described, passive tokens are treated similar to knowledge-based authentication methods, without the need for the user to remember a secret. Active tokens, in turn, have an internal mechanism that produces an answer to a challenge posed by the system during authentication. Most technical systems implement such active tokens via public-key cryptography, as stated by Renaud [Ren05]. A smart card or similar device holds a private key that is used to sign a random message presented by an authentication system. The signature can then be verified using a pre-shared public key. The public key can also be stored on the authentication token and returned at the same time as the signature, along with a trusted third-party certificate.

Bishop, Sullivan, and Ruppel [BSR19] explain how active tokens can be used to produce one-time passwords (OTPs). The token is set up with a secret seed. This seed is then used to generate a hard-to-predict sequence of passwords. Each password is only used once. Besides smart cards, such a system can also be implemented using smartphones. In doing so, a user can easily carry along authentication tokens for several systems at once. Using a specific token per system, in contrast, has the advantage of allowing users to pass on tokens easily or produce multiple copies of the same token and distribute them to an authorized subset. This feature can be advantageous in multi-user systems. A prominent example of this practice is the previously examined door key. A downside of using tokens for authentication is that they can get lost or stop functioning, as Renaud [Ren05] further states.

Finally, one of the most natural forms of authentication, according to Bishop, Sullivan, and Ruppel [BSR19], is biometric authentication. Long before computers or other technical systems came around, humans identified each other based on their visual appearance. Most people can easily identify a familiar voice on the telephone. Also, fingerprints have a long history of matching a subject to an entity in a crime scene. Miller [Mil94] describes biometrics as the automated process of verifying the identity of a human being based on physiological (e.g., iris pattern) or behavioral (e.g., handwritten signature) characteristics.

Technical systems use biometrics to authenticate users by comparing a measured, live characteristic to a template stored during registration, as described by Coventry [Cov05]. Eckert [Eck18] differentiates between two categories of biometric authentication. Static features are measurable without user interaction, as they are mostly physiological. Such static features are, for instance, fingerprints or retinas. In contrast, dynamic features are behavioural attributes a person displays when doing something. Therefore, user interaction is required to measure the property. An example of such a characteristic is typing speed on a keyboard, but also hand-made signatures fall into this category.

Not every trait of human subjects can be used as a biometric authentication feature.

Viable alternatives have to comply with specific requirements, as Eckert [Eck18] states. These requirements are:

- **Universality:** Every potential user needs to exhibit the characteristic in question
- **Uniqueness:** It must be possible to identify a person based on the feature unambiguously
- **Persistence:** The feature should remain unchanged over time
- **Performance:** Measuring the feature and verifying an identity should be feasible with reasonable effort
- **Acceptance:** Users need to approve the usage of their feature for authentication purposes
- **Protection against forgery:** It should be hard for an adversary to impersonate a user based on the characteristic in question

Renaud [Ren05] notes that both knowledge-based authentication, as well as authentication by possession are susceptible to human error. Users might forget a password or lose an authentication token. In contrast, biometrics are less error-prone from a user's point-of-view as virtually no cognitive effort is necessary. However, Coventry [Cov05] highlight other drawbacks of biometric authentication. Most anatomical features of human users can change irrevocably. A reason for such changes can be natural ageing, incidents causing physical harm, or a person deliberately changing a physical trait of themselves. Also, technologies used to measure and compare biometric features can be rather complex, compared to, e.g., password-based authentication, as the authors further describe. This complexity can lead to variable reliability and accuracy of biometric data. Specific hardware sensors might be needed depending on the feature used to authenticate a user, leading to higher cost and effort requirements versus non-biometric authentication methods. Also, revocation of biometric authentication data can be difficult, since the used feature might be inseparably tied to the user. Correspondingly, leaked authentication templates can not easily be replaced.

The presented authentication methods can each be used on its own, as a single factor of authentication. However, it is also possible to combine several principles to verify a user's identity. This technique is referred to as two-factor authentication (2FA) or multi-factor authentication, depending on the number of combined schemes, as stated by Bishop, Sullivan, and Ruppel [BSR19]. The authors describe how combining different authentication methods can strengthen the security of a system and cope for weaknesses of the individual schemes. An example of a 2FA system is a credit card. To retrieve money at an automated teller machine (ATM), the user needs to present the card, and enter a PIN. This system is a combination of knowledge- and possession-based authentication. Each component on its own does not suffice to authenticate a legitimate user. However,

such systems also inherit the drawbacks of its components. In the presented example, both the loss of the card as well as forgetting the PIN renders the service inaccessible for the user. In contrast, some weaknesses can also be limited by combining multiple systems, as stated by Renaud [Ren05]. If knowledge-based authentication is combined with a second factor, for instance, a weaker secret (containing less entropy) can be used while maintaining the same level of security. Low-entropy secrets might in terms be easier to remember for users, making the system more usable.

2.3 Usability and Security

Whitten and Tygar [WT99] were among the first to notice the impact bad usability can have on security. In their usability evaluation of an encrypted e-mail program, they found that flaws in the user interface design might lead to security failures. They also reported that most users were unable to use the system and, therefore, get a security benefit out of it. At first glance, there seems to be a trade-off between security and usability. A classical example involving authentication goes as follows: If a computer never asks users for a password, it does not interrupt the primary task and is, therefore, more usable. In contrast, a computer locked away in a safe that requires a drop of blood to open is arguably more secure, while the usability of such a system leaves room for improvement. With their work, however, the authors showed that improving the usability can actually lead to more secure systems. Since then, a growing number of publications has focused on solving security issues by addressing human factors, see also Section 3.1.

Renaud [Ren05] argues that a user is a fundamental part of a system, as are, e.g., input devices, processing units or network connectors. As such, users can and have been a primary target for attackers. The author demonstrates based on password authentication, how low usability can compromise a theoretically secure system: A password is as secure as its entropy if kept secret. However, high-entropy passwords are difficult for human users to remember, even more so if they use a variety of systems, each requiring its own password. Users cope with this cognitive shortcoming by choosing weaker, but easier to remember passwords or reuse the same password across multiple services, the author further states. Therefore, a security breach of one system can lead to a compromising of another, even if the later was theoretically secure.

Password managers offer a solution to most of the drawbacks passwords suffer from, as Pearman et al. [Pea+19] report. These tools work as a storage for users' credentials and come as both on- and off-line services. By saving passwords in a manager, users are no longer burdened with remembering several secrets for different applications. This practice can facilitate the use of strong, random passwords and discourage password reuse across accounts. Therefore, password managers can improve both the usability as well as the security of knowledge-based authentication methods. However, the authors note that adoption of password managers is sparse and users who employ them might misuse the tools (e.g., by storing reused or guessable passwords). The authors report that barriers for the adoption and secure usage of password managers include lack of

awareness, insufficient threat models, and trust issues.

Adams and Sasse [AS99] argue that, against wide-spread believe, users can be motivated to behave securely. In order to achieve such behaviour, it is necessary to align users' perception of security with measures taken by the system. If users have an incomplete or wrong perception of threats that are relevant to them, they might not understand the purpose of security measures and feel obstructed by them in their primary task. The authors highlight how users' mental models about a system and its security properties are important details designers and developers need to understand in order to provide users with the right guidance and adequate security mechanisms.

Mental models are a general concept originating from psychology that represents a persons constructed reality of a situation, as described by Johnson-Laird, Girotto, and Legrenzi [JGL98]. People construct mental models of every aspect of their life based on their perception, imagination and knowledge. These models are then used to derive expected consequences of actions, to reason, and to explain observations. The authors describe that mental models not only include mental pictures of things but also impossible to visualize concepts such as negation. When it comes to usable security, mental models are a representation of how users think about a system, as described by Wash [Was10]. Users consult their personal mental model when they make decisions while interacting with the system. Mental models of technical systems can be constructed based on a variety of factors, e.g., media reports, past experiences or a system's documentation, see Krombholz et al. [Kro+19].

Renaud, Volkamer, and Renkema-Padmos [RVR14] explain that usability alone does not automatically make a system more secure. Lack of awareness of threats and otherwise incomplete mental models can leave users "not caring" about their security. For instance, users might be aware of client-side threats but not of threats targeting data transfer over a network. To give a more tangible example: A user might be aware of being observed while entering the password but not of a man-in-the-middle (MITM) attack, hence sending the password in the clear over an unprotected channel. The authors also highlight the importance of mitigation strategies employed by users. Users will only adopt such strategies against perceived threats, however, frequently, it is the case that users do not know how they could protect themselves even if they sense a threat.

Kang et al. [Kan+15] describe how users are frequently required to make decisions when using technical systems, especially security- and privacy-related ones. The mental models and perceptions users have of a system strongly influence these decisions. People with more technical knowledge tend to have more distinct mental models of a system, however, no direct connection to better security decisions could be found, as the authors report. They point out that incomplete knowledge about a system can make users overconfident about security in their decision making, which can lead to security-compromising actions. The authors recommend that secure systems should limit the amount of security- and privacy-relevant choices users have to make while completing their tasks.

Ion, Reeder, and Consolvo [IRC15] support this recommendation. They describe how

media reports about security incidents lead to increased awareness among users. Another consequence is a larger amount of recommendations given to users on how to protect themselves. While these recommendations can be valuable to inform users' perceptions and help adapt secure behaviour, the authors highlight a potential problem: there might exists a discrepancy between the security perceptions of expert (e.g., security professionals, computer science graduates) and non-expert users. The authors report results from a user study involving 40 attendees of security conferences and 231 other security professionals as well as 294 non-experts. The results indicate that while expert users mostly recommend technical solutions to security problems, e.g., using a virtual private network (VPN), non-experts fall back to non-technical coping mechanisms, e.g., refraining from using a suspicious program. Anell, Gröber, and Krombholz [AGK20] report that this phenomenon is not limited to mitigation strategies but can be observed for all components of a mental model. This difference in perceptions about secure systems becomes relevant during several stages of system development. Expert users most often design the system, implement it and also operate it. If their perceptions about the security of the system do not align with those of the users, decisions taken during any of the described steps might not hold in practice and compromise the security, the authors further explain.

CHAPTER 3

Related Work

We discuss previous work on authentication systems for smart home assistants and their design space. First, we present some work on user mental models. As we have highlighted in Section 2.3, user mental models are a crucial part of the design space of many systems. They affect how users are interacting with and using the systems. This behavior might not coincide with the intentions developers had when designing the system. Users will only protect themselves if they have threats in mind, rendering security and privacy mechanisms less effective. This concept is pointed out, among others, by Anell, Gröber, and Krombholz [AGK20]. Many authors have studied user mental models for a variety of security and privacy-related systems. We will present some recent studies, highlight both qualitative and quantitative methods used to formalize user mental models, and briefly report on the findings presented in those papers.

Afterward, we will report previous security and privacy-related work in the domain of smart homes with a particular focus on voice-controlled personal assistants. Smart home settings saw an increasing interest in the research community in recent years as the adoption among users, and the supply of new technologies grows. Several papers investigated particular challenges that arise in the highly heterogeneous and frequently changing user settings that are commonly found in smart homes.

This work is not the first researching authentication for voice-controlled smart home assistants. Some alternative authentication schemes have already been proposed by different authors, and will be discussed in Section 3.3. The schemes use different features of the voice channel or the smart home environment to identify authorized users and mitigate specific threats. However, to the best of our knowledge, these schemes did not investigate the design space but are based on the authors' assumptions concerning threat models and user mental models. We want to fill this gap in the literature with this work.

Finally, we report various attacks targeting the voice channel of smart home assistants. These attacks demonstrate new threats for voice-controlled platforms that users might

not be aware of. They also refine the attack surface, which is a crucial part of the design space we need to understand before designing future authentication schemes. For this work, we focus on the voice input channel. Of course, further attack surfaces exist, e.g., the network or hardware parts. Furthermore, we executed selected attacks on our platform of interest (i.e., Amazon Alexa) and highlighted how they affect current and future authentication schemes in Section 4.2.

3.1 Exploring User Mental Models

Krombholz et al. [**Kro+19**] investigated end-user and administrator mental models of the Hypertext Transfer Protocol Secure (HTTPS) with a qualitative study. One goal of their work was, amongst others, to formalize user mental models and threat models. For this, the authors conducted a series of semi-structured interviews, collecting both qualitative and quantitative data. First, participants answered a short screening questionnaire, after which they got sorted into either the end-user group or the expert group. Both groups were subsequently given three drawing tasks with different scenarios:

1) Sending an encrypted message to a communication partner
2) Online shopping over HTTPS
3) Online banking

Participants were invited to share their thoughts during the process in a traditional think-aloud protocol. Following the drawing tasks, the authors asked open-ended questions about cryptography on the Internet and potential attackers from which it protects the users. For the latter, participants were also asked to indicate in their drawings, where such attackers might interfere. The authors aimed to recruit a diverse sample of participants and used three separate channels, namely mailing lists, online forums, and personal contacts. They managed to recruit 30 participants for their study, 18 end-users, and 12 experts. An additional 6 participants were recruited for the pilot interviews, 9 more were used for a post-hoc validity study. The later was done to assure the study setup was not priming participants towards encryption since the first task was about that topic.

For the validity study, participants were given a slightly modified version of the task description without any reference to encryption, with the task of drawing a general scenario of sending encrypted messages being replaced by “encryption in theory” which was moved to be the last of the drawing tasks. The results confirmed the validity of the obtained data. The authors proceeded to analyze data via an inductive approach. The goal was to construct theories through the identification of patterns in the collected data. They used inductive coding and performed descriptive axial coding, as described by Strauss and Corbin [SC15], followed by selective coding to obtain categories and models for data. This approach revealed four different types of user mental models. The authors could observe several differences between the two groups studied, e.g.,

end-user mental models appear to be more conceptual, while experts' mental models are more protocol-based instead. Furthermore, misconceptions about threat models and protocol components where revealed, leading to poor decisions by users regarding the security of the system, putting themselves at risk. Other findings include confusion of encryption and authentication, the assumption of an omnipotent attacker combined with an underestimation of security benefits provided by HTTPS, and a general ignorance or distrust towards security indicators such as the lock icon typically used by browsers to indicate a Transport Layer Security (TLS) protected connection.

Obstacles to the adoption of secure communication tools were studied by **Abu-Salma et al. [Abu+17].** Their work explored user threat models and mental models relevant to end-to-end (E2E) encrypted communication. They chose a qualitative approach. As a first step, they conducted ten unstructured interviews. The topics emerging from the analysis of these interviews shaped the guideline used in subsequent semi-structured interviews, for which another 50 participants were recruited. The interviews included a drawing task where participants were asked to visualize how a communication tool works and how it is different from calling someone and sending a text message (short message service (SMS), E-Mail, or instant message). They found that the vast majority of users did not have a sufficient understanding of E2E encryption, and their mental models of secure communication were inaccurate. For instance, many participants perceived calls as more secure than messages, and SMS were perceived as the most secure form for text messages. A reason given for this, as stated by several participants, was that banks use SMS messages to communicate with customers; therefore, it must be secure. The authors conclude that users value quality-of-service and broad adoption of a specific communication medium more than security, especially if security properties are not well understood. Therefore, they encourage future work to look into securing existing, widely adopted systems in order to facilitate user adoption.

Abu-Salma et al. [Abu+18] built upon the work described above. They tried to validate the findings and expand them via a quantitative study. The authors conducted an online survey with 125 participants from the United Kingdom, asking questions about general mental models of E2E encryption and the understanding of a hypothetical encrypted communications tool. The description of this hypothetical tool was based upon popular applications such as *WhatsApp* and *Telegram*. The tool informed the users that communications via this system are end-to-end encrypted. Participants were asked to answer a series of questions about the tool based on a short scenario description, picked randomly from a list of six scenarios. An example of such a scenario is *"Discussing salary with work supervisor"*. The open-ended questions included in the survey inquired participants, among other things, about entities who could read, listen to or modify messages sent via the tool, or impersonate the user. Examples of different entities were provided to respondents. Data collected via open-ended questions were coded and analyzed using thematic analysis. The authors concluded that a high-level description of secure communication given by the hypothetical tool is too vague and does not inform users sufficiently about the tool's security properties. Findings showed that three-quarters

of participants believed unauthorized entities could access E2E encrypted communications. One-half of respondents believed SMS and landline calls to be at least as secure as E2E encrypted communications, if not more secure. Because of inappropriate mental models of E2E encryption like those, users might revert to insecure channels such as SMS when sending sensitive data. Therefore, it is critical for developers to communicate the security properties in an intelligible manner and allow users to make educated decisions when it comes to secure communications.

In 2010, **Wash [Was10]** reported on folk models of security threats. Folk models is a term used to describe mental models that are shared between users with similar cultural backgrounds. These models are often inaccurate and can, therefore, lead to erroneous security decisions. He conducted a qualitative study consisting of semi-structured interviews. In the first round, the author interviewed 23 users of home computers about past instances of security problems and mitigation strategies employed to protect their computers. After analyzing emerging themes from these interviews, the second round was conducted, including 10 additional respondents. In this second round, Wash also made use of hypothetical scenarios. Topics discovered during the preliminary analysis influenced these scenarios. The author targeted three threats, namely viruses, hackers, and identity theft. For each of those, he gave a short description and an additional piece of information that contradicted mental models identified previously. The author identified eight different folk models for both the concepts of *"viruses"* and *"hackers"*.

An example would be comparing hackers to graffiti artists in that both are perceived as highly technically skilled and lack proper moral restraint. Another observation made by Wash was that one of the reasons botnets were so successful might be because they take advantage of gaps in user mental models. Unlike viruses, for instance, they do not harm the affected computer but target third parties. The author, furthermore, investigated possible correlations between mental models users have and how they choose which security advice to follow and which not. He found that users who lack an understanding of threats intentionally ignore advice because they do not believe it will help them. For instance, users who believe viruses to be *"buggy software"*, do not think running an anti-virus program would help them. They instead try to mitigate this risk by controlling what they install on their system.

Threat models and mitigation strategies of older adults were the topics of work by **Frik et al. [Fri+19].** These user groups are often less tech-savvy than younger adults. They can be less aware of privacy and security risks. They are also a prime target for certain specific attacks. The authors conducted semi-structured interviews with 46 participants, age 65 and above. They asked participants about their security and privacy concerns, how they perceive risks, and what mitigation strategies they apply. Afterward, they used thematic analysis to evaluate their data. They found that threat models and associated misconceptions overlap with those of the younger population. However, older adults often have a harder time mitigating those risks and are reverting to system avoidance. It would, therefore, be profitable for developers to consider these audience's security and privacy concerns. The authors recommend educational approaches and enhanced

usability of protection mechanisms for senior users based on their preferences.

We presented several papers investigating users' mental models on various systems. In our work, we apply similar methods to the smart home assistant platform. Knowing about users' perceptions is an important aspect of system design, as highlighted by the presented work. We explore the design space by investigating the perception of threats and mitigation strategies in the context of authentication for smart home assistants. To the best of our knowledge, no work has examined these particular aspects. In line with work presented here, we chose a qualitative approach based on semi-structured interviews.

3.2 Security and Privacy of Smart Home Assistants

Abdi, Ramokapane, and Such [ARS19] examined the security and privacy concerns of Amazon Alexa and Google Home users. They conducted semi-structured interviews with 17 participants who were using a smart home assistant for at least one month. The interviewer presented participants with four different scenarios involving their device: using a built-in skill like the weather forecast or traffic updates, using a third-party skill like *Spotify*, managing other devices in their home such as smart lightbulbs, and online shopping. Afterward, the authors asked participants about whether they think their device could be exploited and, if so, who poses a danger to them. The authors followed a grounded theory approach. They found that many users have incomplete mental models of how smart home assistants work and what threats users are exposed to when using them. They note that the case of online shopping was particularly interesting, as many users did not use this feature due to security and privacy concerns. We build upon these findings and explore users' perceptions of high-risk tasks on smart home assistants more in-depth. Additionally, we put a particular focus on authentication, since it is among the first security mechanisms users come in touch.

Zeng, Mare, and Roesner [ZMR17] asked users living in a smart home about their attitudes and expectations towards security and privacy. They conducted 15 semi-structured interviews to elicit mental models of smart home technology, threat models, and mitigation strategies. They report that users often have incomplete mental models of how IoT devices interact within the home or with the outside (i.e., the cloud), and mitigation strategies are often adopted from older technologies. When it comes to threats, users show to have varied and sparse models. The work also describes tensions that may arise in multi-user homes where there is a disparity in power users have over their devices. Similar to the authors, we also investigate which mitigation strategies users apply. However, we focus on security-critical tasks such as money transfer. These tasks call for more sophisticated security mechanisms, which are not yet available.

Zeng and Roesner [ZR19] built a prototype smart home app based on design principles proposed by their and other previous work. They used this prototype in a month-long in-home study, including seven households, to explore behaviors and solutions to users' security and privacy issues. They identify several challenges that remain open, amongst others the importance of incorporating voice assistants into access control systems. The

authors argue that smart home assistants are given unrestricted access to most IoT devices in a smart home and can, therefore, be used to bypass access control mechanisms imposed by the studied prototype smartphone app. They highlight the importance of sophisticated voice-based authentication in order to apply access controls consistently.

He et al. [He+18] conducted a user study with 425 participants investigating which factors are essential for users when it comes to access control in a smart home environment. The authors argue that current access control mechanisms only distinguish between full and no access, while some feature a guest account with limited capabilities. However, smart home environments are composed of heterogeneous individuals interacting with the same multi-user devices. The authors identify several key aspects upon which users base their preferences of access control. For example, elementary-school-aged children should only be able to access any capability while an adult is nearby while visiting babysitters might be able to regulate the light if they are inside the home. The authors, furthermore, discuss the importance of user authentication in order to enable access control mechanisms. They note that smart voice assistants are, besides smartphones, a central entry point for smart homes and can, therefore, be charged with user authentication. However, the authors point out that, while audio authentication can potentially identify contextual factors important for access control policies (e.g., persons nearby), current mechanisms are insufficient and may introduce privacy issues. In this work, we improve the current state of authentication by exploring users' perceptions of threats and highlighting important aspects of future authentication schemes. These contributions are valuable for the design of safer authentication systems, which in turn enable improved access control functionalities.

Cho [Cho19] examined whether different input modalities (voice vs. text) had an impact on users' perception when asking smart voice assistants sensitive health-related information. The author also investigated the influence of different devices (smartphone vs. smart home assistant) in such a scenario. 53 participants were instructed to ask a voice assistant both high and low sensitive health-related questions. The modalities (voice on smartphone, text on smartphone, and voice on smart home assistant) were tested in a between-subjects setup. The results suggest that voice interaction leads to a significant enhancement in users' perception of social presence, but only for less sensitive information. At the same time, there was no significant difference between smartphones and smart home assistants. Overall, voice interactions made users feel more like having a social conversation in conditions where the questions asked were less sensitive. This characteristic was mainly observed for users with low privacy concerns. The findings made by the authors can aid us in formalizing crucial properties of authentication systems for smart home assistants, which we examine in this work.

3.3 Authentication Schemes for Smart Home Assistants

Feng, Fawaz, and Shin [FFS17] proposed a novel authentication technique called *VAuth* for voice assistants. Their method uses body-surface vibrations to detect the

origin of a speech command received by a voice assistant. For this purpose, they use an accelerometer that is in contact with a user's skin. It registers the person's voice through vibrations rather than over the air. The device then transmits the registered signals via Bluetooth to the voice assistant. Here, an additional piece of software compares both the signal from the accelerometer as well as the one received by the voice assistant's built-in microphone. Commands are only executed if there is a sufficient correlation between the two. This system provides continuous authentication.

In contrast to other schemes like passwords or PINs, which authenticate a user at the beginning of a session, continuous authentication does not interrupt the primary task. It authenticates every action taken by the user. The authors built prototypes using off-the-shelf wearables that already include accelerometers, namely earbuds, necklaces, and glasses. They chose items people were already familiar with, so users were not burdened by yet another device they had to carry. The results were promising; the prototypes were able to detect voice commands with an accuracy of 97%, while the false-positive rate was at 0.1%. The authors carried out a usability study via Amazon Mechanical Turk (MTurk). 952 participants, United States residents with voice assistant experience, were asked to fill out a questionnaire, rating statements on a 7-point Likert scale (see Likert [Lik32]). The survey included questions about the perceived security of voice assistants, willingness to use wearables with built-in security technology as well as a general-purpose usability questionnaire. They, furthermore, assessed the systems under different attacker scenarios. Their threat model included attackers interfering with the audio channel of victims' voice assistants. Incentives are stealing private information and executing unauthorized commands. Attacks can consist of inaudible or mangled voice commands, replaying of pre-registered audio commands by the user, impersonation by mimicking a trusted user's voice, and interfering with the accelerometer via audio signals such as loud music. The authors report that their system is only vulnerable in the scenario in which an attacker can play loud sounds within close distance of a user wearing the *VAuth* device.

Blue et al. [Blu+18] proposed another authentication scheme termed "Two Microphone Authentication". It aims at detecting the presence of a particular user by combining direction of arrival (DOA) with robust sound hashs (RSHs). DOA is a technique to determine the direction of an audio source, using at least two microphones. The differences in time these two microphones receive the same signal is used to calculate an estimated origin of the signal. RSHs are used to determine if the signals received by the microphones are indeed identical. Since the audio channel is susceptible to small changes in signals (e.g., background noise), classical cryptographic hash functions cannot be used. RSH produce so-called speech digests derived from unique features such as words spoken or semantics of a sentence. The authors make use of personal smartphones in their authentication protocol as the second microphone. Their security model assumes that a user is in possession of his smartphone at all times, allowing them to authenticate the user via proximity to the phone. Both the smartphone's built-in microphone and the voice assistant receive voice commands issued by the user. The mobile phone calculates

the RSHs and sends it to the voice assistant along with a timestamp of when the signal was received, authenticated by a pre-shared key. The voice assistant checks if the RSHs of the received commands match. Also, the DOA is computed to check if the command's origin is closer to the smartphone than the voice assistant.

Lei et al. [Lei+18] proposed physical presence as an authentication factor. They observed that attacks on voice user interfaces (VUIs) are mainly carried out while no user is present. Attackers could exploit Bluetooth speakers or smart TVs to inject voice commands, as shown by proof-of-concept attacks presented in the authors' work. They argue that any user present would notice an ongoing attack and prevent it from succeeding. Their proposed scheme is called Virtual Security Button (VSButton). The name was chosen to show the resemblance to a physical button that needs to be pressed before a VUI accepts commands. The proposed method makes use of Wi-Fi technology already used by Smart Home Assistants. The authors describe how it is possible to detect human movement inside a room by observing changes in the channel state information (CSI) of Wi-Fi networks. These changes are introduced by reflections that differ depending on the location of an object or a person. Changes in the location are assumed to be due to human motion, which can be as small as moving an arm 20 centimeters. It is also possible to differentiate between motions occurring inside the room from those behind walls. Therefore, attackers have to gain physical access to the room, i.e., the house, in which a VUI is located to interact with it. The authors have tested the quality of their authentication method with a user study, including 6 individuals in a real-world setting.

Zhang, Tan, and Yang [ZTY17] developed a system for speaker liveliness detection. They wanted to mitigate the threat of *Replay Attacks* on VUIs were attackers use pre-recorded audio from an authorized user to bypass speaker authentication, see also Section 4.2.2. Their system extracts features in Doppler shifts and uses them to distinguish audio generated by a loudspeaker from humans speaking directly to an interface. This technology could be integrated into voice-based authentication systems, its only hardware-requirements being a speaker and a microphone.

Kwak et al. [Kwa+19] pursue a similar goal. They propose a machine-learning-based system capable of differentiating between genuine user commands and malicious commands issued by an attacker. They note that 87.45% of users issue less than 20 voice commands per month. Their system uses text-converted utterances as a feature to train a global model, which can then be further adapted to specific users. It could be used to complement authentication systems and aid in the detection of attacks on VUIs.

The authentication schemes we presented build upon various assumptions about how users interact with smart home assistants and what perceptions of threats they have. In our work, we investigate these perceptions with active users of the system and formalize concrete hypotheses based on qualitative methods. Therefore, we lay the groundwork for improved systems that take both users' and security needs into account.

3.4 Attacks on Smart Home Assistants

Several attacks have been presented on smart home assistants. **Zhang et al. [Zha+17]** described the "DolphinAttack", which exploits the non-linearity of commonly used microphones. This technique allows attackers to emit inaudible audio signals in the ultrasonic spectrum, which are processed by smart speakers and interpreted in the same way as human speech. Therefore, it enables stealthy attacks on smart home assistants, which are impossible to detect by bystanders without dedicated technical equipment. The authors carried out several different attacks against Apple Siri running on a smartphone. This system applies speaker recognition to accept the wake word ("Hey Siri") only from previously trained voices. Therefore, an attacker must bypass this biometric authentication. The authors conducted a series of experiments to test the feasibility of different attack scenarios. The authors used a TTS system from Google to train Siri, posing as a legit user in this scenario. Other TTS software from nine different vendors was used to generate the wake word and a command (*"call 0123456789"*). Out of 89 different types of voices used, 39 resulted in successful recognition of the word. These results show that simple brute-force attacks on speaker recognition are possible.

All TTS-issued commands were accepted. Therefore, the authors concluded that only the wake word is authenticated but not subsequent commands. This finding enables another attack, which the authors described as concatenative synthesis. In this scenario, an attacker can record arbitrary words spoken by a legitimate user, however, not the wake word. Sampling these records and extracting phonemes allows the attacker to build necessary texts that then get accepted as if spoken by the user. Using the words "**h**e", "c**a**ke", "**ci**ty", and "ca**rry**", it is possible to construct "Hey Siri". Such attack vectors can be used in combination with an ultrasound carrier to inject inaudible voice commands into Siri successfully. The authors also performed tests in an experimental setup, including different languages and background noise levels.

Roy et al. [Roy+18] presented an advanced attack, built on the same principles, that extends the range of a successful attack from 5ft to 25ft. They use an array of speakers operating with higher power usage. This technique limits frequency leakage into the audible spectrum introduced by the exploited non-linearity of the microphones. This setup prevents the attack signal from becoming audible even over more considerable distances. The authors also propose defensive strategies that can protect from inaudible attacks. Signal forensic methods can be used to detect artifacts produced by an attack signal but not by an ordinary human voice. Such an artifact could be, for instance, frequencies below 50Hz Since human voice is unable to produce such low frequencies, this can be a good indicator of leakage produced by high-frequency attacks.

Sugawara et al. [Sug+20] built on top of this with their attack named *"LightCommands"*. It exploits a vulnerability in micro-electro-mechanical systems (MEMS) microphones, which allows attackers to inject commands via potent light sources such as lasers. These commands are by their very nature inaudible and can be injected from a noticeable distance. In several experiments, the authors demonstrate that carefully

modulated light beams aimed at microphones of smart speakers like Google Home result in command input as if it originated from an audio source. These results, of course, have severe implications on smart home assistants, since it allows attackers to interact with such systems without having direct physical access to the device. The authors demonstrate the feasibility of such an attack by injecting commands over a distance of 75 m. They managed to attack a system located inside an office from another building across the street. Only a direct line of sight between the attacker and the microphone of the targeted device is necessary. Attacks are also inaudible to users standing nearby. Users can only observe reflections of the laser on surfaces, the activation sound of the smart speaker, or, depending on the model used, a light signal indicating that the device is recording voice input. The response produced by the device following such an attack is still being output through the device's speakers. However, in most cases, such feedback is only issued once the smart home assistant has processed a command. The authors note that strong, possibly continuous, authentication can be an effective mitigation mechanism for this kind of attacks.

Kumar et al. [Kum+18] built the *Skill Squatting* attack on the observation that many users, especially those speaking with an accent, experience frequent misinterpretations by Alexa. An investigation of 11,460 speech samples conducted by the authors revealed that such misinterpretations appear consistently. This behavior can be leveraged by an attacker trying to trick Alexa into opening a malicious skill instead of the one intended by the user. To do so, the attacker creates a public skill with a specially crafted invocation name. This invocation name is used by Alexa to determine the user's intent and redirect the control flow accordingly. Recognition errors made by Alexa can be caused, among others, by homophones, e.g., "sail" vs. "sale", compound words, e.g., "outdoors" vs. "out doors" or phonetic confusion, e.g., "dime" vs. "time". The authors used publicly available speech sample databases to perform two-fold cross-validation to test how vulnerable different word pairs are to *Skill Squatting*. The attack succeeded at least once for 25 out of 27 pairs tested. In 3 cases, the attack had a success rate of 100%. This attack can also be targeted at specific users, as the authors point out, and is then referred to as *Spear Skill Squatting*. An attacker can leverage, for instance, specific words in the targeted user's language to obtain better results. The authors describe a significant increase in the success rate when targeting specific demographic groups defined by gender or region of origin.

Yuan et al. [Yua+18] described an attack called *CommanderSong*. They used techniques from attacks on image recognition software and adapted them to the voice recognition engines used in personal assistants on smartphones or smart speakers. The attack consists of perturbations introduced to carrier signals to make the language model employed by the speech recognition engine recognize a command that is then executed by the device. Such carrier signals can be songs or audio tracks of videos. This vector allows for a widespread attack on users while also being stealthy. The authors conducted a survey on MTurk, where participants where asked whether they could notice anything abnormal about songs manipulated to contain voice commands. The results show that

users noticed some form of abnormality 6-33% of times, depending on the song chosen as a carrier. However, no participant was able to identify the injected command correctly.

The related work we presented gives us valuable insights into real-world threats that target smart home assistants. Any authentication system needs to consider such risks and protect users accordingly. In this work, we investigate what threats users believe are relevant to them in high-risk scenarios. We highlight potential gaps in their perception and explore which aspects are decisive to give users a feeling of security.

CHAPTER 4

Attacks on Voice-Controlled Authentication Systems

We highlight several existing attacks that target voice input systems, with a particular focus on smart home assistants. We describe these attacks and derive implications that result from applying them to smart home assistant authentication. We also execute a selection of attacks on our platform of study, i.e., Amazon Alexa, to test their feasibility and necessary efforts. Most attacks we cover have been conducted on other smart home assistants or different voice-input systems. Due to similar hardware (see Sugawara et al. [Sug+20]) and software (see Zhang et al. [Zha+17]) characteristics, we can assume that most can be transferred across platforms. We test this assumption for some of the presented attacks in Section 4.2.

4.1 Examples of Attacks

We present a variety of attacks on authentication schemes that target voice input systems and are, therefore, relevant for smart home assistants. Furthermore, we also highlight incidents involving smart home assistants that have implications for these systems' security, especially for authentication. As described in Section 1.2, we focus on attacks via the voice input channel. The attacks we cover can be grouped into two main categories:

- Faking a user's voice: mimicking a user's speech might allow an attacker to bypass voice-based biometrics authentication systems.
- Inaudible command injection: by injecting commands in an inaudible way, attackers can avoid detection by humans nearby, enabling attacks on smart home assistants while the user is in the house.

In this work, we do not consider attacks targeting the hardware of smart home assistants directly. Also, attacks on the firmware and network connections of such systems are out of scope.

4.1.1 Faking a User's Voice

One of the most natural authentication mechanisms for VUIs is voice-based biometrics authentication, according to Bishop, Sullivan, and Ruppel [BSR19]. This technique extracts characteristic features from users' utterances and compares them to data acquired during the voice profile registration. It relies on the assumption that each user's voice is unique, given certain criteria that can be measured. If an attacker succeeds in presenting a system with an utterance having the same (or at least similar enough) characteristics as a legitimate user's voice, she can bypass the authentication. Most systems require users to say a certain word or sentence each time they want to authenticate. This phrase does not have to be secret, since authentication is only done on voice specific features. See Section 2.2 for further information. For smart home assistants this activation command is typically referred to as *wake word*, see Section 2.1.

A classical attack on such a system is the *Replay Attack*. This technique uses a recording of a user's authentication with the system and replays it to gain access, as described by Janicki, Alegre, and Evans [JAE16]. Therefore, an attacker needs to obtain a speech sample of the user interacting with the system of sufficient quality and replay it to the system without getting detected. While this attack is of relatively low technicality, it might be difficult for an attacker to obtain the necessary speech sample. Depending on how familiar the attacker is with her target, this may be a more or less easy task. While recording another user living in the same home is often straightforward, specific speech samples of unfamiliar targets might prove to be more difficult to obtain, e.g., a celebrity. Voice recorders are widely available nowadays, primarily smartphones with the appropriate software. These devices can also function as a speaker and replay the adversarial sample. More sophisticated tools exist on the market, which can improve the quality of the recording. Dedicated wiretapping tools allow stealthy recording due to the microphone's small size or concealment in everyday items. Background noise can influence the quality of a recording. An attacker might need to filter it out or shield the microphone during recording.

Certain authentication systems use a different text (or pool of texts) for every authentication action, e.g., presenting the user with a text they must read back. A simple *Replay Attack* is not possible, since the recorded phrase might not correspond to the one currently required by the system. Zhang et al. [Zha+17] present an attack named *voice sampling*. The authors were able to construct phrases with a user's voice by concatenating together pieces of other words spoken by that user. Such pieces are called *phonemes*. The English language, not considering regional dialects, is composed of about 44 such phonemes. For instance, the word "voice" comprises the four phonemes |vɔɪs|. By extracting every possible phoneme from a user's voice samples, given a sufficiently large number of recordings, an attacker can theoretically construct every phrase in the

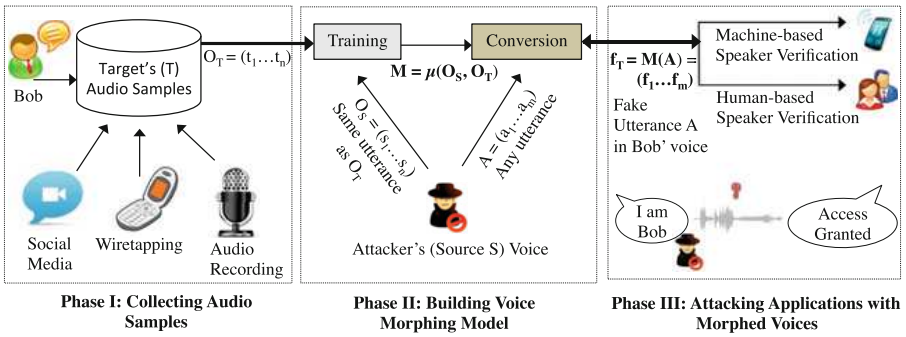

Figure 4.1: During Phase I, the attacker collects audio samples $O_T = (t_1 \ldots t_n)$ from a targeted user T (i.e., Bob). This acquisition can be done via social media (e.g., videos uploaded by the target), wiretapping the user's phone, or record live audio if the attacker has physical access to the target. The acquired samples are used during Phase II to train a voice morphing model along with syntactically identical samples $O_S = (s_1 \ldots s_n)$ spoken by the attacker (source) S. The model $M = \mu(O_S, O_T)$ resulting from this training can then be used to convert an arbitrary utterance $A = (a_1 \ldots a_m)$ from the attacker's voice to the target's voice. The resulting fake utterance $f_T = M(A) = (f_1 \ldots f_m)$ is then used in Phase III to deceive both machine- and human-based speaker verification. For instance, an attacker could generate the phrase "I am Bob" in the targeted user's voice. This utterance would be verified by a voice recognition system and grant access to the attacker. (reprinted from Mukhopadhyay, Shirvanian, and Saxena [MSS15])

specific language. In practice, varying audio quality (due to background noise) and the rate of speaking can impede this process. However, the authors noticed an important property of the voice assistant they were examining (i.e., Apple Siri). The system only authenticates the wake word but does not check subsequent commands. Therefore, it suffices to obtain voice samples for the phonemes of "Hey Siri", which is composed of |heɪ| and |sɪri|. The authors accomplish this by recording the words "he" (|hi|), "cake" (|keɪk|), "city" (|ˈsɪti|), and "carry" (|ˈkari|), extracting the required phonemes and assembling them as needed to construct the wake word. The authors successfully impersonated a user against the voice assistant on an iPhone 4S.

The *voice sampling* attack, as described above, becomes more extensive if the authentication system requires longer or phonetically more complex phrases. A significant amount of time is needed to extract the necessary phonemes from recordings and concatenate them together. Mukhopadhyay, Shirvanian, and Saxena [MSS15] highlight a different technique called *voice morphing*. The authors use voice samples of a target (i.e., a user) and attacker samples of the same phrases and use them to train a model. Figure 4.1 depicts this process in detail. The authors used an open-source voice conversion software [Pro20]. The trained model can then be fed with new phrases spoken by the attacker and output those phrases in the voice of the targeted user. The authors tested their method against the Spear open-source speaker recognition tool, developed by Khoury, Shafey, and Marcel [KSM14].

The results showed that the attack succeeds with an averaged probability above $80-90\%$, for most tested targets. The authors claimed that only a few minutes' worth of speech samples from the targeted user are required to achieve similar results.

Furthermore, the authors presented participants with fake speech samples generated by the described method within an online user study. Almost 50% of participants could not distinguish generated samples from a genuine targeted speaker (e.g., a celebrity). The authors note that the success rate can be improved using a larger amount of high-quality audio samples and the attacker trying to mimic the target's voice as much as possible.

Jia et al. [Jia+18] improve the efficiency of speech synthesis significantly by using transfer learning. Their proposed method combines three individually trained networks: a speaker encoder network, a synthesis network and a vocoder network. The authors claim that their system can create arbitrary synthetic utterances in a targeted voice, unseen during training, from a few seconds of untranscribed audio data from that voice. The authors evaluated the generated voice in a user study with regard to naturalness and similarity to a genuine speech sample from that voice. Depending on the model used for training, the resulting mean opinion score (MOS) (see Streijl, Winkler, and Hands [SWH16]) was 3.28 ($\sigma = 0.07$) and 3.03 ($\sigma = 0.09$) respectively, indicating a good similarity. The authors note that their system might be abused in order to impersonate users, possibly against voice-based authentication systems. However, they also demonstrate how a system might mitigate this threat by distinguishing synthesized from natural speech. They developed an evaluation-only system with a similar setup to the previously described one. By using a synthesized training set of 1200 speakers, their model was able to distinguish synthesized speech correctly with an equal error rate (EER) of 2.86[1].

4.1.2 Hidden and Inaudible Command Injection

Smart home assistants are typically located inside the home of their users. In order to interact with the system, an attacker, therefore, might need to be in physical proximity of the smart home assistant to interact with it and carry out attacks over the audio channel. While audio waves can travel through walls and windows if they are loud enough, such interactions can be detected more easily by inhabitants or neighbours. An attacker might also try to get access to the house by tricking the user and interact with the system at close range. However, this interaction can still be easily detected by the user due to the openness of the voice input channel. We highlight several presented attacks on smart home assistants that allow inaudible and other forms of hidden command injections. Such techniques allow attackers to interact with a system without drawing the attention of nearby humans.

Several papers brought up the idea of attacking voice-controlled systems via hidden voice commands, e.g., Carlini et al. [Car+16], Vaidya et al. [Vai+15], or Roy, Hassanieh,

[1] EER is the value of error rate (out of 100) at which *false acceptance rate* and *false rejection rate* are equal. Hence, it gives the probability for both a fraudulent entity to be verified as legitimate and a legitimate user being rejected, as described by Scheuermann, Schwiderski-Grosche, and Struif [SSS00].

and R. Choudhury [RHR17]. These techniques are sometimes referred to as *mumble attacks*. Sounds used by such attacks, while still audible, cannot be identified as voice commands by humans, but are processed by smart home assistants. This method exploits vulnerabilities of the machine learning techniques that build the base of STT systems used in voice-controlled smart home assistants. Malicious audio can be played from any source equipped with a speaker. The attack distance depends on the used sound level. Since the attack is audible, nearby humans might still get suspicious about the incomprehensible noise they hear, even more so, the louder the sound is.

Yuan et al. [Yua+18] improved on this technique with their attack called *CommanderSong*. They achieved higher stealthiness by incorporating the hidden voice commands into common songs. In a proof-of-concept described in the paper, the authors use the open-source speech recognition system Kaldi[Pov]. By analyzing the acoustic model used by the software, the authors were able to identify the key features used to determine the outcome of the deep neural network. They modified ordinary songs in a way that the model extracted the desired features while the song was still recognizable to humans. They achieved up to 96% success rate using off-the-shelf speakers. The distance between speaker and microphone was 1.5 m. During a user study conducted by the authors, an average of 65.2% of participants, who listened to the song, noticed some anomalies, however, most attributed them to noisy speakers. No participants correctly identified the hidden voice command present in the audio samples, even when listening to it repeatedly. The authors were also able to transfer the attack with selected commands to the iFLYTEK[iFL] and DeepSpeech[Moz20] speech recognition systems.

Zhang et al. [Zha+17] were among the first to propose an inaudible attack on voice-controlled systems. Their attack is called *DolphinAttack*. The authors noticed that MEMS microphones used in most modern smartphones and also smart home assistants exhibit non-linear properties when receiving sound in the ultrasonic frequency range (i.e., > 20 kHz). They developed an attack were standard voice commands are modulated on to ultrasound carrier signals. This kind of high-frequency sound is no longer audible to humans. However, most microphones will still process such a signal and perform demodulation, re-constructing the initial low-frequency voice command. Any remaining high frequencies are filtered out by a low-pass filter employed by most voice-controlled systems. Therefore, any speech recognition software presented with the input interprets it as if it was a genuine voice input.

The authors tested their attack against several off-the-shelf voice-controlled systems (among which also Amazon Alexa on an Echo Dot). Their attack setup included a signal generator taking care of the modulation, and a wide-band dynamic ultrasound speaker. They were able to attack most systems successfully. For the Amazon Echo Dot, the maximum distance at which the attack succeeded was 1.65 m. This distance could theoretically be increased by using more powerful equipment. Furthermore, the authors developed a portable attack setup, consisting of a smartphone, an amplifier, an ultrasound transducer, and an additional battery. All but the first could be acquired for less than $3. This portable attack device was effective within 27 cm of the target.

The authors propose hardware- and software-based defence mechanisms. Microphones should be designed to suppress signals in the ultrasonic (i.e., inaudible) range. Legacy systems could actively search for modulated signals, extract the baseband and subtract it from the processed signal. Software-based methods could employ machine learning to detect anomalies in the input signal, preferably in the range of $500 - 1000$ Hz. Roy et al. [Roy+18] improved the attack range of this technique by adjusting the hardware setup. Instead of a single speaker, they used an array of physically separated ultrasound speakers and a custom-made amplifier. Their improved setup achieves an increase in attack distance beyond 7.6 m. They managed to successfully attack smart home assistants from outside the house, through an open window. The authors noted that their setup could be effective over an even larger distance if one increases the power of the amplifier, without leaking into the audible frequency range. This trade-off between attack range and audibility was a major weakness of the original setup, according to the authors. They described a possible adversarial scenario were an attacker simultaneously attacks several smart home assistants in a neighbourhood, from a car parked on the street.

Sugawara et al. [Sug+20] developed a different kind of inaudible attack on smart home assistants called *LightCommands*. Rather than using sound, this attack exploits a phenomenon of MEMS microphones were powerful light signals aimed at the diaphragm are recognized as audio signals. By carefully modulating the amplitude of the laser beam, the authors were able to convert sound signals into light signals, which can then be registered by a microphone. This technique can be used to attack a variety of smart home assistants and other voice-controlled devices. An attacker only requires direct line of sight to the microphone of the target, but no physical access to the device. The authors report that, depending on the targeted model, minimum activation power varied between $0.5 - 60$ mW (from a distance of 30 cm). They tested several devices using a 5 mW laser point, as this is the maximum allowed power output in the United States for commercially available products. Some devices, such as the 1st Generation Amazon Echo Dot or the Google Home Mini, could be activated from a distance of over 110 m. The authors note that attack range is only limited by the attacker's power budget, optics and aiming capabilities.

In a proof-of-concept attack, the authors show a potential attack vector using the *LightCommands* technique. They set up a 5 mW laser pointer, which was mounted to a telephoto lens on a tripod head, on top of a tower. Using a telescope, the authors aimed the laser beam at the microphone of a Google Home Mini located inside an office building 28 m below their position and across the street. The total distance between source and target was 75 m. Furthermore, the light beam had to pass through a closed double-pane glass window. Figure 4.2 shows the setup. At the bottom left is a depiction of the target as seen through the telescope. On the right, the beam of light is visible. The attack was successful without device- or window-specific adjustments. This proof-of-concept shows that this attack is possible under real-world conditions. All materials used in this setup were commercially available. A low-cost toolkit, effective on short distances, can be built with $18, as the authors note.

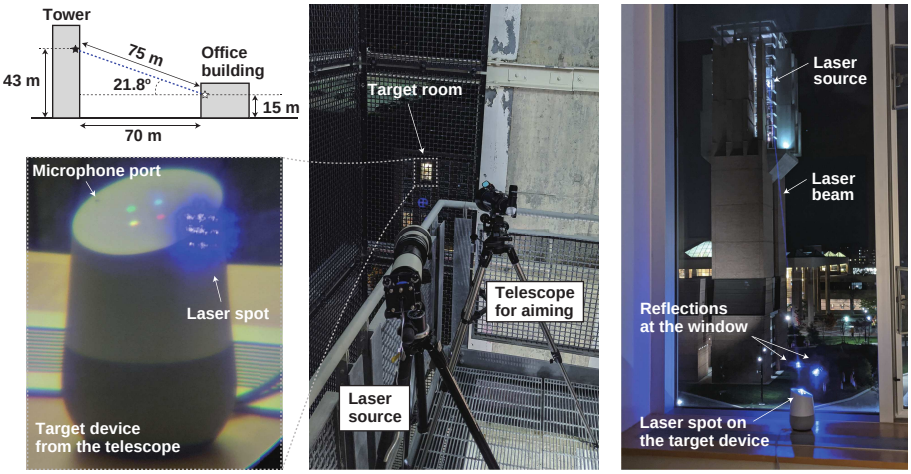

Figure 4.2: Depiction of the *LightCommands* attack under realistic conditions. (Top left) The attacking setup was placed on a tower, 43 m above ground level, while the target was located inside an office building 70 m across the street, 15 m high. The total distance the laser had to travel was 75 m, hitting the target at an angle of 21.8°. (Bottom left) The targeted device (Google Home [Gooc]) as seen through the telescope part of the attacking setup. The attacker has to aim the laser spot precisely at the microphone port to inject commands. (Middle) The attacking setup included the laser source mounted to a long-distance lens and the telescope used to aim the beam. The targeted room is visible in the background. (Right) The attack as seen from the targeted room. The laser source located in the tower is emitting a beam towards the targeted device. This beam is partially visible due reflections from microparticles in the air, reflections at the window, and on the surface of the target device. (reprinted from Sugawara et al. [Sug+20])

However, a user could still be able to detect this attack. First, it is possible to notice the laser beam both from reflections in the air or on the glass as well as on the targeted device itself, where it leaves a visible spot. Second, a user might notice the device responding to the received command. The authors propose solutions for both cases. Since the wavelength of the laser beam did not have any measurable effect on the attack efficiency, an attacker might use light in the infrared spectrum, which is not visible for humans. A special camera is then needed to aim the beam correctly. For the second case, the attacker might issue a command to lower the volume of the smart home assistant or completely turn off voice output. The only remaining feedback is light-emitting diodes (LEDs) turning on when a command is received, which are present on some devices. The primary defence against this kind of attack would require changes to the design of the microphone's aperture, such that it is shielded from light beams, while still letting sound waves pass through. Also, software-based defences are possible, e.g., comparing the input of multiple, physically separated microphones.

4.1.3 Malicious Skills

Mitev, Miettinen, and Sadeghi [MMS19] used ultrasound voice commands in combination with a malicious skill to perform a MITM attack on Amazon Alexa. The attack is termed *Lyexa*. A pre-requirement for this attack is that the attacker is controlling an IoT device equipped with a microphone and speaker, located close to the targeted Alexa device. The attack works as follows: the attacker-controlled IoT device continuously listens for the wake word of Alexa, in the same way a genuine Alexa device does. If the wake word is detected, the malicious device starts emitting ultrasonic signals to jam the Alexa device and prevent it from understanding the subsequent command. Meanwhile, the malicious device records the command and sends it to an attacker-controlled back-end server. At the same time, the malicious device issues another inaudible command to invoke an attacker-provided skill on the Alexa platform. The back-end server relays the command to the Alexa voice service (AVS) application programming interface (API). The response is then echoed back to the user by the malicious skill running on Alexa. This procedure results in a passive MITM, which can be turned into an active attack if the back-end server modifies the request or response.

Kumar et al. [Kum+18] presented a different attack using malicious skills. They noticed systematic interpretation errors made by AVS. Most of these errors are due to homophones (words with a different spelling but same pronunciation) or words with similar phonetic structure. Examples of such words are “sail” vs “sale” (both pronounced |seɪl|) or “wet” (|wɛt|) vs “what” (|wɒt|). The authors proposed an attack called *Skill Squatting* that exploits such misinterpretations. They describe how an attacker can craft a malicious skill with a name phonetically similar to a targeted genuine skill and upload it to the official Alexa skill platform. If the user wants to invoke the genuine skill, due to interpretation errors during speech recognition, Alexa might instead invoke the malicious skill, with high probability. The attacker can craft the skill to mimic the interaction of the genuine one, giving the user little chance to notice the attack.

The authors showed that several skills potentially vulnerable to this attack (due to existing similar words) are currently available for the Alexa platform. The research also identified eight pairs of skills with phonetically similar names, which might be an indicator of *Skill Squatting* being already exploited in the wild, however, no clear evidence of this could be found. Finally, the authors report on a variation of this method, called *Spear Skill Squatting*, that targets specific users. By leveraging characteristics of a specific demographic group’s speech, e.g., dialects, misinterpretations can happen more often and in a more predictable way. These adoptions can significantly improve the likelihood of a malicious skill getting invoked instead of the user-intended one.

4.2 Executing Selected Attacks

We wanted to test if some of the attacks presented in Section 4.1 work on current Amazon Alexa devices. Considering that in the meantime the hard- or firmware might have been updated, attacks might no longer work or might not translate to Alexa from other smart

home assistants. This investigation also gives us a better understanding of the difficulty of executing the attacks in a realistic scenario.

4.2.1 Experimental Setup for Execution of Attacks

Our attacker model for these experiments is a low-tech attacker limited to a small budget. The attacker can interact with the smart home assistant for a short time unattended by any user. Furthermore, the attacker might be an acquaintance of the targeted user and possess some insider knowledge about them. Such an attacker could be a typical social acquaintance of a user who gets unsupervised access to the home to, e.g., water the plants while the user is on holiday. Another potential attacker fitting the description would be a housekeeper. Due to the given attacker model, there is no need for the attack to be stealthy since no bystanders are present. We assume the attacker can issue several commands and also observe the response. An attacker might be motivated by monetary gains such as ordering goods on the user's account.

Our initial setup included a 2nd Generation Amazon Echo Dot running firmware version 64764212. We also confirmed the results later on with a 3rd Generation Amazon Echo Dot (firmware version 2919948420). The device was located in our office with typical, low-level background noise. Alexa was used with a German account. Therefore, we had to set the interaction language also to German, since online shopping is only possible in the principal language of the account. For some attacks, we used a Samsung Galaxy A3 2017 smartphone to record and replay the audio. We set up an Amazon account for testing purposes and activated the voice shopping feature. We used online shopping via Alexa as an example of a high-risk task that is already available for this platform in Germany and Austria.

For our tested firmware version, Alexa supports two different authentication mechanisms when performing online shopping. First, a 4-digit PIN needs to be set up before a user can use this feature. Second, a user has the option to register a voice profile. This setup requires a user to speak predefined phrases to Alexa, which will train the voice recognition model for that account. If a voice profile was installed, users have the option to allow voice purchases for registered voices only. They can also choose whether recognized voice profiles need to input the PIN always or only on the first purchase. These options lead to a total of three different scenarios of authentication for voice purchases via Alexa:

1. Knowledge-based authentication via the PIN
2. Biometric authentication via the voice profile
3. A combination of both

Since getting a PIN from a user might be as straight forward as eavesdropping on an interaction with Alexa, we focus on the second authentication scenario. For our experiments, we assume the user has set up a voice profile and already made a purchase

using the PIN before, allowing biometric-only authentication for this profile. The goal is to *fake* a user's voice and bypass the authentication. For testing purposes, we used the Alexa command *"Alexa, who am I?"*. This command behaves similar to the `whoami` command present on most operating systems. Alexa will answer with the name of the user or will say "I'm not sure who's speaking." if no registered user was recognized. We did not use the shopping feature for our experiments, since the only items eligible for voice purchase at the time of testing were e-books. For this item category, the refund process is very cumbersome and can only be done via e-mail to Amazon's billing department. Therefore, we verified that results can be transferred to this feature as well only once at the end.

4.2.2 Selected Attacks to be Executed

We used the technical setup as described in Section 4.2.1 and asked one of our colleagues to register his voice as a user voice profile for Alexa, while the main researcher posed as the attacker. We chose three attack vectors we found relevant for our described attacker model, due to their low technical requirements, low budget needs, and swift feasibility:

- *Mimic Attack:* Mimic a user's voice without technical equipment
- *Replay Attack:* Registering and replaying a user interaction using a smartphone
- *Synthesizing Attack:* Synthesize a registered voice using freely available software

Mimic Attack We started by testing if it is possible to imitate the voice of the registered user without technical support. After listening to the user invoking Alexa several times, the principal researcher tried to mimic the phrase. To see how difficult the attack was with a user of different gender, we registered a female user with an audibly higher-pitched voice. We also registered a deeper-pitched, male voice from a third colleague. During our experiments, we noticed that it was sufficient to mimic the wake word (i.e., *"Alexa"*) to trick the voice recognition. We, therefore, suspect that Alexa only uses this wake word to match a user to a voice profile, while subsequent commands are not checked.

Replay Attack We performed a *Replay Attack* as described in Section 4.1.1 by recording a user saying *"Alexa, who am I?"* using a smartphone and a built-in audio recorder. We then used the same smartphone to replay the recorded voice sample, and Alexa verified the targeted user. Next, we tested our hypothesis that only the wake word is used to identify the user. We prepared recordings of a registered user saying only the wake word. We then replayed these recordings to Alexa using a smartphone, followed by a command spoken at the moment by the attacker. Furthermore, we tested a more general approach, recording the user saying a word which contains *"Alexa"*, in our case *"Alexander"*. We edited the recording using the open-source tool *Audacity*[Aud]. We cut

out the relevant part and tuned the speed of the sample to sound more similar to the wake word.

Synthesizing Attack Finally, we tested if we can achieve similar results as with the *Mimic Attack* using synthetic voice samples. We used the openly available demo version of Google TTS API [Good]. We tested different voice types, and adjusted both speed and pitch in order to mimic the targeted voice profiles' wake word utterances.

4.2.3 Results from Executing Selected Attacks

We executed the attacks depicted in Section 4.2.2 using the experimental setup described in Section 4.2.1.

For the **Mimic Attack**, the attacker immediately succeeded on the first try. We repeated the experiment several times. The attacker always managed to pose as the user within a maximum of five tries. Note that for the initial experiment both the user and the attacker were male. We obtained similar results for the female and deeper-pitched voice targets. The attacker managed to be identified as those users again within a maximum of five tries.

For the **Replay Attack**, we successfully authenticated as the targeted user, using a recording of only the wake word, confirming our hypothesis. Therefore, we were able to replicate the findings made by Zhang et al. [Zha+17] and Sugawara et al. [Sug+20] on the Alexa platform. We obtained similar results with the modified version using a recording of a longer word and cutting out relevant parts to form the wake word, which was successful throughout several repetitions.

Finally, for the **Synthesizing Attack**, we were able to artificially generate wake words for one of our registered voice profiles (male, higher-pitched voice), while it was unsuccessful for the remaining targets. However, we expect this attack to become more successful as more voices get added to the API and a wider variety of adjustment options becomes available.

4.3 Implications for Smart Home Assistant Authentication

From the attacks we described in Section 4.1 and Section 4.2, we can deduce several implications for authentication systems running on smart home assistants. Two essential security properties for these systems might not be as strong as assumed:

Attacks only from inside Smart home assistants are typically located inside a user's house. This boundary can be seen as the first layer of defence, in that an attacker needs to penetrate it first, before being able to interact with the system. For this, we assume no trivial gateways such as an open window or door are present. The *LightCommands* attack

presented in Section 4.1 demonstrates the feasibility of malicious interaction from outside a house. All that is needed for this attack to succeed is a direct line of sight. Smart home assistants are often kept in open spaces such as the living room and can remain turned on, even when no user is at home or during the night. Yuan et al. [Yua+18] highlight the feasibility of attacks through speaker equipped devices such as TVs, radios or laptops. Again, an attacker does not need physical access to a smart home assistant to interact with it. Finally, a trivial attack from the outside could be an attacker shouting or using loudspeakers to send voice commands to a device through closed windows or walls.

Users notice attacks when present While a person can trivially detect a straightforward attack involving shouted voice commands while near the smart home assistant, both *LightCommands* and *CommanderSong* can be performed stealthily. Zhang et al. [Zha+17] demonstrated that inaudible command injection is possible. Therefore, a user might not notice a malicious interaction is taking place until after the smart home assistant executed the command. Depending on the scenario, the damage might not be reversible at that point, e.g., using Alexa to open a garage door.

As highlighted in Section 4.2, smart home assistants can be protected by different kinds of authentication mechanisms. The voice-entered PINs suffer from similar weaknesses as typed-in PINs. Shoulder surfing is a classical eavesdropping attack on PINs on mobile phones and personal computers, see Roth, Richter, and Freidinger [RRF04]. Eavesdropping on voice PINs can be as easy as being present while a user interacts with the system. Mitev, Miettinen, and Sadeghi [MMS19] demonstrated how a remote attacker can exploit captured IoT devices to listen in on users over the Internet. Sugawara et al. [Sug+20] describe how an attacker can use lasers to eavesdrop on people from a distance, e.g., by picking up vibrations on windows, created by sound waves.

Regarding voice-based biometric authentication, we presented several attacks that can successfully bypass such an authentication system by either replaying a legitimate user's voice or mimicking it using more or less technology-reliant techniques. Taking these findings into consideration, we can formalize several possible attacker models which could affect users of smart home voice assistants:

Casual attacker Due to the possibly low barrier provided by authentication, casual attacks by acquaintances of the user can be possible. Such acquaintances could gain a user's PIN as well as a recording of their voice during an interaction by merely using a smartphone located close-by the smart home assistant. It could then be possible for the attacker to carry out a *Replay Attack* while the user left the room for a short period. Since smart home assistants can be found in spaces typically open to guests (e.g., the living room), this does not require special access privileges in a user's house. This kind of attacker could be motivated by financial gains or a wish to cause inconveniences to the user (i.e., a prank). This attack vector does only require basic technical skills and minimal preparation time. Even if no recording can be made, mimicking the familiar voice of a well-known user can also lead to a successful exploit.

Targeted attacker Another possible attacker model is a targeted attacker with a high motivation to harm the user or enrich themselves. This attacker does not have to be familiar with the user nor interact with them directly. The *LightCommands* attack enables interaction with a smart home assistant from a distance. Similar laser equipment can be used to eavesdrop on a user to retrieve the PIN. Of course, eavesdropping on a user is not only useful during an interaction with the smart home assistant. Kumar et al. [Kum+18] demonstrate how arbitrary speech samples can be used to construct a wake word and, subsequently, bypass the biometric authentication. An attacker could get a hold of such samples by following the user around public places, waiting for any voice interaction with other people to occur. Depending on the targeted subject, speech samples might already be publicly available through social media websites. Such openly accessible speech samples are particularly likely for public figures such as politicians or celebrities. This attack vector comes with high demands to budget and technical knowledge of the attacker.

Broadcast attacker The last attacker model presented here is a broadcast attacker. In contrast to the previously described attacker models, the target is not a single specific user, but rather several unknown users. By spreading the attack to as many targets as possible, an attacker can gain an advantage even if only a low fraction of attacks succeeds. *CommanderSong* is a technique that allows a malicious sample to be distributed over mass media such as radio, TV or Internet platforms. An attacker could construct a song used as background music in a video, and embed a command targeting arbitrary users, e.g., “Alexa, transfer $10 to evil account.” Even a sequence of commands is possible, allowing an adversary to cope with multi-step interaction required by Alexa. Regarding authentication, an attacker could count on users not having set up any authentication mechanisms. By broadcasting PINs with a high likelihood of being employed by several users or using a variety of synthesized speech samples sounding similar enough to many users, such an attack can succeed with non-negligible probability. This attack requires medium technical knowledge and medium effort, while the gains vary with the reach of the attack as well as the expected number of vulnerable devices.

Finally, we want to highlight a finding made during research on Alexa’s authentication mechanisms. The Amazon developer guideline [Amaa] states that speaker recognition should only be used for personalizing a skill. Personalization is a technique that allows developers to distinguish users inside a skill to present different contents and interaction modes according to the recognized voice profile. The guideline states explicitly that voice profiles can not be used for authentication. However, we perceive this as conflicting with Alexa’s current implementation of the online shopping feature. Under certain circumstances, as stated above, authentication depends on speaker recognition only.

CHAPTER 5

Conducting a User Study to Investigate Users' Perceptions of Smart Home Assistant Authentication

We describe the design of our user study and give details about how we conducted the study, from recruiting participants, to conducting the interviews, to analyzing the data. The purpose of our user study is to explore user mental models and threat models concerning voice-controlled smart home assistants. We are also interested in uncovering what mitigation strategies users apply to protect themselves in high-risk situations. Finally, we want to identify important aspects of the design of future authentication schemes for voice-controlled smart home assistants from the users' point-of-view.

We chose semi-structured interviews because they proved to be a useful tool for investigating users' perceptions, as was shown by e.g., Bernd, Abu-Salma, and Frik [BAF20] or Zeng, Mare, and Roesner [ZMR17]. We transcribed the interviews for data analysis, where we followed the principles of thematic analysis, described by Braun and Clarke [BC06], since it is a flexible, relatively easy-to-learn method that is adequate for researchers with little experience of qualitative research.

The research questions we answer with this study are:

- Which kind of attackers and threats are users concerned about when performing high-risk tasks via voice-controlled assistants in a smart home environment?
- Which potential mitigation strategies do users apply to protect themselves?

- Which properties does an authentication system for voice assistants need in order to be perceived as secure by users?

5.1 Study Design

As we have learned from previous work, described in Section 3.3 and Section 3.4, current authentication mechanisms are not satisfying the security and usability needs of users. These circumstances become even more significant once high-risk tasks, such as transactions or physical access control come to the voice-controlled smart assistant platform. This study aims at laying the necessary groundwork for future authentication mechanisms. As discussed in Section 2.3, users' perceptions about a system are essential for user-centered design and should be know in advance, especially when it comes to security-critical applications.

In line with comparable previous work, we use semi-structured interviews in order to describe users' perceptions of threats and mitigation strategies. Furthermore, we explore which properties are important for users to perceive a voice-based authentication system as secure and usable. Semi-structured interviews give us the flexibility to go deeper into concepts that we can barely anticipate since we do not have a pre-established theory we can rely on or test. Instead, we explore the design space of authentication schemes for voice-controlled smart home assistants, from threat models to mitigation strategies, to design ideas from the users' perspective.

5.1.1 Scenarios and Vignettes

In our study, we make use of vignettes. Vignettes are short, pictorial descriptions of situations, that can be used in interviews to encourage discussion about scenarios without actually playing them out, as described by Barter and Renold [BR99]. Reineck et al. [Rei+17] state that vignettes have an advantage over abstract survey questions in that they are closer to reality. Furthermore, they indicate that vignettes might reduce social desirability bias since interviewers can ask questions less directly.

Another decisive factor for us was that most of the examined functionalities are not available in Germany or Austria at the time of writing. Vignettes allowed us to describe the scenarios we wanted participants to immerse into, which would have been more difficult using traditional survey methods, especially in a lab environment. In order to facilitate immersion even further, we included a picture in every vignette, which is also common practice, according to Reineck et al. [Rei+17].

Several other studies have used scenarios as a tool in the security and privacy domain. Abu-Salma et al. [Abu+18] used scenarios to survey participants about a hypothetical E2E encrypted messaging tool. The authors' goal was to assess user mental models of such communication tools. Wash [Was10] presented respondents with hypothetical scenarios such as *"finding out you have a virus"* as part of a semi-structured interview. The author then added information that contradicted previously identified mental models.

For example, since it was unclear how users would feel about viruses that were created on purpose, Wash informed study participants that *"the virus in question was written by the Russian mafia"*. Krombholz et al. [Kro+19] used scenarios as the basis for their drawing tasks to evaluate user mental models. Abdi, Ramokapane, and Such [ARS19] employed scenarios during their interviews with smart home users to uncover perceived threats and other security and privacy concerns of that group.

5.1.2 Expectations for User Mental Models

In the spirit of constructivism, we want to discuss some expectations we had going into this study. We do so because we believe that, similar to Krombholz et al. [Kro+19] and Braun and Clarke [BC06], the personal views of researchers shape every part of a study, from study design to data analysis to reporting. In our case, our expectations particularly influenced the design of the scenarios.

We believe that the most influential factors for user threat models in the context of smart home voice assistants are:

- **Presence of bystanders** We expect users to perceive different kinds of bystanders as more or less of a threat. We assume differences to exist in the perception of partners, children, friends, and strangers, with every category possibly being subdivided into more fine-grained attributes.
- **Location** We anticipate different levels of perceived threat between indoor and outdoor scenarios, where users feel more at risk in the latter case.
- **Task** We expect different tasks performed via a VUI to influence users' security and privacy concerns. We assume distinctions to exist between money related actions and physical access scenarios (e.g., unlocking a door). In the former case, we anticipate the amount and nature of the transaction (e.g., active sending of money vs. passive checking of balance) to affect users' threat models.

When it comes to gaps in users' perceptions of threats, we expect people to be aware of the threats posed by people who are present during the invocation of a security-critical task. However, we expect people to neglect the threat of other IoT devices listening in on their actions. Similarly, we expect people to be aware, to a certain point, of the threat posed by people with physical access to their smart home assistant, but not of attackers able to inject commands from outside the house, as described in Section 4.1. Finally, we expect people to be skeptical about the voice code as a mean of authentication. However, we assume users have incorrect assumptions about how secure this authentication method is and against what kind of threats it can protect users.

Regarding mitigation strategies that users employ, we expect to find that people refrain from using the system at all or use security-critical features only when bystanders are not present. Some users might also make use of the whisper mode provided by Alexa to mitigate the threat of other people overhearing a sensitive conversation.

5.1.3 Scenario Design

We designed the scenarios we used for this study to cover different applications for smart home voice assistants that bring along increased security and privacy risks. We chose tasks that we assume to be of different severity when it comes to security. We also selected functionalities that are already available to a certain extent in specific markets or are likely to become available in future releases of Amazon Alexa and other voice-controlled smart home assistants. We discuss the different tasks below.

Transfer a small amount of money Transferring small amounts of money is a task available via several platforms (i.e., online banking, *PayPal*). Therefore, we expect most users to be familiar with the concept or to have done this via any platform at least once. We chose a small amount of money since we assume the perceived risk to be on the lower end for users. We set the amount to be 20 Euro precisely because several banks and credit card providers in our target area (i.e., Germany and Austria) offer contactless payments for amounts up to 25 Euro without the need to enter a PIN or other additional authentication[1][Dig]. It suffices to own the card as a type of possession-based authentication.

Pay a reoccurring bill Furthermore, we included the task of paying a reoccurring bill (in our case, the utility bill). We chose this task because it is similar to paying a credit card bill, a feature available on the Alexa platform through Capital One [Cap], amongst others. We modified it to include another party besides Amazon and the bank, namely the energy provider. We also expect this to be a more cumbersome task for people, therefore having an interest in resolving it with minimal hassle. Since the task is reoccurring, users might have an additional stimulus to automate it.

Unlock the front door Another feature already available in some markets is unlocking doors, e.g., via August [Aug]. We chose it as a representative task involving physical security, as opposed to the other tasks involving personal finances. We chose the front door because it allows us to transfer interaction outside of the house, another factor we expect to be influential for users' risk perception.

Check the transaction history As the fourth task, we selected checking a user's account history for a specific transaction. We assume this process to be the least risky from a security perspective because no actual transaction is taking place. The information requested is, however, still potentially sensitive to the user from a privacy point-of-view.

We also selected four different situations in which interaction with Alexa is taking place. These interactions vary in two aspects: the presence of other people and the

[1]Note that since the conduction of our study, this limit has been raised to 50 Euro [honl]

location, namely inside or outside the house. The situations used in the scenarios are the following:

Dinner party with friends The first situation included in our user study was a dinner party. In this scenario, the user invited several friends over to their home. This suggests that the people are, to a certain extent, familiar with each other. The interaction is happening while the people are at the table.

Watch TV with partner Next, we constructed a situation where the user is watching TV with their partner, suggesting close familiarity, while no other people are immediately present. Both are inside their home. We chose the activity of watching TV to have another potentially smart IoT device present during the interaction, without pointing this fact out in a specific way.

Come home from grocery shopping Another situation we selected was returning home from shopping groceries. In this scenario, we moved the interaction outside the house. There are no people immediately present besides the user. The user's hands are full with the grocery bags, making a hands-free interaction more appealing.

Have dirty hands due to gardening work Finally, we included a situation where the user's hands are dirty due to them working in the garden. We chose this setting as another representative of circumstances in which hands-free interaction is more desirable than traditional typing interfaces since the user's hands are dirty after just coming back inside from working in the garden. The situation includes kids running around the house, therefore, being present, while possibly not paying immediate attention.

We combined the tasks depicted above with these situations to create scenarios that are as realistic as possible for most users, possibly also familiar to some of them. We aimed at including only necessary details in order not to confuse participants and make them lose focus. We also designed the interaction with Alexa to be minimalist. Notably, we chose not to have Alexa respond to the user after accepting the voice code. We did so because we wanted to leave this part open for participants to explore. The interaction with Alexa and the request for a voice code are in line with currently available interaction patterns, e.g., when doing online shopping via the platform. We chose to use a different voice code for each scenario since we expected this to be more realistic than having the same code for each action. However, this is another topic we want to explore in this study.

You gathered some friends for a dinner party at your place. In the middle of eating you remember that you owe Kim 20€ for the lunch she paid the other day. You want to settle this right away.
You say: *„Alexa, transfer 20€ to Kim!"*
Alexa responds with: *„OK, to transfer money, tell me your voice code!"*
You: *„My code is 8915."*
Alexa accepts the code and the transaction succeeds.

Figure 5.1: Vignette for scenario "Dinner"

Scenario "Dinner" In the scenario "Dinner", depicted in Figure 5.1, we combined the task of transferring a small amount of money with the situation of a dinner party with friends. In our experience, friends are a prime target for small, fast transactions. The use of Alexa can be convenient since the user is sitting at the table, therefore, using a laptop might be undesirable. Several people are around who might overhear the user's interaction with the smart home assistant. However, these people are, to a certain point, familiar to the user, as they were invited by them. The corresponding image shows a laid table with several people around, chatting in a light atmosphere.

You are in your living room watching TV when your partner asks, if you have already paid the utility bill this month. Since you have in fact not done so yet, you decided to do it right away using your Alexa device.
You say: *"Alexa, pay the utility bill!"*
Alexa answers: *"OK, to pay it, tell me your code!"*
You: *"6858"*
Alexa accepts the code and the payment is processed.

Figure 5.2: Vignette for scenario "TV"

Scenario "TV" The people involved in the scenario "TV" are the user and their partner. We selected the activity of paying a reoccurring bill since this is a typical task that is of interest to both partners, while also being less casual and completed less frequently. The picture associated with this scenario depicts two people sitting in a living room, on a couch, in front of a running TV, see Figure 5.2.

You have just taken all your groceries out of the car and are about to take them inside. The front door is locked. Your hands are full and you don't want to put everything down again so you ask Alexa to do open it for you.
You say: *"Alexa, unlock the front door!"*
Alexa answers: *"OK, to unlock the door, tell me your voice code!"*
You: *"3071"*
Alexa confirms the code and the door is unlocked.

Figure 5.3: Vignette for scenario "Door"

Scenario "Door" Figure 5.3 shows the third scenario, "Door". For this, we combined the task of unlocking the front door with the scenario of coming back from grocery shopping. The task involved is a typical task that a user performs while outside the house. The situation includes the user handling several bags, therefore, being unable to handle a key. This setting should justify the use of a voice-controlled smart home assistant. No other people are immediately present in the scene. The picture shows a person carrying bags of groceries next to a car, a blue sky in the background indicates that the scene takes place outside.

You just came back from working in the garden. Your kids run around the house screaming. They are already very excited for the upcoming school trip. That's when the question comes to your mind: have you already paid for that? You want to check if the transaction is there in your online-banking.

Figure 5.4: Vignette for scenario "Hands"

Scenario "Hands" For the last scenario, "Hands", we coupled the task of checking a transaction history with the situation of having dirty hands due to previous gardening work. We included children in this scenario as potential bystanders. We described them as running around, screaming, meaning they do not pay immediate attention to the user while still being present. Also, this scenario does not feature any dialog with Amazon Alexa. We do not refer to Alexa in this description, nor include a device in the image. We did so to leave room for the interviewee to imagine how an interaction could play out. At the same time, this allows us to explore alternative interaction mechanisms, potentially not involving Alexa. Figure 5.4 displays the combined vignette.

5.1.4 Interview Guideline

We embedded the scenarios described in the previous section into semi-structured interviews. We designed a guideline for those interviews, providing a loose structure, and outlining the questions we asked. Since we were carrying out semi-structured interviews, we did not follow this guideline with complete accuracy or asked questions precisely as described here. As the interview proceeded and participants started to share their thoughts, we drifted off this guideline and explored interesting topics more in-depth, asking suitable follow-up questions. At some point, we got back to the guideline only to leave it again at a later point, following more topics we deemed worthy as the interviewees brought them up. The final version of the interview guideline is in Appendix A.1.

The guideline is divided into three parts. First, we included some questions that introduced the topic to the participants in order to prepare them for the interview. These questions mostly revolved around participants' Alexa usage in general, e.g., we asked participants how long they already know about Alexa, on which devices they had previously used Alexa, and what some typical tasks were they used Alexa for. We can also use the data obtained in this way in the analysis and to describe our sample. The final question of the first block was asking participants whether they have used the online shopping feature of Amazon Alexa before. If participants had done so, we asked whether they encountered any problems in the process. Otherwise, we asked participants for reasons why they have not made use of this feature yet. This question sets the ground for our further questions as it already includes a security-critical task with which participants might be familiar. The task also may include authentication steps[2]. We chose to include the last question in the data analysis since several interesting topics were already brought up at this point, e.g., participants stating reasons why they did not use online shopping on Alexa.

In the second part of our interview guideline, we presented participants with the scenarios described in Section 5.1.3. After letting them read the description text and look at the image, we asked participants which problems they think could arise in such a situation. We did not ask about security-related problems in order to not prime participants in a specific direction. Only if no security-related problems were talked about at all, we asked more directly, e.g., whether they thought the voice code included in the scenario was useful or not. We explored problems we deemed interesting more in-depth by asking why interviewees think that was problematic. Afterward, we explored threats that participants were able to identify and asked them if they could think of any other actors that might pose a threat in such a situation. Lastly, we asked participants about mitigation strategies. We investigated what interviewees thought might be useful to them and which mitigation strategies they would apply in the given scenario, with the threats described above in mind. We also explored changes to the authentication procedure more in-depth, if participants talked about any alternative authentication method they would prefer. This process was repeated for each of the four scenarios. Afterward, we asked participants

[2]The default settings of Amazon Alexa allow online shopping without any authentication. However, it is possible to activate authentication by voice code. It is, furthermore, possible to allow recognized voice profiles to place orders without saying a voice code [Amae; Amad]

to sum up all four situations and to think about possible similarities and differences between the scenarios. We included this question to make participants recapitulate all scenarios, possibly applying the knowledge they gained from a following scenario to those seen earlier. It also helps participants to focus on details which they might not have noticed before (e.g., the number of people present in the scenario) and think about the consequences introduced by said factors.

In the third and final part of our interview guideline, we included some demographic questions, mostly used to describe our sample. Including demographic questions at the end helped us minimizing participants' fatigue typically encountered after longer interview sessions, as these kinds of questions can usually be answered without noteworthy mental effort. We included two standardized scales in this section, namely the affinity for technology interaction (ATI) scale by Franke, Attig, and Wessel [FAW19] and the concerns for information privacy (CFIP) scale by Smith, Milberg, and Burke [SMB96]. The former is designed to assess a person's tendency to engage in intensive technology interaction actively. The later is an instrument for measuring an individual's privacy concerns, mostly regarding the privacy practices of larger organizations. For the ATI, an official German version is available on the authors' website [FAW], for the CFIP, we used a translated version provided by Harborth and Pape [HP18]. Of course, for the interviews conducted in English, we used the original version of both scales. We further included standard demographic questions such as age, gender, and level of education. We asked participants how many people live in their household and how many of those use Amazon Alexa. This data can help us gain more insights into the ecosystem of voice-controlled smart home assistants in practice, e.g., if the same Amazon Alexa device is shared among multiple users.

After completing the interview, we asked participants whether they had any questions remaining, we reiterated the purpose of the study and why the interview guideline and the vignettes were designed as presented. To the best of our knowledge, we also tried to explain the current situation regarding security, and most of all, authentication, on Amazon Alexa, and comparable VUIs. We also informed participants how we planned to proceed further with compensation and how they could get in touch with us if they wanted to be informed about the study's final results.

5.1.5 Interview Procedure

For the on-site interviews, we made an appointment with the participants via e-mail for a 1-hour time slot. Afterward, we invited them to our research facility, where we booked a suitable conference room to interview in. In the beginning, we greeted the interviewees and offered a glass of water before asking them if they had any questions they wanted to discuss before the interview. Next, we handed over the information sheet for the study explaining the essential topic of the study, stating participants' rights, giving details about their compensation, and asking consent for the gathering and processing of participants' data. We asked participants whether they are okay with us making audio recordings of the interview and explained how those are going to be processed and used for analysis

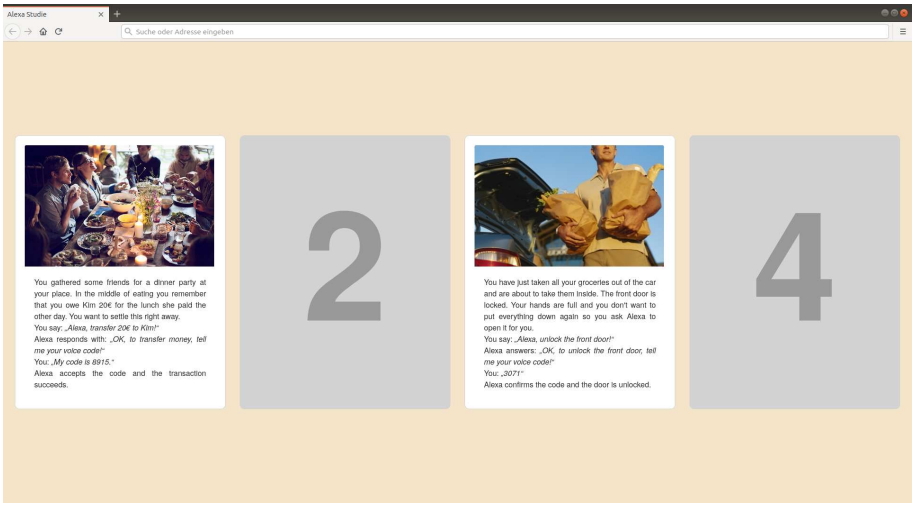

Figure 5.5: Website used for remote interviews. Two scenario cards have already been flipped over, while two are face down. By clicking on a card, a flip animation is shown and a large version is displayed until a closing button is pressed.

and when their data will be deleted after completion of the study. Participants received a copy of this information sheet, including contact details of the principal researcher, which they could use if any further questions arise after the session or if they want to withdraw consent.

For the remote interviews, the process was similar. We agreed on a time-slot with the participants and asked them to find a quiet location with a stable Internet connection. Furthermore, we required them to use a computer, laptop, or tablet with a minimum screen size of 10 inches, in order to be able to see our vignettes well enough. We sent them a digital version of the information sheet which they signed and returned a copy back to us. We recorded the remote interviews with the same recording device as the on-site interview to evaluate data uniformly. We left it up to participants whether they wanted to turn on their camera or not, while we always used that feature to engage with interviewees on a more personal level.

During the interviews, we provided on-site participants with pen and paper, allowing them to express themselves also via drawing or writing. Remote interviewees could send us any graphical artifacts they produced via e-mail. We followed through with the interview according to our guideline, as described in the previous section. For on-site interviews, we printed each vignette onto a size A3 laminated paper card and presented them to the interviewees in random order. We shuffled the cards face down and let participants choose one. After asking the relevant questions from the guideline, we put aside the card and repeated it until participants had talked about each scenario. However, we explained to participants that they could go back to a previously seen vignette if they wished to add something. For the remote interviews, we aimed at giving interviewees a similar

experience. We constructed a web page displaying four face-down cards numbered 1-4, which would turn around and show the face side full-screen when clicked on. Figure 5.5 shows an example screenshot. Again we let participants choose the order of scenarios.

During the conduction of our study, we had the possibility to interview a blind Amazon Alexa user who contacted us through one of our recruiting channels. We adapted our interview material to enable participation without the need to see and read the vignettes. The main adaption was switching out the printed vignette cards with audio recordings. In these audio recordings, we first described the picture displayed on top of the card. Afterward, we read the corresponding text. We assumed this to be the most natural order of examining our vignette cards. Both descriptions were made into separate audio files, leaving us with two files per scenario. Doing so allowed us to replay each description individually if the participant wished to listen to it again. The audio recordings were narrated by a different researcher from the one conducting the interviews to have a clear distinction between vignette and interview questions. For the parts of Amazon Alexa, we used a computer-generated voice similar to the original Alexa voice. Note that data obtained from this interview was processed and evaluated in the same manner as other interview data.

At the end of each on-site interview, we handed participants a tablet on which they could fill out a short questionnaire composed of demographic questions, as described in Section 5.1.4. We implemented the questionnaire using the Google Docs Form[Goob] feature. For remote interviews, we sent participants a link to the form at this point.

After each interview, we stopped the audio recording and asked participants if they had any questions about the study or our work in general. We explained to participants how we planned to evaluate our study and how we intended to publish the results. Several participants expressed the wish to see the outcome of our research and provided us with contact details to send them the finished paper. Approximately two months after the interviews, we sent participants the compensation in the form of a digital voucher via e-mail and asked them to sign a confirmation of receipt.

5.1.6 Pilot Testing

We pilot tested our study design by conducting four interviews with participants we convenience sampled from our circle of acquaintances, among those two active users of voice-controlled smart home assistants (Amazon Alexa and Apple Siri). We did two interviews each for the on-site and remote setup. Furthermore, we asked colleagues from our research facility for feedback. We also used the pilot interviews to get a good estimation of how long interviews are going to take. Since most pilot interviews took around 40 minutes, we decided to use 1-hour time-slots for each participant, accounting for pre- and post-interview procedures. After each pilot interview, we asked participants whether anything was unclear to them or if specific parts of the interview left them confused or unsure how to respond. We also asked them about their personal opinion of the interview, what they liked, and disliked.

As a result of these pilot tests, we made minor modifications to the wording of the scenario descriptions and interview questions. One question was dropped from the interview guideline as most participants did not understand the question or did not know how they should answer it. We cut another question because we deemed it insignificant to our research questions, thus unnecessarily protracting the interview.

We made significant adaptions to the "Door" vignette. We changed the scenario from using Amazon Alexa to open a garage door to the current version of opening the front door. We chose to do so based on feedback received during pilot testing. It became clear that people might have different understandings of the implications of an attacker gaining access to their garage door, mostly based on the garage setting they are familiar with. Some garages might have direct access to an apartment while others are not connected to the main building.

Furthermore, participants had different settings in mind regarding how public this scenario is, i.e., how easily others might listen in on their conversation. This circumstance was due to different understandings of where the interaction occurs, with some users expecting to be already inside their garage when accessing Alexa rather than outside in a public space. To reflect that critical aspect of our scenario better, we changed the interaction to opening the front door and adapted the image to partially show a blue sky, making it more evident that the person in the scenario was still outside. These changes were introduced before the fourth and final pilot interview, after which we deemed no further changes necessary.

We also examined our tool set for this study, such as the voice-recorder, the vignette cards, and the website. Both participants and colleagues gave us very positive feedback on the overall scenario design. Interviewees from the pilot tests stated that working with the presented materials was pleasant. We checked both audio quality and ease of use of our recording setup for both on-site and remote interviews and deemed it sufficient.

5.2 Recruitment

We expected the number of Alexa users in the areas we conducted our study to be relatively low. Therefore, we implemented various recruiting strategies in order to get participants for our interviews. As described before, we conducted both in-person interviews as well as remote interviews over video chat. We recruited participants for in-person interviews via flyers posted all around campus of Saarland University. We made appointments via e-mail. Ultimately, we recruited five participants via this channel, all of which were based in Saarland. For the remote interviews, we recruited people by advertising the study through relevant mailing lists of Saarland University and TU Wien. Three Alexa users participated in our study after responding to these posts, two based in Saarland and one based in Vienna. We also applied convenience sampling (i.e., recruiting *friends of friends*), which yielded another six participants, of which two were based in Saarland, three in Vienna, and one in South Tyrol, Italy. Last, we also asked participants after each interview if they knew other potential participants (i.e., Amazon

Alexa users), a method known as *snowball sampling*, as described by Lazar, Feng, and Hochheiser [LFH17]. We expected this to be another sound strategy since users of smart home assistants possibly got to know about the system from social acquaintances who already have one. This recruitment channel yielded two participants, both based in Vienna.

Participants could choose whether they wanted to have the interview in German or English. Thirteen went for German, three participants preferred to talk English. We were able to recruit seven female and nine male participants. Our participants' age ranged from 18 to 55, with a mean of 29.31 ($\sigma = 10.69$, median = 26.5). Fourteen participants had at least completed high school, with seven holding a bachelor's degree (or equivalent) and two holding a master's degree. We furthermore investigated our participants' tendency to actively engage in intensive technology interaction using the ATI scale consisting of six-point Likert items labeled *(1) "completely disagree", (2) "largely disagree", (3) "slightly disagree", (4) "slightly agree", (5) "largely agree"* and (6) *"completely agree"*. The mean result was 4.1 ($\sigma = 0.76$, median = 3.83), individual participant's scores ranged from 3 to 5.33. To assess people's privacy concerns, especially when dealing with large companies such as, in the presented case, Amazon, we used the CFIP scale. This scale consists of 15 seven-point Likert items anchored with *(1) "strongly disagree"* and *(7) "strongly agree"*. The overall mean was 5.76 ($\sigma = 0.71$, median = 5.93). A complete overview of the demographic data of our study's participants is displayed in Table 5.1.

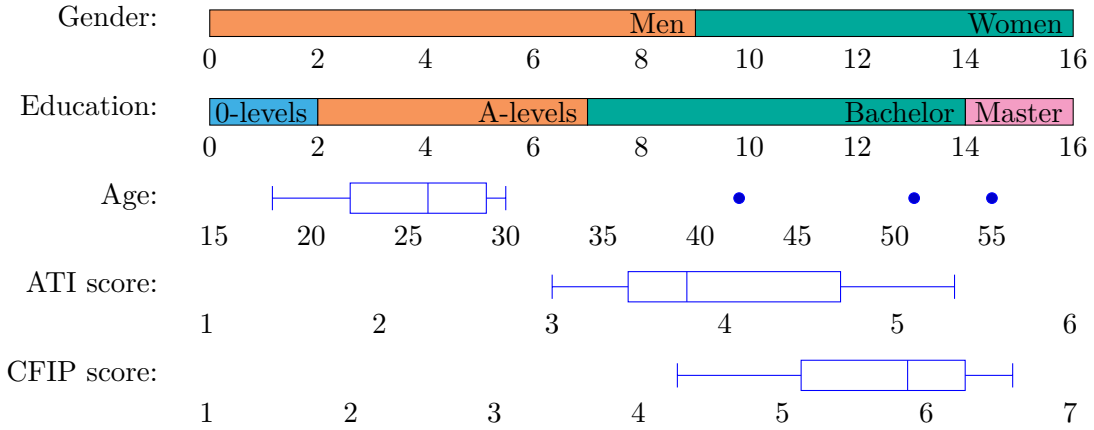

Table 5.1: Results of the demographic questionnaire: An ATI score above 3.5 indicates an above average affinity for technology interaction with higher values indicating stronger affinity, see Franke, Attig, and Wessel [FAW19]. For the CFIP scale, higher scores indicate that a subject is more concerned about their information privacy with regard to organizational privacy practices, see Smith, Milberg, and Burke [SMB96]

5.3 Data Analysis

We chose to analyze our data using a thematic analysis approach, as described by Braun and Clarke [BC06]. They argue that this technique is widely used in a variety of fields. However, different terms are used, and several variations exist, as the method provides researchers with greater freedom than, e.g., grounded theory, as Braun and Clarke state. The authors describe thematic analysis as "a method for identifying, analyzing, and reporting patterns (themes) within data."

As a first step, we familiarized ourselves with the data obtained from the interviews. This process started when transcribing the interviews. The leading researcher produced transcriptions with the help of a colleague. We chose to transcribe the interviews ourselves rather than use an external service to get an initial feeling about the data. We transcribed the data at an orthographic level, including non-verbal utterances only when we deemed them essential for the semantic of a phrase (e.g., a participant laughing while saying something, indicating it was a joke or not meant in all seriousness). We did not transcribe other features such as pauses, non-verbal clues such as nervous fidgeting or occurrences like coughs. Afterward, we read and re-read the data to get an even better understanding, taking notes of interesting details and basic patterns.

The notes we took in the first phase served as the groundwork for the open coding that followed. We started coding the interviews one after the other. Since the involved researchers spoke German as their first language, we decided to code in that language and later translate codes for the final report. We did not translate interviews conducted in English, but coded them with the same codes as German interviews. The first pass of open coding looked as follows:

- Researcher A marked quotes and handed the document over to researcher B together with the current codebook used.
- Both researchers coded the document independently from each other.
- We merged the coded documents and calculated the inter-coder agreement based on Krippendorff's alpha [Kri11].
- We discussed potential mismatches and the current code structure. Code renamings and merges only happened as consequences of these discussions.
- We re-coded the interview, if necessary.

We repeated this process for a quarter of all interviews (i.e., four). In this phase, we started coding with a more descriptive approach, Braun and Clarke use the term *"bottom-up"*, being broad in terms of aspects coded. As we progressed, the approach shifted more towards a deductive way as we focused more on coding for our research questions. We did so primarily for RQ2 and RQ3.

After the first round of open coding, we performed axial coding, grouping codes into categories. In this stage, we reduced our codebook significantly, merging and splitting codes that were not crucial for our research questions. The Krippendorff's alpha after this step was 0.941, indicating a high agreement among the coders. We also started searching for themes. We used a graphic tool [Lat] to lay out all our codes. We then tried to collate codes together, forming categories and clustering categories according to their semantics.

We tested the resulting codebook by independently coding another four interviews. For those interviews, Krippendorff's alpha was 0.83, indicating a strong agreement between the two coders. We held a final discussion about the codebook and preliminary themes we identified during the second pass of coding.

Finally, the leading researcher coded the remaining half of the interviews using the codebook agreed upon in the previous discussion. At this stage, we introduced new codes very sparsely, since the existing codebook covered most aspects of the data. After this final pass of coding, the themes identified were revisited and adapted if necessary. We also checked back with all the data to review if themes were consistent with the whole data set and coherent internally. Where necessary, we re-coded extracts of data, especially if significant aspects of our themes had been missed in a previous pass of coding.

5.4 Ethical Considerations

Dealing with human research subjects entails taking into consideration the ethical aspects of the study. We designed our study in such a way that we can burden participants in the least possible way. We followed the principle of data economy when it comes to collecting personal data, i.e., we aimed at only collecting data necessary to answer our research questions. Regarding the storage and processing of said data, we followed legal guidelines from countries we conducted our research in as well as ethical research guidelines from TU Wien [Wie07].

We informed participants about the data we planned to collect, how we planned to process it and when we were going to delete it as part of a consent form handed out and signed by each participant before each interview. Furthermore, the consent sheet included information about the topic of the study, legal disclaimers about participants' rights, as well as information on how participants would be compensated for their time and efforts. Participation was voluntary in any case, including termination of the interview without giving reasons. Interviewees were also able to resign from answering questions if they did not want to respond. We provided the participants with contact details of the primary researcher, whom they could contact regarding any questions arising after the interview session, and also if they wanted to withdraw their consent to the processing of their data. This procedure is in line with the recommendations given by Lazar, Feng, and Hochheiser [LFH17].

We compensated participants with 15 Euro per interview in the form of an Amazon

voucher. We chose this method since we were interested in interviewing users of Amazon Alexa, for which an Amazon account is needed to set up the device. The amount of compensation is in line with similar studies, e.g., [Kro+19; ZMR17], considering the time spent by the participants as well as the specific requirements we had while recruiting (i.e., Amazon Alexa users only).

Our study design and consent sheet got approved by Saarland University's ethical review board (ERB). We chose to do so since most of our work is conducted on the university's campus.

CHAPTER 6

Exploring the Design Space of Secure and Usable Smart Home Assistant Authentication

We conducted our study and evaluated the data, as described in Chapter 5. We now present the findings made while exploring the design space of smart home assistant authentication. When analyzing our data, we focused on answering our research questions stated in Section 1.2.

First, we cover the users' perception of threat. Users were concerned about different attackers that could affect them in the presented scenarios. They also talked about trust in certain groups of people or entities. We describe a variety of factors that play into users' perception of threat and security.

Next, we report the various mitigation strategies users were coming up with in order to protect themselves. These mitigation strategies can provide us with a better understanding of how systems are used in practice and which practices users adopt to mitigate the threats that concern them. Previous work has shown that certain mitigation strategies might not be ideal or expose users to even greater danger if they are based on incomplete or wrong assumptions about the system, see Section 3.1.

Finally, we present essential properties for secure and usable smart home assistant authentication we discovered during our data analysis. These properties can help users to better adapt to a new system. By taking into account users' expectations, several human errors can be prevented, making the system both more user friendly and secure, as highlighted in Section 2.3.

The majority of our interviews (13/16) were conducted in German. Also, we coded the data with German codes. For this report, we translated all codes and relevant quotes

Few	< 4
Some	$4 - 7$
Several	$7 - 9$
Most	$9 - 12$
Almost all	$13 - 15$
All	16

Table 6.1: Approximate number of people marked by each quantifier

to English, if necessary. Participants are labeled P1-16. Note that throughout this work, we refrain from explicitly indicating the number of participants talking about each concept during our interviews. Instead, we use vague quantifiers to give an estimation of the prevalence of a theme. Braun and Clarke [BC06] discuss this issue and state that *"thematic analysis [...] does not provide a quantified measure."* In order to make our analysis process more transparent, we provide approximate numbers of participants that were represented by a certain quantifier in Table 6.1. Furthermore, codes depicted in Table 6.2 and Table 6.3, as well as the full codebook in Table A.1 are sorted alphabetically and include the overall number of quotes associated with each code. As stated by Braun and Clarke [BC06], a higher frequency does not indicate a greater importance of a concept and vice versa.

6.1 Perception of Threats

Users brought up different types of actors and threats when we asked them about potential problems they could encounter in the given scenarios. We describe these threat models of users on a per-code basis. Table 6.2 lists all codes assigned to the category *Attackers and Threats* with the total number of occurrences throughout the interviews.

6.1.1 Amazon as a Threat

The most discussed threat actor in our interviews was Amazon. Participants' primary concern was that they would share personal data with the company operating Alexa. Such personal data includes, but was not limited to, account balance, credit score, and power usage. Either these data would be given directly to Alexa as part of the interaction, or Amazon could analyze metadata, such as frequency of interaction with a particular skill, and derive information valuable to them. Participants were concerned that this privacy infringement could lead to targeted advertisements or worse deals for users. Some interviewees also suspected that Amazon could sell their data to third parties. One thing users were willing to do in this context was paying for Amazon internal services (e.g., streaming service subscriptions) since it is all part of the same ecosystem and no new information could be gained from such a transaction. P9 states: *"Of course using it to*

Code	Frequency
Accidents as threat	49
Amazon listening in on conversations	8
Bystanders as threat	51
Criminals as threat	42
Cyberattacks as threat	40
Insiders as threat	39
Malicious skills as threat	2
Pranks as threat	19
Sharing data with Amazon undesirable	95

Table 6.2: Final codes in the category *Attackers and Threats*

pay Amazon services, like say I needed to pay for Amazon Prime or pay for a Prime video or pay my Amazon card. I could see that being easy cause it's all build into the same system, I don't know." While most participants talked about Amazon as a threat actor when talking about scenarios involving personal finances, a few users also considered it undesirable giving a company direct control over their front door and, therefore, access control over their home. P8 explains their view as follows: *"I don't know if I would be so excited if Amazon had this access code to my house. I have the feeling that I would give too many permissions to Alexa. Just the thought that Alexa then theoretically has everything to open my front door, that doesn't sound like a great idea to me."*

Furthermore, participants highlighted that Amazon could suffer a data breach and leak personal data to adversaries. Therefore, users preferred not to share this data in the first place, since most believed Amazon would store customer data from voice interaction for an extended period. In general, users perceived Alexa as being more susceptible to leaks compared to smartphones. P6 describes an ideal interaction as follows: *"[T]he perfect transaction would be as if Alexa weren't there. This means that nothing is lost in terms of data with Alexa, as if we were standing at the bank counter or, for example, as if we were doing it ourselves via an app on the smartphone, on the device itself."* This perception was in part founded on participants' awareness that all interaction data passing through an Alexa device t to the cloud to perform the STT and TTS conversions. Therefore, even when using a third-party skill, data is still accessible by Amazon. P16 states in this context: *"So you have the skill that calls it up and Alexa vocalizes it, and with this vocalization, the data is there [accessible] again."*

Some participants also stated that Amazon could potentially record a user's voice code. This code could then be used to harm the user, e.g., by making transactions from a user's account. Again, this concern was based on the perception that Amazon stores all

processed data in the cloud. A few participants also suspected Amazon employees to occasionally listen to users' conversations as part of a quality assurance process. P16 explains that: *"[I]f an employee comes across this conversation and evaluates it because something special has happened there somehow or some flags are raised in the background, and then comes across this code, he could theoretically use it to harm me."*

6.1.2 Bystanders as Threat

A primary concern of users overall was that other people might be able to overhear them interacting with Alexa. Users brought up several kinds of bystanders, who might be close-by in certain situations and accidentally or deliberately pick up the voice code. Bystanders could listen in on conversations through an open window or even thin walls. This occurrence can happen unconsciously to the user, as P2 states: *"[W]e don't know who is hearing us at that time, so the dinner is happening somewhere in the ground floor and somebody is outside listening probably. It can happen, there is a possibility."* Also, unfamiliar people inside the house could be a threat, e.g., craftsmen or plumbers. P3 provides an example of how such a scenario could take place: *"Anyone who is in the house, who does not visit the house regularly like a craftsman or something, or I don't know, who ever is to visit, say the landlord is visiting and you may be in separate rooms for a moment, after this code was used, so they [...] could do something."*

However, people were mostly concerned about bystanders when the interaction was taking place outside, as was the case for the "Door" scenario. Due to it being a more open, public space, it was more likely for people casually passing by (P14: *"could be anybody that's riding down the street, one their bicycle or taking a walk or yeah, could be anybody"*) or deliberately eavesdropping on the user (P4: *"Well imagine you are in front of your door and the neighbour, who always wanted your PlayStation, is there in front of their door until you say your code"*). These examples highlight that participants were in disagreement about the motivations of such potential attackers. While a few brought up specifically monetary gains, most were unsure how a bystander would act upon learning the code. P11 explained how the threat of bystanders outside the home could be especially relevant for blind users: *"As a blind person, I see the danger that someone else hears it, because someone could stand one meter next to me without me noticing. Because when I'm at home and it's, e.g., as in [the 'TV' scenario where] I would pay the bill, then I know who is there, so I don't know, that my partner is there or that I am alone, that now nobody else is so close to me and hears that."*

Almost all participants described neighbours as a potential bystander threat. They can be close-by at all times, and participants noted them in all four scenarios. Especially for neighbours, we noticed different levels of concern depending on the residence of the user. People living on the countryside considered neighbours less of a threat than those living in a city. This perception was both due to larger distances between housing units and stronger mutual trust. P15 summarises this: *"Since we come from the village, you have built trust in your neighbours, I don't think anything could happen here, but like in*

some cities, if I have a neighbour I don't like and he wants to prank me, then I see the neighbours in such cities as a problem."

6.1.3 Insiders as Threat

Besides unacquainted bystanders, participants also perceived close acquaintances, insiders, as a threat. Such insiders are familiar with the user to a varying extent and might have access to private areas of the user. We discovered several conflicting perceptions when it comes to insiders as a threat. Friends were a group of insiders that most participants trusted in general. Some participants did not view friends as a threat, e.g., in the "Dinner" scenario. P2 stated that *"it's really not a big deal, I would say, if people are present, because here [in the 'Dinner' scenario] only they are present, and they are friends and family so I don't think I would feel less secure"*, however, the same participant stated that *"even if friends are present I don't trust them with my door, my code."* Several participants shared the opinion that, although there exists a bond of trust between them and their friends, this does not extend to security- and privacy-related affairs, e.g., P7 explained that *"I trust my friends, but not with my money."*

An additional issue with friends was expressed by several users, namely company that friends might bring along. This group came up, especially when talking about the "Dinner" scenario. Participants explained that friends might bring along company less familiar to the user, e.g., their partners. When asked about different levels of trust within their circle of friends, P3 responded: *"Well, it wouldn't be so bad with my closest friends, it's normal that you trust your closest circle of friends more, but that when you have a bigger party and then friends of friends or new friends just come along or new partners or their children, or anyone else, you don't know well, then I find it difficult to do [a transaction] by voice control when everyone is seated at the table."* Mostly student participants brought up flatmates as a potential threat. They share several aspects with friends, in that they can have access to a user's apartment and might be present during day-to-day interaction with a smart home assistant. The level of trust towards flatmates appeared to be generally lower than towards friends.

Another type of acquaintance enjoying a more extensive amount of trust is a partner. Most users explained that they trust their partners, even going as far as sharing their authentication codes with them. This practice might have practical reasons, as many participants noted for the "TV" scenario. P5 states: *"I assume that you will have agreed upon that. You live together and you pay electricity bills together so that you share stuff anyway, e.g., this voice code. I would do that too. So I'd think that's okay if two people knew that."* In contrast to this complete trust, other participants stated that they do not want to share sensitive information with partners. P16 points out that *"of course, even if it's the partner who enjoys a special trust status, it's still a problem in terms of security."* This problem can become even more significant if the relationship terminates. A few participants voiced that a previous partner might become a threat, should there be bad blood between them.

Several participants talked about their family as a fully trusted entity. This group included most often parents, siblings and kids. The latter, however, were suspected of causing trouble, depending mostly on their age. Primarily young teenagers were perceived as a threat, in that they could misuse the system. P4 explains that *"Children are usually quite bright and register everything, they absorb everything, soak everything up like a sponge and then I think they could use that somehow, the voice code, [to] make transfers, top up the cell phone. They could get pretty good access to the account."*

6.1.4 Pranks as Threat

When talking about the motivation of insider attackers, several participants explained that they would expect their friends or kids to play pranks on them, using Alexa. Such a prank could be, e.g., setting the alarm for the middle of the night. Participants described that they had made similar experiences with other technology in the past, and played such pranks to their friends also themselves. P6 stated that they currently did not use the online shopping feature of Alexa due to security concerns. When we asked about potential threats they had in mind, the answer was: *"I don't even think anyone would do that maliciously, but I remember, I think that was two or three years ago, that I and a few colleagues went to a friend's house who also owns an Alexa. And that's when we discovered that this shopping list tool works. And then we put all sorts of nonsense on the shopping list. And so just the general worry that you could buy things using voice commands. And then next time, when I host a party, three days later I would get all sorts of nonsense delivered to my doorstep."* Similar worries were expressed by some participants about their kids, who could shop online, without malicious intentions.

6.1.5 Criminals as Threat

While harm caused by pranks might be limited, participants were also concerned about the potentially more harmful threat of criminals. This threat actor was brought up primarily when participants talked about the scenario "Door", which involves using Alexa to unlock the front door. Almost all participants described burglars as a threat. They expected the front door to be a prime target and Alexa to be an easier entry point than traditional door locks. In this context, some users highlighted that reconnaissance of Alexa was easy. Burglars could easily find out if a device was active, as P7 states *"I mean you can easily find out whether Alexa is on the doorstep or not. Just say 'Alexa' once."* Furthermore, adversaries could also find out which skills are activated on Alexa by trying out various commands.

Again, the biggest fear of participants was that criminals could find out their voice code and use it to bypass the authentication. This attack could be executed by simply eavesdropping on a user, like P9 describes: *"[T]here could be somebody parking across the street just listening, waiting, because they know you have an Alexa, they know you do this, just waiting for that PIN, so that they can just come in, tomorrow when you're at work, and have their way with your things, which is not a very drastic threat, or likely threat,*

more than likely this would not happen, but there is always the chance, nobody ever thinks things will happen to them." A few participants also described more technical methods to capture the PIN. Burglars could use equipment such as microphones or cameras and strategically place them near the users' door. Such devices could operate stealthily if hidden, e.g., inside a bush. Also, drones were noted in this context. P4 describes how a *Replay attack* could play out: *"[A] burglar can just place a camera somewhere or a hidden microphone or something and then have my voice, which can then say to Alexa: unlock the front door."*

In addition to this basic attack, a more sophisticated variation is the voice sampling attack, where a phrase is composed of single utterances collected over a whole conversation or multiple recordings. P9 describes such an attack as follows: *"[S]o this whole interview, I'm sure you could make millions of phrases of what I said today basically just take each word, chop it, put it together, and I don't know if the Alexa software is intelligent enough on the voice profile to be like: this is not one flowing sentence, this is multiple statements put together into a sentence."* Furthermore, participants were aware of low-entropy voice codes being susceptible to brute-force attacks. P2 explains this as follows: *"For example, if say really smart bugger, so they can get at least this information that this door opens with a 4-digit code, and then they can keep on trying until it opens."*

A few participants also brought up criminals as a threat in the scenarios that involve access to the user's bank account. Criminals could execute unauthorized transactions from the account. However, most participants expected to be protected, to a certain degree, by consumers laws. P7 explained their views as follows: *"I mean, when it comes to costs, I can still say, I'll block the account now if something strange happens and I lose track and there are some transfers abroad or whatever, where I can't check where they're coming from or why they're there now. Then I say, well, I'm going to block my account or my card and then I have secured it."*

6.1.6 Cyberattacks as Threat

Another threat talked about by almost all participants were cyberattacks or hackers. Participants expected adversaries to be able to interact maliciously with their Alexa device over the Internet. Several participants stated that news articles about cyberattacks made them aware of this threat, e.g., P2: *"So there was news about people that can use Alexa, this is how they can breach inside your home, right. They get into the device and then into the Wi-Fi and then what is connected to that. Because sometimes people use Alexa as, say, their security camera is linked to Alexa, lights and electricity."* Hence, Alexa can be used as an entry point to the smart home ecosystem. Also, other participants explained that, due to Alexa, systems that were not at risk of cyberattacks before can now become so. An example given was the front door, which is controlled via Alexa in one of the scenarios. P7 gives an example of this: *"By making my door electronically accessible through a mechanism, I give hackers the opportunity to attack my front door, which is otherwise normally, at my home, not electronically attackable."* This aspect of

Alexa being yet another attack surface was brought up by most participants in almost all scenarios.

Another possible attack surface described was the cloud. Participants expressed a general uncertainty about the security of the Amazon cloud ecosystem. The main concern was data leakage, e.g., account credentials or credit card details, possibly due to cyberattacks. Participants also talked about attacks on the firmware running on Alexa. P7 though that injecting voice commands might be done via remote access to the device: *"Somewhere the process happens between the microphone that records, and the processor, which then processes the recorded text, so to speak, or the command. And then you can get the processor to process a command without having to enter anything into the microphone."*

As for the previously described threats, participants were also concerned about cyberattackers getting a hold of their voice code. Almost all participants stated that they believe, Alexa is always listening. This perception was again reinforced by news articles covering the topic. P9 explains how this permanent monitoring can facilitate cyberattacks: *"[F]or a device that constantly hears everything you are saying, anybody could somehow get into that device and catch that code, you know if you use that code on plenty of other accounts, I don't think I would personally use it for that, I guess I don't like using it to buy things."*

The quote includes an aspect not discussed yet, namely code reuse. Some participants talked about this throughout the interviews, stating that they assume users to reuse voice codes across applications, both on the Alexa platform and other systems. P9 expected users to use their credit card's PIN also as a voice code for Alexa: *"And I guarantee you if you have the voice code to open your front door that's gonna be your four-digit PIN for your card, it could be for plenty of things in your life."* So an attacker getting a hold of a user's voice code could use it to attack other accounts of that person as well. One participant highlighted IoT devices as a potential threat in this context. Devices like smartphones, smart TVs or even smart toys could be used to both eavesdrop on a user, and emit malicious commands towards a smart home assistant. P3 also suspected personal computers to be a target for such an attack: *"[I]f someone hacked into the computer from outside and has access to both the microphone and the speakers, they can hear my code and use it later. That is of course dangerous. If I then tell the device to open the front door, suddenly, while I'm on vacation, someone is inside of my house."*

6.1.7 Malicious Skills as Threat

A few participants brought up malicious skills that could be a threat in the scenarios. An important factor in this context was trust towards institutions. If a skill was provided by Amazon or an official account of, e.g., the user's bank, it was assumed to be safe, whereas skills provided by unknown third parties were perceived as being more dangerous. Past experiences with smartphone apps mostly influenced this perception. P2 explains this as follows: *"Well, if the function is from the official Amazon developers, it's pretty safe, while as the third-party apps it can be developed by anyone so, in that case it has to be tagged with, say, official, it's exactly how we have apps. So we have some apps which are*

authentic by Google or some bank apps which are from the bank and there are some apps which are completely not safe." However, participants failed to give a concrete example of how an attacker might use a skill to harm the user.

6.1.8 Accidents as Threat

Finally, almost all participants were concerned about accidents affecting the interaction with Alexa. This threat could manifest in Alexa executing commands without the user's intention. Alexa was suspected to potentially interpret a user's general conversation as a command. Also, children might accidentally activate Alexa. P7 states that: *"I can also accidentally, sometimes it happens that you accidentally say something, which she then picks up. Yeah, or my kids babble something in there. [...] so the children don't necessarily think about it, play something, imitate me, whatever. And then they say: Alexa, pay the grocery bill, whatever."* Some participants stated that this perception comes from past experiences they had with their Alexa, where voice commands got executed without their intention. Possible sources of such commands were conversations between the user and other people, and also TV commercials, as P13 described: *"I've often had the case that Alexa, when there's an advertisement or I'm talking to people on Discord, Alexa understands something and then does something that I actually didn't say to her."*

Several users were worried about technical faults interfering with their Alexa interaction. Especially when talking about the "Door" scenario, participants brought up power outages as a denial-of-service threat. Another category of technical faults coming up during the interviews were bugs in Alexa's software leading to unauthorized operations or data leakage, as described by P1: *"What also comes into play is that when you use the device, the device is always on and you do not know* 100% *whether the device has some error while you are not present and that something then happens without your knowledge and without your consent. That would surely be something where you think there could be some bug and the data can go anywhere, can land anywhere."*

The most reported case of technical fault leading to undesired actions by Alexa was the failure of the STT system. As a consequence, Alexa misinterprets a user's command and does not execute it or executes it with different parameters. As an example of this behaviour, participants stated transferring a wrong amount in the scenario "Dinner" or paying the wrong bill in the scenario "TV". Participants again reported relevant experiences they made in the past when Alexa did not understand them correctly. P1 explained that: *"Because there are always technical errors, you always have to take them into account. So far, no machine is without faults. [...] Alexa could just get me wrong at that moment and doesn't understand 20 but 200,000. Alexa is getting better and better, but you can't say that she understands you* 100%. *Most of the time, when she has the feeling that she has not understood you completely, she asks. But sometimes she just didn't understand and then guessed what I could have meant. Especially with transfers or 'I want to open my house' [that is] dangerous."*

Such misunderstandings could be favoured by background noise. This noise could be introduced, e.g., by other people talking, kids screaming, or a running TV, as P12 explains: *"That the voice transmission of the language code is somehow disrupted by the noises from the television, or that Alexa does not understand it properly. Because I've already had the experience with Alexa that you ask her something and then she answers something completely different."* This threat was in general perceived as more dangerous the higher the stakes were. While accidents were only viewed as an inconvenience when performing low-risk tasks such as turning on the light, the situation can become more dangerous if they occur during high-risk tasks, such as transaction, as P6 states: *"Because, whether I have to say twice to turn on the light and Alexa doesn't do it because she didn't understand it or she should please pay my electricity bill and she doesn't understand and she accidentally pays it fourfold, that makes a difference."*

6.1.9 Summary

We addressed RQ2 by reporting perceptions of threats and threat actors users of Amazon Alexa were concerned about when performing security-sensitive tasks. We found that most users perceived different kinds of bystanders as threats. Both familiar (e.g., family, friends) and less familiar (e.g., neighbours, casual visitors) people could be present during an interaction with Alexa, and the voice code used for authentication could be easily obtained by a motivated eavesdropper but also by an accidental listener.

There could be various motivations for adversaries. When talking about insider threat actors, such as friends or children, participants were concerned about pranks at their expense. While these pranks usually do not cause major harm, they present an inconvenience users want to avoid. Participants also considered more serious attackers such as criminals, both on- and off-line. Due to the high stakes included in the scenarios used in our study, criminals could be motivated by potential financial gains. Offline attackers were perceived as most dangerous in the context of the "Door" scenario. Physical access to their home was stated to be a primary protection target for most users. Alexa was perceived as a less-secure attack surface than regular door locks. This impression was due to weak authentication mechanisms, for which several possible attack vectors were highlighted:

- Eavesdropping on a user's interaction from a public space
- Replay attacks using recordings made with hidden microphones
- Brute-force attacks on small-entropy voice codes

In the context of personal finance scenarios, participants were concerned about cyber-attackers interfering with their device over the Internet. These attackers could exploit vulnerabilities in Alexa to eavesdrop on a user's voice code or inject malicious commands directly, in both cases bypassing the authentication. Some users also suspected attackers

of being capable of using other IoT devices to listen in on users and interfere with smart home assistants. Some users also brought up malicious skills as a possible attack vector for cyberattackers.

Due to past experiences, participants worried about technical faults impeding a secure interaction with Alexa. Examples highlighted include:

- Misunderstandings, possibly due to background noise, that can lead to command executions using wrong parameters
- Denial-of-service
- Unintentional voice commands getting executed

Finally, almost all users expressed privacy concerns when it comes to sharing data with Amazon. High-risk tasks such as money transfers can involve sensitive data which participants were uncomfortable sharing with a company they suspected of employing targeted advertisement or selling data to third parties. Storing user data renders data leaks on the back end of the system possible, potentially due to cyberattacks. Finally, some users also explained that Amazon or its employees might eavesdrop on a user's voice code and use it against their will.

6.2 Mitigation Strategies

In our user study, we also investigated which mitigation strategies users of smart home assistants employ to protect themselves from harm. We descriptively present the results on a per-code basis. The relevant category in our codebook is called *Users' Mitigation Strategies*. Table 6.3 depicts all codes included in this category and the total number of quotes for each code.

6.2.1 Refrain from Using the System due to Security Reasons

The most prominent mitigation strategy, brought up by almost all participants, was refraining from using the system. Participants stated that they would not use the system in its current form, due to security concerns. We observed this across all scenarios. In the context of the "Door" scenario, P6 states that: *"So I think language is too insecure for me to open a front door or to open my house."* Users would fall back to using known methods such as keys, and do things manually rather than via Alexa. P16 explains that: *"If you assume the scenario that all houses are smart houses, then I would turn off the function and manually open the door or whatever you do."*

Also, in the context of the other scenarios, participants stated that they would prefer doing the task via a computer or smartphone. P11 explains how this behaviour can mitigate the threat of technical errors on Alexa: *"I think I'd rather do it myself and see*

Code	Frequency
Build trust in security mechanism via trial-and-error	6
Change voice code regularly	11
Move to another room to use Alexa	3
Refrain from using the system due to security reasons	37
Take time for important actions	6
Users notice acoustic attacks on their Alexa if they are present	3
Voice code protects against unauthorized access	27
Voice interaction inappropriate in specific social situations	31
Whispering the voice code protects against eavesdropping	3

Table 6.3: Final codes in the category *Users' Mitigation Strategies*

if it really worked, because I can well imagine that she [Alexa] says, 'Yes, it's done' and then it's not done. Well, I wouldn't do it like that, I'd rather sit down at the computer and do it that way." We observed several participants weighing up the comfort gained by performing tasks via Alexa with security risks. P6 described this as follows: *"I [would] wash my hands for a moment or I put down my shopping bag. Or I don't do it with a voice command but with [near-field communication (NFC)]. And that would not be worth the risk to me in all of these 4 scenarios. So I wouldn't use any of the skill described here because the effort- or the comfort-to-risk ratio is not profitable for me."* Again, participants perceived Alexa as yet another attack vector.

Most participants based their perception of insecurity on the voice code. Users were not satisfied with the security provided by this authentication method, as, e.g., P8 highlights: *"I would have to think about it again, because now there are even more people present who could hear this voice code. Because as it [the interaction] works here, it sounds very easy to complete this payment, especially only with this voice code, so that is a weak security if it really only depends on a voice code."* Participants were overall not comfortable performing high-risk tasks, as described in our scenarios, using current state-of-the-art authentication mechanisms.

6.2.2 Move to Another Room to Use Alexa

A few participants talked about leaving the room if bystanders were present and using Alexa at a different location where no other people are around. This strategy can mitigate the threat of bystanders eavesdropping on the interaction. It could, therefore, enable the voice interaction, which would not be possible if a user did not want to share information with other people around, as P8 explains: *"In this scenario ['TV'] someone is actually sitting next to you, so you are not able to make this payment at all if somehow*

acquaintances were there who should not hear this voice code." Although this mitigation strategy does not work well in the context of the "Door" scenario, P11 explained how they could protect themselves from eavesdroppers when unlocking the front door: *"I could protect myself, when I am sitting in the car, that while I am still sitting in the car, I enter the code via the app and the door will open, then if the way to the door is short, it is already open and I can enter."*

6.2.3 Voice Interaction Inappropriate in Specific Social Situations

Most participants explained that, even outside of security concerns, they would not feel comfortable using Alexa for the presented tasks in specific social situations. An essential factor in this decision was which people were present in that situation. Regarding casual acquaintances, participants agreed that they did not want to perform security-sensitive tasks while such bystanders were present. However, participants had different opinions when it comes to people they have a closer relationship with. While some stated that they had no problem performing the tasks if only close friends or family were present, others explained that, although security might not be an issue, they would not feel comfortable talking about financial matters in front of them. P4 explained it as follows: *"When it comes to the circle of friends, I think that is a circle of trust, so to speak, I don't think of it as a major threat. However, money is always a delicate topic and you don't want to address that in front of everyone."*

Also, some participants expressed reluctance towards saying the voice code out loud, while people were present, even without any immediate security risk. P1 states in this regard: *"It is surely weird or bad when you are among several people and say your code at that moment. You don't say your ATM card code in front of everyone, so to speak. So now I am giving away my code, sure I can change it, but it is still strange to say that in front of everyone."* Another inconvenience that can arise when several people are present during an interaction with the smart home assistant is background noise. This background noise, as discussed earlier in Section 6.1, could lead to malfunctions of the STT system of Alexa. Therefore, it could be necessary to ask for some quiet, drawing attention to the user. In order to cope with these situations, participants wished for a less noticeable interaction mechanism, as P1 explained: *"Because when I'm talking to Alexa to transfer money, all the other 12 people have to keep their mouths shut, then it's much more pleasant for me to quietly take out my cell phone for a moment and make the transfer and say: Yeah, Kim, it's done."*

Finally, a few participants expressed a general unease giving orders to a computer in front of others. P10 explained their perceptions as follows: *"It feels a bit awkward sometimes. When I tell Alexa: Play some music! But maybe that's a matter of getting used to it. Because when I'm alone, it might be easier because it feels less odd. But if someone is around, I find that a bit strange."* When asked about the origin of this perceived awkwardness, the participant stated that: *"It's kind of uncomfortable that you might be showing that you entrust a computer with something. Or maybe you admit, that sounds a bit mean now, that you're lazy."* This association of using Alexa with being lazy was

shared between several interviewees. P2 points out that: *"Alexa is there that you can be lazy."* Users were worried about being perceived as lazy by other people, due to their Alexa usage. Performing tasks in a more traditional way, via a computer or smartphone, was recognized as doing it in person, in contrast to instructing Alexa. When it comes to authentication, user's favored discreet mechanisms that would allow them to enter their credentials in secret, without attracting the attention of nearby people.

6.2.4 Take Time for Important Actions

A few participants highlighted the importance of taking the time to perform important actions as a possible mitigation strategy for the threat of accidents occurring during the authentication. Rather than executing tasks spontaneously, as it is often the case when using Alexa, users preferred to carry out the operation during a dedicated time slot and in a quiet surrounding. P12 expressed this as follows: *"For me, certain processes also have their fixed place where I say, I don't do that just casually. Instead I sit down and have a certain amount of peace or can focus on it. Let's put it this way. I think it's about this focus, which I can then add."* An ideal system would let them double check the data they entered before submitting, e.g., credentials.

Several participants also expressed a need for a clearer overview when performing certain tasks. Voice interaction can be cumbersome if several options are presented to the users, as P9 voiced when talking about the "Hands" scenario: *"[I]t would be kind of be annoying to hear Alexa just list of transactions like a robot, [..] it would be easier to just read."* Visual feedback was desirable for most participants, especially when performing tasks involving complex parameters, such as money transfers, as P5 highlights: *"[W]ith a transfer you have to check everything beforehand before sending it off. I would rather have it in black and white before I do it."* In such scenarios, users wished for a more tangible overview, which they could get from screen-equipped devices. P7 stated that: *"[O]n the smartphone it is clearer again for me. I have, how do you call it, the visual control."*

6.2.5 Change Voice Code Regularly

A few participants expressed that they would change the voice code regularly. By doing so, they could mitigate the threat of insiders like friends and children. When asked about how often they would do so, P5 responded: *"I think I'd even change that every few days. Yes, very regularly."* Other participants stated they would adopt a new voice code every month. Furthermore, frequently changing the voice code could also prevent *Replay Attacks*, as P3 explains: *"That [Replay Attack] doesn't work anymore if I change the code regularly. Because then, in the meantime, the combination of numbers has hopefully changed so that the device no longer reacts."*

Moreover, participants perceived longer, higher-entropy codes as more secure. Such complex voice codes could be harder to remember for a casual eavesdropper and, therefore, protect users from pranks and insider threats. However, motivated attackers using technical equipment could still bypass this authentication system. P3 explained their

perceptions in this context as follows: *"Change the code regularly to reduce the risk, maybe a longer one, four characters can be remembered relatively easily when heard once, but when I'm visiting somewhere and someone says 12-15 characters, similar to a secure password, at least it is more difficult to remember. It is of course not secure against external attackers, then it does not matter how long the code is."* Users were aware of a trade-off between having longer, presumably more secure, voice codes and the difficulty to remember them. Therefore, they would adapt the complexity to the desired security level for any given task. Also, passphrases could be used to circumvent this trade-off by being both of high-entropy and easy to remember, as P2 stated: *"4-digit code would be easier to remember and less chances to forget it. Longer codes have that problem, but phrases again can bring back that [ease of use], we won't do much mistakes when we remember the phrase. It's like movie dialogues, we don't forget them easily [...] I would prefer: if less risk is involved then smaller codes, higher risk, bigger codes."* Participants expected higher-entropy codes to also be an effective counter measure against brute-force attacks.

6.2.6 Whispering the Voice Code Protects Against Eavesdropping

A few participants brought up using the whisper mode of Alexa to minimize the threat of eavesdropping. The whisper mode is a built-in feature that allows a user to speak to the device in a soft voice. The device will use a softer version of its default voice to answer. By using this interaction mode, e.g., for entering the voice code, eavesdropping would become harder, although still possible, as P6 explains: *"[U]sing whisper mode with Alexa, in which you can also whisper very softly with Alexa, which could only minimize the risk but not switch it off."* Participants also explained that this interaction mode could be used to keep interaction with Alexa more discreet, without drawing much attention to the user. Such an interaction can be desirable in certain social situations, as discussed above. However, whispering to Alexa can also have negative effects on the elegance attributed to Alexa interactions, potentially making users feel awkward as P6 describes: *"To kneel down for your own door and whisper so that it opens, I do not believe that fits in with the Alexa lifestyle."*

6.2.7 Users Notice Acoustic Attacks on Their Alexa if They Are Present

A few participants stated that users could notice malicious interactions with their device if they are in the same room or home. P7 explains that: *"[I]nside the house no real sound can get through. If someone stands in your garden and yells ALEXA, Alexa may still hear it, but I don't think so. And then you probably hear it too."* Participants expected that attacks would primarily happen while a user is away from home. Therefore, different security measures are needed depending on the presence or absence of users. P6 proposed a guard mode, where weak authentication mechanisms such as a voice code would be turned off: *"If I were worried that someone would have picked up the language code.*

Because if I had switched on guard mode, I couldn't use it anymore." This guard mode could then be activated when a user leaves the house.

6.2.8 Building Up Trust by Trial-and-Error of the Security Mechanism

Most participants agreed that they built up trust towards an authentication system while using it. Positive experiences, or lack of negative ones, can give users a sense of security, even if a system might still be technically vulnerable. P1 explained this as follows: *"Because in the end there may be some [security] issues with the payment method I'm currently using, but I've just done it so often and I'm so familiar with it that I feel safer because of that."* Therefore, a new authentication system may suffer from poor trust in the beginning and needs to earn its users' trust. P1 states in this context: *"So I think the uncertainty is certainly due to the fact that it is something new. If you have never done it before, you are unsure whether it is safe and whether it will work."* This process can be accelerated by employing a trial-and-error strategy. P10 explained how they would test a voice-based biometric authentication this way: *"I think I would sit in front of it quite often and try it out while disguising my voice. To see whether she [Alexa] recognizes it or not."*

6.2.9 Voice Code Protects Against Unauthorized Access

Participants also provided insights into their perception of the voice code and which threats it could help mitigate. Most participants stated that the voice code, as depicted by our scenarios, did provide some sort of protection. Similar to other knowledge-based authentication mechanisms, this protection was strongest while only the user knew the secret, as P9 explains: *"I mean the code right now it would make me feel more at ease, pretty good at ease, because it's a code that only you would know, hopefully, unless you are telling everybody."*

A voice code could also be a viable mitigation technique for friends or other people trying to prank the user, as P9 describes: *"Just gives another layer of security, so no, not, like my friend couldn't just come into my house and be like: Alexa, pay the utility bill."* In general, the voice code was perceived as another layer of security, providing minimal protection which users preferred to having no security mechanism in place at all, as P16 states: *"Yes, in any case, financial transactions and querying private data are privacy-relevant areas and I would definitely want to protect myself from free access by third parties. Such a code is just a relatively low-threshold and simple protection."*

Furthermore, having a voice code as authentication mechanisms can prevent unintentional voice commands from being executed. Users having a conversation, children playing around, or even TV commercials including relevant commands, could accidentally activate Alexa. P7 described how accidents could happen and included that a voice code could prevent user's from unintentionally interacting with another person's device: *"[The voice code] partially protects against it. I say it out loud, but I would never do that with a cell*

phone [code]. [...] Somehow that doesn't seem to make sense to me, but on the other hand, if someone else uses Alexa, I don't know, let's say when I borrow my brother's. Well, that doesn't happen to me then. I can also accidentally, sometimes you accidentally say something, which she then picks up. Yeah, or my kids are messing around in there. So it makes sense."

6.2.10 Summary

We reported the results relevant to RQ3. We described mitigation strategies talked about by our study's participants. Users largely agreed that they would refrain from using a system which they perceive as insecure, especially if it is viewed as non-essential. Alexa was generally perceived as a luxury article in that it facilitates tasks for its users but does not enable previously inaccessible features. Therefore, users consider Alexa as just another attack surface when it comes to security- and privacy-critical tasks. Users preferred to employ fall-back systems such as computers or smartphones to perform such tasks since they already use these devices in their day-to-day life, and they have a certain level of trust in them.

To build up trust and better understand which threats can be mitigated by an authentication system, users employed a trial-and-error strategy. By trying out the system under what they perceived as typical attacking conditions, such as disguising one's voice, users could gain trust in a new authentication technique. Most of our participants had not used voice-based authentication before and did not fully trust the system without hands-on experience.

When it comes to voice codes as used in most of our scenarios, participants were in general concerned about eavesdropping. In order to mitigate this threat, different techniques were proposed:

- Move to another room to use Alexa
- Use Alexa's whisper mode
- Change voice code regularly
- Use more complex, hard-to-remember codes

Participants stated they would move to a different room and interact with Alexa there if several bystanders were present as it is the case, for instance, in the "Dinner" scenario. In this context, users also explained that there exist specific situations in which interaction via voice was undesirable. Some stated that this was due to an awkward feeling when talking to a computer, which can be perceived as admitting to being "lazy" because a user does not carry out tasks themselves but mandate a computer. Furthermore, interaction over voice can draw unwanted attention to the user. Especially for money-related tasks, participants expressed a desire for discreet interaction options. Using the whisper mode

of Alexa can be a less obtrusive operation mode. Participants also stated that it could be used to mitigate eavesdropping. However, this input feature was perceived as less elegant and, therefore, not fitting into what some users described as the Alexa lifestyle.

Another mitigation strategy for eavesdropping was changing the code regularly. By doing so, participants expected that a leaked code would no longer be valid during an attack. Note that a few participants believed that they could recognize attacks on their device while they are present. Therefore, attacks would mainly occur while a user was away from home. Changing-frequency of the voice code varied from every few days to months. Participants also stated that more complex codes might be more secure than the four-digit PINs used in our scenarios. By including a wider variety of characters and choosing longer sequences, the threat of non-technical eavesdroppers could be mitigated, since overhearing a complex code and also memorizing it was perceived as hard. This technique was also described as an effective mitigation strategy for brute-force attacks. However, more complex codes were also harder to remember for users. In order to circumvent this drawback, passphrases were proposed, e.g., quotes from movies. Such phrases combine facile recall of shorter codes with large entropy of longer ones.

Finally, participants talked about mitigation strategies for accidentally executed voice commands. In this context, several participants noted that a voice code could be an effective technique. Users believed a voice code would not be uttered casually, as can be the case for standard commands. A voice code was also perceived as a low-threshold security mechanism, capable of preventing casual pranks and providing fair protection as long as only the user knows it. Another mitigation strategy for accidents was to take one's time for important tasks. Alexa interactions most often happen casually, when a user's mind might be occupied with another task as well. When using a computer, participants stated that they were intentionally sitting down in front of it and focused their attention on the task at hand. This interaction pattern could also be transferred to smart home assistants.

6.3 Important Properties of Secure and Usable Smart Home Assistant Authentication Systems

We present important aspects of authentication systems for smart home assistants we identified during the open and axial coding phases of our analysis. These properties showed to be crucial for users' perception of security when performing high-risk tasks via voice commands. We go on to discuss these findings in Section 7.2, where we provide design implications based on the gained insights.

6.3.1 Building Trust

Trust showed to be an important aspect for users' perception of security when it comes to authentication systems. Most users stated that they had not used voice authentication mechanisms in the past. Therefore, they perceived this technology as new and unfamiliar.

Users did not trust the system out-of-the-box. However, they described several ways of how trust can be built-up for such a novel system. Almost all participants describe how trust can be transferred to a system if a trusted entity backed it. In the scenarios including personal finance tasks, mostly banks were highlighted as being trusted by the users, but also *PayPal* and energy providers were brought up in this context. Trust in these parties was also rooted in past decisions, as P10 stated: *"I trust my bank, otherwise, I wouldn't be with my bank."* Participants stated that they would have more trust in a system if it was directly provided by a trusted third party. P8 explains: *"So if it really came from the bank, I'd trust the whole thing more, then I'd be more inclined to use it."* Another reason for this trust in established institutions was that users believed, third parties would have an interest in keeping systems secure due to potential consequences, possibly of legal nature, in case of an incident. P9 explained this follows: *"[T]he bank obviously has the best of interest to keep my account secure, because obviously if somebody gets into that account, takes all that money, they have to refund me that money, if they prove that it was fraud. Then they have to do the investigation, track down that fraud and they have to go after that money. So really they have more a massive interest and the security would be a little tighter."*

We identified another way how participants gained trust in a system. Positive experiences made in the past, combined with a lack of harmful incidents, strengthened users' perception of security about a system. Since most participants lacked this kind of experience with Alexa, they were reserved for performing high-risk tasks via this platform, as P1 explains: *"If I imagined that I would make transfers like this from now on, it would be quite a big hurdle for me, simply because it is unfamiliar at the beginning and because you don't know whether you can trust it now. That would be the thought at the beginning."* Several participants expressed that they would prefer using a classic typing interface, as found on smartphones or laptops, over voice commands. They attributed this preference to being used to that interaction method, as P4 stated: *"I'd rather type in. Maybe I'm old school, I'd rather type in than doing that via voice commands. But yes it could be in the future, you get used to everything. I'm someone who prefers to type it in."* Other participants also expressed that they would expect voice interaction to become more common in the future and, therefore, new users might feel less reluctant to use the system for high-risk tasks. P10 draws a comparison to smartphones, which, at some point, also were novel systems: *"I can already imagine that for the new generation, it will become quite normal for them. Just as it is not normal for my mom to do banking on her phone, she does it on the computer. So it will perhaps be normal for the next generation to tell Alexa such things."*

Good experiences with a system in the past led to a higher trust in its security. Similarly, the trust could be lost if users witness security incidents, as P9 highlights: *"If I was like 'Oh, man, I really trust this!' And then you came by and said like 'Send me 80 dollar! 8915' and it [Alexa] was like 'Okay' and I'm like 'I trusted you Alexa!'"* As reported in Section 6.2, applying a trial-and-error strategy to authentication can facilitate experiencing a system in a shorter period. Another similar strategy to build trust, as

described by some of our participants, was checking reviews and ratings by other users of a system. Reading up on experiences made by other people can have a similar effect as experiencing something first-hand when it comes to trust. P1 expressed this as follows: *"What always helps are things like customer reviews or experience reports, [...] if you read that everything works, you have so and so many people who rated this. If the reviews are consistently positive, that would certainly build up trust."* Participants also brought up talking to acquaintances who already used a system.

6.3.2 Transparency and Agency

Almost all participants stated that transparency is essential when it comes to the perception of security. A transparent system can enable users to make educated decisions about their interaction. Several participants noted that this property did not transfer well from computers or smartphones to smart home assistants. This attitude was partially due to the interaction via voice rendering it more difficult for a user to get a good overview of the current state. Visual interaction enables users to absorb information much quicker, as stated by P7: *"So when I order on the [personal computer (PC)], I have several options that I can grasp directly and it is simply easier for me to take in with my eyes than to listen with concentration."* Using a computer also conveyed a feeling of being in control, which was an important characteristic when it comes to high-risk tasks, as P14 explains when asked about distinctions between using Alexa versus a computer: *"For me, I guess the difference is I feel secure that I'm the only one that knows the PIN and, yeah, I have control to log in or log out, that I have control that I'm not leaving anything open, that I leave myself potentially open for a threat."* Using Alexa, in contrast, was perceived as surrendering the agency over a task to another party. The user was no longer the active part and could only hope for the successful execution of the process, as P10 expresses: *"I say it to a computer, because I give a computer the power to manage my money. If I click it, it's still: I clicked this, I decided to pay for it."* The participant justified this feeling with a non-transparent control flow: *"It's weird if I don't see when something happens. Because I say something and it happens. And then I just can't understand whether it was done correctly."*

A potential factor for this perceived loss of agency was the personification of Alexa. Several participants compared Alexa to a human operator and stated that voice commands felt like giving orders to an employee. This perception entailed that Alexa could be affected by human error, as stated by P7: *"Suppose I had a butler and I always had to tell the butler, yes, open the front door. I can't trust that* 100% *either. Clearly, somehow, there is a large basic trust. But even then it's kind of uncomfortable when you have in the back of your mind: what if he didn't do it, what if he forgot about it?"* In this context, participants expressed the wish for more transparent control flow. This could lead to a better understanding of which entities are involved and how tasks are distributed within a system. P10 provided an example of this: *"I tell Alexa to please check whether I have transferred it. With the skill, she'll pass it on to my bank in that case. And then the bank looks it up. I would want that Alexa said that too. [...] That she also says: 'Okay, I'll*

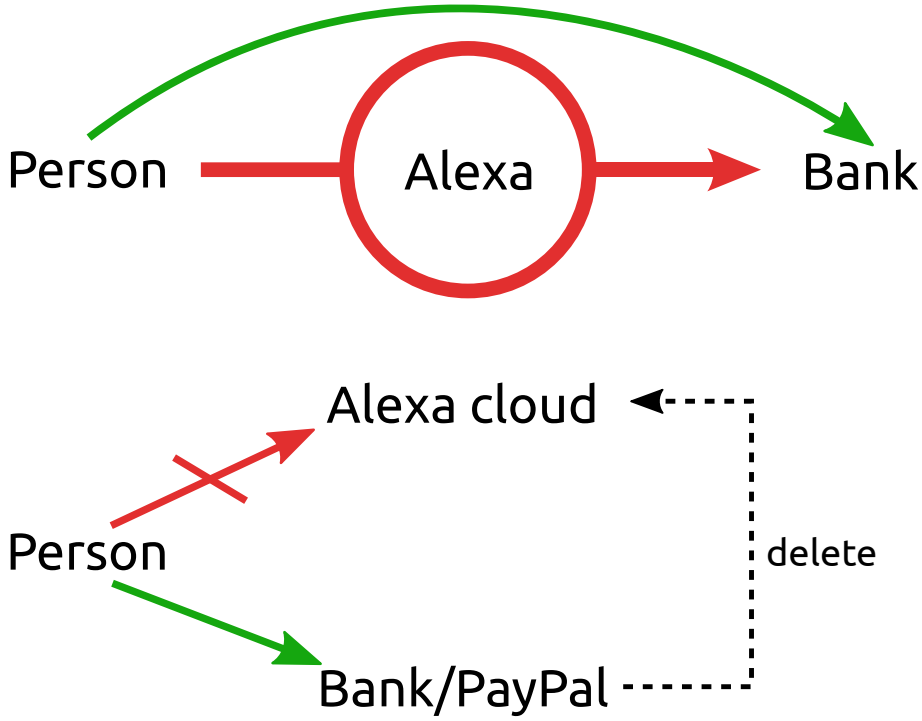

Figure 6.1: Ideal data flow when performing online banking via Alexa, as depicted by P6 (Top) An ideal data flow would not allow Alexa or Amazon to intercept any data, as if a person would talk directly to their bank (*"as if I was standing at the bank counter"*) (Bottom) A technical solution to this could involve a user's device decoupling from the Alexa cloud after the initial connection to a third party (e.g., a bank or *PayPal*) was established. The processing of audio data would then be handled by the service provider rather than Amazon. Alternatively the third party could request the deletion of all relevant data from the cloud on the user's behalf after each interaction. The original drawing created by P6 can be found in Figure A.1.

pass this on to the bank'. Then she thinks about it for a moment. I don't know if Alexa has to think about it, but then she thinks about it for a moment and then says: 'According to the information from your bank, you have not yet transferred that.'" Participants envisioned an ideal flow of information to be as direct as possible, limiting potential data leakage towards Amazon. Their main concern was information privacy. P6 depicted their idea and a potential solution in Figure 6.1.

6.3.3 Risk Assessment of Authentication

Participants provided insights into how they assess risks in the context of authentication. Based on this assessment, users derived different requirements for an authentication system. One such factor was the location of interaction. Some locations might call for stronger security measures than others when it comes to authentication on a smart home

assistant. Most participants agreed that the most distinctive difference was between interactions occurring in a public space (e.g., in front of the door) and those taking place in a private space (e.g., the user's home). P8 stated that: *"In this scenario ['Door'] it is about the environment, that you are outside. I think I would feel a bit safer in these two scenarios that I just brought out ['TV', 'Hands'], because you are in the privacy of your own home. You then think you have more control over it. So who gets this data, who hears it, and so on."* Therefore, an interaction taking place inside a user's home could use a weaker authentication mechanism, as expressed by P9: *"Obviously, if you're inside the house [...] I believe the voice recognition and the code would suffice plenty. Cause you are already in the house, it's you, you can tell by the voice and then you also give him [Alexa] a code. So I would be okay with."*

A few participants also described different zones inside a home relevant for authorization decisions. P3 provided an example: *"Transactions or something like that are only allowed from the study, while for the device hanging in the children's room, or in the hallway area where everyone has access, only certain things work there."* Another factor taken into account as part of a risk assessment by the users was being at home vs being away. As reported in Section 6.2, participants perceived the threat of security breaches to be more prominent while the user had left home. Authentication systems could follow this assessment and apply stronger methods during the vacancy period.

Furthermore, we found that participants weighed the perceived risk against the effort needed to authenticate. If an interaction was considered to be low-risk, users were comfortable with using weaker authentication mechanisms (or no authentication at all). We observed this, especially in the context of the "Hands". Several participants stated that they would prefer having no voice code when checking transactions. In contrast, most participants agreed that in order to execute transactions, an authentication step should be in place. P9 explains this as follows: *"[If] your bank would let you pay the utility bill you have already set up with your bank, and then your bank also lets you view transactions, I would be okay with [having no PIN] to view transactions [...] to make transaction, have the PIN. That made me definitely feel better."* P14 could imagine using the currently available voice code authentication to perform low- to medium-risk tasks, but not for high-risks ones: *"I would be fine with using a voice code to see my transaction history, even my account balance, even though that's also quite personal information, but to make an account transaction, I don't think Alexa should be allowed to do that."* Some participants also expressed different requirements of protection depending on the amount of money transferred, where low amounts could be sent without strong authentication. A few participants explained that, since absolute security did not exist, there has to be a trade-off. P16 described this as follows: *"It just always depends on how much effort I want to put into it, there will be no absolute privacy with such a system."*

6.3.4 Perception of Authentication Methods

We report on participants' perceptions of authentication methods, split according to the distinction presented in Section 2.2. We discuss how these insights can help designing

new authentication schemes for smart home assistants in Section 7.2.

Knowledge-Based Authentication

In Section 6.2, we reported users' perceptions of the security benefits of a voice code authentication, as included in our scenarios. Participants thought of the voice code as a low-level barrier that could primarily mitigate casual attacks and pranks by familiar people. If every skill had its own voice code, users were worried about not being able to remember all of them. P10 stated that: *"It would be tough if every command had a different code, because who can remember so many codes? But of course, it would be more secure. Because then I have a different code for practically everything I want to do. But then you just have to memorize them."* This factor can lead to users reusing their voice code across applications, as P16 describes: *"Just as there should be a separate password for each account, there should be a separate PIN or code for each skill. As we all know ourselves, we will use the same everywhere."*

Participants also described similar coping mechanisms, e.g., only modify one or two digits between codes. P9 highlights how weaknesses of voice codes could potentially affect the security of other, unrelated systems: *"In a negative way, somebody could get that code, say that's the exact same code I use at the ATM or check card and somehow they could go get into your bank statements."* A few participants were unsure whether the voice code would be specifically set up for the Alexa platform, or use the regular authentication method, e.g., for their online banking. P10 expressed this as follows: *"[W]hen I log into my bank app, I have to put in a PIN. So maybe that could be Alexa actually entering that PIN into the app to getting access to the system."* One participant stated that, while the voice code was not an acceptable authentication method for high-risk tasks, it could serve as duress mitigation. By setting up a code for threatening situations, a user could say that code instead of their usual authentication code, upon which the system could initiate an emergency routine, e.g., call the police.

Possession-Based Authentication

Several participants stated that they would like to use token-based authentication with Alexa. Tokens would not be susceptible to the openness of the voice input channel. Participants described different kinds of tokens they could imagine themselves using in combination with a smart home assistant. Hardware tokens could be used to detect the physical presence of a user, which could be matched with the location of the Alexa device. If the token was close-by, it could be assumed that a voice command was made by a legitimate user. P2 gives an example where a smartphone is used as the token: *"So if my mobile phone is in, Alexa can connect to my mobile phone, it's in the same location as I am communicating, then I guess it's fine."* It could also be possible for a token device to be equipped with a microphone and relay the command directly to the smart home assistant. P9 explained how a *Fitbit* could be used as a registered authentication token: *"[L]et's say your Fitbit that is also connected to Alexa, if when you get near that door Alexa can actually read 'Hey, yes, it's you.'"* They also propose a scheme where a

device is registered as trusted, which allows for weaker authentication mechanisms when accessing Alexa via this device.

Another way how smartphones could be used for authentication purposes with Alexa was push notifications. Some participants expressed that getting a notification requiring confirmation whenever a security-sensitive voice command was executed could be a secure authentication mechanism. P4 explains that: *"[T]he best for me is if I can do this via the app, where I start an additional app where I can then enter it [the voice code]. I feel like I am in a more protected environment when I can just type it in on my mobile phone."* Several participants also brought up using OTPs instead of static voice codes. These OTPs could be received via a trusted device, such as a smartphone. P14 explained how interaction with Alexa could look like, if OTP authentication was used in combination with a smartphone: *"[Y]ou would say 'Alexa, open my front door' [...] Alexa says 'Okay, generating your code now' And then it flashes a four-digit number on your [smartphone's] screen and says 'State your code to unlock the door' And then it's not always the same code every time."* A drawback of token-based authentication described by a few participants were *Replay Attacks*, were an attacker registers the signal sent by a token and replays it later, leading to a potential authentication bypass, see also Section 4.2.2.

However, some participants explained that using a dedicated device, such as a smartphone, for authentication can defeat the purpose of Alexa, in that a user could then perform a task on the smartphone in the first place. P6, talking about the "Hands" scenario, stated: *"I think that's impractical, because if I have to pick up a smartphone to verify myself, then I could check it right away, via an app."* Tokens might also not be compatible with *hands-free* interaction desired in some situations, as P9 described: *"[A]nything more than a voice verification, like, say, a thumbprint on your phone, at that point you are defeating the point of the Alexa being able to talk to a virtual assistant, now that you have to involve physical things to actually pay, so at that point you just log into your phone and do it."*

Biometric Authentication

Most users stated that authentication via voice biometrics was their preferred method of authentication, due to the naturalness of the interaction. Also, no additional effort would be required from a user to authenticate. P6 states that: *"I really believe that distinguishing by voice would be the smartest way that Alexa could classify a user."* However, some users had doubts about the current state of voice recognition accuracy concerning Alexa, mostly due to past experiences. When asked about their trust in the security of voice recognition on Alexa, P16 stated: *"Currently no, not satisfied. You can tell the difference, it recognizes you by your voice, but even this recognition sometimes doesn't work, and I think that's very rudimentary. It's nice that you can see that this feature is under development, but it is far from mature."* Some participants also required voice recognition to be able to distinguish live voice from machine emitted sounds. Otherwise, *Replay Attacks* could break the authentication system, as P6 explains: *"Security would only be given if this banking skill were to work with speech recognition*

that filters electronic voices. This means that Alexa would not accept a recording of you in order to carry out payments for you." Another drawback highlighted by interviewees in the context of voice recognition was annoyance caused by false negatives, i.e., Alexa not recognizing a legitimate user. P12 explains how this could be due to variances in a human's voice: *"It's just difficult because the voice is often different, let's say when you have a cold, for example. Voice sounds different in the morning than in the evening. So I think that's technically difficult to do."*

In the context of the "Door" scenario, some users also brought up face recognition as a potential authentication mechanism used in combination with a smart home assistant. By using a camera, a device could measure facial features or scan a user's iris as an authentication method. This technique could maintain a hands-free interaction, as P16 explains: *"The scenario is that your hands are full, that is to say, sending a unique authentication code to your mobile phone is then in vain, alternatively, since you are in front of the door, other biometric data would then also be possible so that in this case you could use eye lasers, but that might be an overkill."*

Multi-Factor Authentication

Combining some of the above-described methods to form a stronger multi-factor authentication was brought up by most study participants. Users perceived a direct relationship between more authentication factors and better security. P9 explains how a multi-factor scheme could look like: *"Just makes it, I mean, two layers of security is better than one. [...] You have your mobile device, then the code and the voice profile, So I would have three phases, so someone would have to steal your mobile device or watch, would have to learn your code and would have to sound like you."* Among the most popular combinations among our participants was having a voice code and voice-based biometric authentication. This perception was influenced by other high-risk systems users were familiar with, which also use multi-factor authentication, as P6 explains: *"[I]f I want to complete a transaction via online banking, it never works with the use of a single device. Because you either have a small [transaction authentication number (TAN)] device into which you insert your bank card to verify this transaction [...] or you have your mobile phone connected to it, then you get a verification SMS with a code."*

CHAPTER 7

Discussion

We discuss our findings from Chapter 6 and their implications, and give some recommendations for the design of upcoming authentication systems for smart home assistants. Furthermore, we discuss some limitations our study might be subject to.

7.1 Relating Actual and Perceived Threats

We discuss how the threats perceived by our study participants, as presented in Section 6.1, relate to actual attacks and threats, covered in Section 3.4 and Section 4.1.

7.1.1 Eavesdropping

In our study, most participants were concerned about others overhearing or eavesdropping on their interaction with Alexa. Voice is a rather open channel, and conversations can be understood within a radius of several meters. Eavesdropping on a conversation can be seen as an auditory variant of shoulder surfing, a practice including an attacker observing a user's screen, see Roth, Richter, and Freidinger [RRF04]. As Eiband et al. [Eib+17] report, most people might have experienced a shoulder surfing incident, either actively (i.e., as an observer) or passively. Similar, people might have made eavesdropping experiences in the past, making them aware of this threat. When it comes to bystanders, such as neighbours, participants expressed a certain distrust towards them. However, most failed to explain how such a person would act upon learning their authentication code. More concrete motivations of attackers were stated while discussing the "Door" scenario. Burglars are a threat most participants were aware of, having clear intentions of harming a user. This awareness could stem from extensive news coverage of general burglary incidents.

7.1.2 Insiders as Threat Actors

When it comes to insiders as a threat, participants were less sceptical as friends and family members were described as a trusted group of people. The biggest perceived threat originating from these actors were pranks, which were thought of more as an annoyance than an actual threat. Also, children were perceived as a minor threat, due to their motivations being of mostly non-malicious nature and limited damage potential. Only a few participants talked about intimate partners as a threat, while most stated that they trust their partner and would have no problem telling them their voice code. However, previous work by Levy and Schneier [LS20] and Marques et al. [Mar+16] suggests that adversarial behaviour between family members happens on a relatively large scale.

7.1.3 Security Breaches Due to Misinterpretations

A problem many participants had experienced first-hand when using Alexa were accidental command executions due to a misinterpretation of the wake word or the voice command. Participants reported that conversations or audio from a computer activated Alexa. Schönherr et al. [Sch+20] found over 1000 triggers for Amazon Alexa, Google Assistant and other smart home assistants in TV-shows, news or audiobooks. For instance, the utterance "a lesson" activated Alexa. On several occasions, subsequent words were taken as a voice command by Alexa, leading to various subroutines getting executed. Users believed that such accidental activation could have negative consequences, e.g., Alexa ordering products. Also, erroneous STT conversion during interaction with Alexa was brought up, such as transferring a wrong amount of money. However, no participants expressed that such unintentional activation or misunderstandings could be exploited by an attacker.

As described in Section 4.1, the *CommanderSong* attack is based on a similar concept and is capable of activating Alexa without uttering the genuine activation word. A similar presented attack (*Skill Squatting*) capitalizes on misinterpretations made by the STT in combination with homophones to execute attacker-controlled skills instead of genuine user skills. This technique might be used to capture a user's authentication credentials and perform a MITM attack. However, only a few participants were aware of skills being used as an attacking tool, while all of them failed to provide a concrete explanation of how such an attack could be performed.

7.1.4 Attacks on Voice Recognition

Almost all participants talked about voice recognition as an authentication method for Alexa. Most interviewees were in favour of this technique as it was perceived as a natural and unobtrusive way of authenticating users. Participants were aware of some, mostly low-tech, attacks on voice biometrics. Most users brought up the *Replay Attack* where a recording of them could be used to bypass authentication, see also Section 4.2. Participants expected attackers to employ readily available devices such as microphones to capture a user's interaction. Due to the widespread usage of smartphones, most people

have access to a device capable of both recording and replaying sounds. While most users thought an attacker would need to capture a user's voice in situ while authenticating with Alexa, a few participants were aware of the *voice sampling* attack where arbitrary recordings of users can be used to form the necessary utterances for authentication, as covered in Section 4.1. This technique simplifies the attack vector since an attacker does not have to go near a user's smart home assistant to place a recording device but can try to capture a conversation in a more public space. Also, mitigation techniques such as using the whisper mode are no longer effective.

A few users also talked about non-technical attacks on voice recognition, where an attacker might try to mimic the user's voice. We demonstrated the feasibility of such an attack in Section 4.2.3. Interviewees stated that this attack would be a go-to candidate in their efforts to test the security of a system via trial-and-error, as it was easy to execute and provided quick feedback. No participants talked about how synthetic voice could be used to attack a system, as we illustrated in Section 4.2.2. However, mostly due to awareness of the *Replay Attack*, the majority of participants required secure voice recognition to be able to distinguish machine playback from live human speech.

7.1.5 Injecting Commands without the User Noticing

When it comes to inaudible attacks, no participants demonstrated to be aware of such techniques. The *DolphinAttack* can be used to interact with a voice assistant while other people are nearby. In contrast, the *LightCommands* method enables interacting with smart speakers without having physical access. We covered the technical aspects of both techniques in Section 4.1. These attacks can undermine important attributes of users' perception of security in the context of smart home assistants:

First, participants expected to notice other people interacting with their device if they were nearby. For instance, the perceived threat of an attacker yelling voice commands through an open window was assumed to be detectable by a user nearby. If an attacker were to use ultra-sonic frequencies or light signals instead, this might no longer be the case. Therefore, the perception of threat in such situations might not match the actual risks users can be exposed to. Users had different requirements for the security of authentication mechanisms depending on if they were at home vs when they were away.

Second, interviewees stated that the house functions as a physical barrier to attackers. They had different perceptions of threat depending on whether the interaction was taking place inside or outside of it. For smart home assistants located inside a closed house (i.e., no open windows or other entry points) users believed an attacker would have to gain access before being able to interact with the device. Locking the door and closing the windows when going to bed at night might be perceived as a sufficient protection strategy. However, as the *LightCommands* attack demonstrated, it is possible for an attacker to inject commands through closed windows and from a considerable distance, as long as a line-of-sight to the device existed, see Section 4.1. Again, there might be a gap in users' perception and actual threats.

7.1.6 Cyberattacks

Most participants stated that cybercriminals or hackers are a potential threat. They expected such actors to be able to bypass authentication by attacking the firmware, both over the network and locally. Participants did not know how to protect themselves from this kind of threat. Mitigation strategies, such as keeping systems up to date, were not talked about during the interviews. This is in line with the findings made by Anell, Gröber, and Krombholz [AGK20]. In their study, only experts brought up updates as a countermeasure to generic security threats. Hackers were also believed to be able to attack the back end system of Alexa (i.e., the cloud), due to uncertainties about the security of the infrastructure. As for most systems, potential consequences stated by users include leaking credentials or sensitive user data. Participants also stated that a smart home assistant could be used as an entry point to the broader smart home ecosystem. Some participants explained that, since smart home assistants can control other IoT devices, there has to be some connection. An attack capable of controlling a smart home assistant was perceived as being able to control other IoT devices and compromise the system to a larger extent.

7.1.7 Non-Targeted Attacks

Participants in our study were mostly concerned about targeted or opportunistic attacks (e.g., children capitalizing on an accidentally overheard voice code). Only in the context of the "Door" scenario, non-targeted attacks were brought up. Users stated that burglars could explore potential targets by walking up and down a street and checking for smart home controlled doors, possibly by repeating the wake word for reconnaissance. Participants also described attackers going from door to door and trying voice codes at random, eventually succeeding due to chance. However, as highlighted by Yuan et al. [Yua+18], large scale attacks on smart home assistants are possible, e.g., through malicious songs playing on the radio. With security-sensitive tasks such as making money transfers getting adopted more and more, the incentives for such widespread attacks grow.

7.2 Design Implications

Finally, we provide design recommendations based on the findings reported in Chapter 6. These aspects were crucial for participants to feel protected during security-sensitive tasks presented in the scenarios.

7.2.1 Rely on Voice Recognition as an Intuitive and Trustworthy Authentication Method

Most participants agreed that voice recognition was the most desirable authentication mechanism for smart home assistants. It was perceived as a natural way of authentication, as it is one of the default methods humans apply when verifying a familiar person, for

instance, when talking on the phone. A few participants even expected conversational agents to be already capable of doing so. Some smart home assistants employ a form of voice recognition to distinguish users, however, it is not recommended as an authentication mechanism yet[Amaa]. Participants were aware of some potential threats and shortcomings of voice recognition that would need to be addressed before users can trust such a system. The most prominent feature being a live detection to distinguish human voices from speaker playback.

7.2.2 Include Demo Mode to Let Users Experience the Effectiveness of the Authentication Method

We observed that users initially mistrust new authentication mechanisms they had not used before. If such a novel scheme is designed, we recommend including a demo mode which participants can use to try out the authentication process. Most state-of-the-art systems will block access once a threshold of unsuccessful authentication attempts is reached. Such systems are, therefore, not suitable for users to test different adversarial techniques. By including a separate sand-boxed mode that allows unlimited tries, users might be able to build up trust faster and also get a better understanding of a novel interaction mechanism. Any such demo mode must have the same look-and-feel as the standard authentication process, the only difference being that upon successfully authenticating, no real user data is accessible. The user should also be informed whether an authentication attempt was successful or not.

7.2.3 Provide Unobtrusive Authentication for Social Situations

We found that participants felt uncomfortable using conspicuous authentication mechanisms in certain social situations, e.g., during a dinner with friends. When designing authentication for tasks projected to be performed in such social settings, we recommend including an unobtrusive authentication mechanism, that allows users to perform security-sensitive tasks without drawing much attention. While conventional voice recognition has shown to be a desirable option, it might not work for settings including several bystanders. Voice recognition is conspicuous unless considerable background noise (e.g., loud music during a party) is present, in which case the STT system of the smart home assistant might have difficulties understanding the user correctly, leading to failed authentication attempts. Participants reported having experienced such erroneous behaviour before, see Section 6.1.8. An implementation of a new system could also automatically identify the current social situation a user is part of during authentication by, e.g., detecting other persons nearby or measuring the level of background noise. The system could then dynamically adapt the authentication process according to predefined rules for different situations.

7.2.4 Maintain Low-Effort Interaction

As we have found, crucial characteristics of smart home assistant interaction for users were the effortlessness and straightforwardness with which tasks can be executed. Users reported that the main reason to use a smart home assistant was that interaction with these devices was faster and more effortless compared to computers or smartphones. If an authentication mechanism takes away these features, e.g., by making users access one of the described screen-equipped devices, our participants would not be willing to adopt this technology. As the benefit of voice interaction was diminished by requiring interaction with other devices, users were no longer willing to take on the perceived additional risks that are coming with smart home assistants. Also, participants felt that if, for instance, smartphone interaction was required during authentication with a smart home assistant, they could use the smartphone to perform the task in the first place. Therefore, for use cases already available on traditional platforms, the design process of new systems has to account for this risk-benefit analysis made by the users and reduce the effort needed to authenticate to an adequate amount. Such low-effort interaction could be provided by continuous authentication mechanisms, as described, e.g., by Feng, Fawaz, and Shin [FFS17].

7.2.5 Keep the Authentication Process Transparent

During our study, we found that participants were unsure about how the flow of information during an authentication process looked like. Especially which party performed the verification of the presented authentication information in scenarios involving third parties (e.g., banks) was not clear to all users. While some participants believed Amazon would authenticate the user and then get permission to access their account, others perceived Alexa as some sort of butler, who takes a user's credentials and uses them to log into a banking application on the user's behalf. Since it was not evident from Alexa's output whether it came from Amazon or a third party and conversational agents were attributed with human characteristics, the perception was reinforced that Alexa uses third-party systems just like a human user would do. To enhance transparency and make control flow transfers from the Alexa back end to third-party skills easier to detect, we propose using different voices for each subsystem. This way, a user could instantly notice once the third-party takes over and at which point built-in routines control the interaction again. A similar mechanism to provide this transparency to the user could be having Alexa announce handing over control to a skill and reporting back once a request has gone through. This practice could lead to better mental models and, consequently, more educated decisions made by users.

7.2.6 Account for Varying Requirements

Users stated that they have varying requirements for authentication mechanisms depending on two main factors. The first was the location of the interaction. For scenarios taking place outside, users were concerned about a larger number of threats than for interactions

happening inside the home. The second factor was whether a principal user was present or not. Users were under the impression that they can detect malicious behaviour when they were nearby. Therefore, fewer threats were relevant in such circumstances. Also, the task carried out affected the users' security requirements. Most participants agreed that information requests were less security-sensitive compared to tasks involving money or physical access.

We provide an example for three different security levels that can be derived based on three contextual factors: the presence of a principal user, the location of the interaction, and the kind of task.

- **Low:** A low security level is given while a principal user is present. Only trusted people are within hearing range. Tasks entailing a low security level revolve around information requests with required previous knowledge, e.g., checking if a specific transaction was executed.
- **Medium:** A medium security level is given for interactions of a principal user while outside the home. Several unfamiliar people could be able to overhear voice commands, possibly also without the user noticing. Tasks at the medium security level include arbitrary information requests, e.g., checking an account balance.
- **High:** A high security level is given for interactions while the principal user is out of the house. Also, if untrusted people (e.g., a tradesperson) have access to the smart home assistant while the user is at home but not in the same room, the security level can be classified as high. Among the tasks entailing a high security level are physical access control (e.g., unlocking a door) as well as money transactions.

If the principal user was away from home, the smart home assistant should still be accessible or remain turned on, however, security mechanisms should become more restrictive, especially when it comes to authentication. A possible feature accounting for these varying requirements could be a guard mode, that, if turned on, requires stronger authentication to turn back off. A real-life example would be an alarm system that requires a code to be disarmed. A user could turn on the guard mode if they leave the house or go to bed at night. Upon their return, they authenticate once using a strong, possibly multi-factor, authentication mechanism to turn guard mode off and switch back to the default authentication method, that could be weaker and less obtrusive.

7.3 Limitations

As described in Section 5.1.1, we used vignettes and scenarios in our interviews to investigate users' perceptions in the context of security-sensitive tasks via smart home assistants. As soon as the tasks included in our scenarios become widely available, and users had the chance to experience them first-hand, an in situ study could provide further insights that could be missed in a lab setting.

As reported in Section 5.2, our sample was almost balanced concerning gender. We also managed to recruit participants with a variety of educational backgrounds. However, the age distribution of participants was skewed towards younger participants, hence, older users of smart home assistants might be underrepresented. We also measured participants' affinity for technology interaction using the ATI scale, as well as their concerns for information privacy using the CFIP scale. Our sample participants score slightly above average for the former, compared to the results of Franke, Attig, and Wessel [FAW19]. CFIP scores indicate that our participants were highly concerned about their information privacy, compared to the results of Rose [Ros06], indicating a further potential underrepresentation. As this study is exploratory in nature, our goal was not to formalize hypotheses that can be generalized for the whole population of interest. Any future work doing so should aim for a more balanced sample in this regard.

Most decisions we took during study design and data analysis can be influenced by expectations and conceptions we had going into this work. Scenarios and questions presented to participants during the interviews might be subject to bias. We cope with this by making our expectations explicit in Section 5.1.2, in line with previous work, e.g., Krombholz et al. [Kro+19] or Braun and Clarke [BC06]. Furthermore, we reprint the guideline that we used during the semi-structured interviews in Appendix A.1.

CHAPTER 8

Conclusion and Future Work

We wrap up this thesis by providing a conclusion summarizing the methods used in this work and the results that we obtained. Furthermore, we describe possible directions of future work building upon our findings.

8.1 Conclusion

In this thesis, we explored the design space of authentication for smart home assistants. New, security-sensitive tasks are coming to the voice interaction platform and call for appropriate authentication schemes enabling privacy-preserving mechanisms and protecting users from unauthorized access. State-of-the-art authentication for smart home assistants, consisting of vocalized PINs and biometric voice recognition was shown to be insufficient to protect users from casual insiders and targeted attacks by cybercriminals alike. A few alternative authentication schemes have been proposed in the past, however, to the best of our knowledge, no previous work has investigated the requirements for authentication systems from a users' point-of-view.

We closed this gap in the literature by reporting the results of a qualitative user study focusing on security-sensitive tasks on Amazon Alexa. During semi-structured interviews, we questioned participants about (1) their perceptions of threats, (2) mitigation strategies, and (3) design aspects which are crucial to a secure interaction experience, in the context of four scenarios involving high-risk tasks. The findings provided by a thematic analysis show that users are primarily concerned about bystanders that can eavesdrop on their interaction with Alexa, impacting their security and privacy. Furthermore, we highlight that the adoption of high-risk tasks using current state-of-the-art authentication mechanisms can be hindered by a lack of trust towards the voice code technology, as most users stated they would refrain from using the system as a coping strategy. Users strongly favoured biometric voice recognition as it was perceived as a natural and unobtrusive form

of authentication, however, most users noted that current systems were not satisfying their security requirements, e.g., due to being vulnerable to *Replay Attacks*.

Based on the insights gained from our user study, we provided design recommendations for future authentication systems. One such recommendation is based on a key finding that users have context-dependent requirements for authentication on smart home assistants. Based on the location of the interaction (e.g., inside the home vs outside) and the presence of bystanders (e.g., family members vs casual acquaintances) users felt more or less at risk. Since users requested interaction with smart home assistants to be effortless and straightforward, authentication methods should get in the way of the primary tasks as little as possible. Hence, the design of authentication systems needs to take these specific requirements into account and provide context-sensitive mechanisms with adequate security guarantees.

We concluded that authentication for smart home assistants is still an open field. Providing more sophisticated authentication mechanisms can enable security-sensitive tasks and potentially aid specific user groups such as people with visual impairments.

8.2 Future Work

Authentication for smart home assistants remains a wide-open field. As this study is exploratory, future work can evaluate the findings on a broader basis. Qualitative results can only provide us with an idea of the problem space and hypotheses building upon it need to be tested using quantitative measures. A possible quantitative follow up to our work could, e.g., implement the demo mode described in Section 7.2 and investigate the effects trial-and-error strategies have on users' perceptions of security and trust.

In our interview study, we used scenarios to compensate for unavailable functionality in our region. As security-sensitive tasks, such as online banking, become available to more and more users, it can be beneficial to study users' perceptions outside of a lab environment, e.g., via a diary study. This method can provide additional insights, which a lab study fails to provide.

Our study featured a diverse sample of participants, however, most were based in central Europe at the time of interviewing. To gain additional insights and to see how the results translate to other communities, future work could investigate the perceptions of different populations. As previous work, e.g., by Sambasivan et al. [Sam+18], has shown that it can be beneficial to investigate the security and privacy needs of diverse user groups. A promising exemplar of such an understudied user group are visually impaired users. Voice interaction has the potential to be usable without additional assistive technologies and might, therefore, become an interaction mechanism of choice for this group. As we found during our interviews, visually impaired users might have different perceptions of threats and mitigation strategies, e.g., when it comes to bystanders.

Finally, as we reported in Section 7.2, voice recognition is a desirable method of authentication for most users. However, as highlighted in Chapter 4, several exploits exist for

state-of-the-art voice biometrics. Therefore, future work can improve this authentication method further to ensure better security standards while keeping the already present usability aspects. A potential direction for improvement could be a liveliness detection able to distinguish human speech from speaker output, as desired by several study participants, see Section 6.1.

List of Figures

2.1 Examples of smart speakers hosting smart home assistants 6
2.2 High-level overview of a smart home assistant ecosystem 7

4.1 The *voice morphing* attack in detail 29
4.2 The *LightCommands* attack under realistic conditions 33

5.1 Vignette for scenario “Dinner” 46
5.2 Vignette for scenario “TV” 47
5.3 Vignette for scenario “Door” 48
5.4 Vignette for scenario “Hands” 49
5.5 Website used for remote interviews 52

6.1 Ideal data flow as depicted by P6 79

A.1 Original version of Figure 6.1, created by P6 122

List of Tables

5.1 Results of the demographic questionnaire 55

6.1 Approximate number of people marked by each quantifier 60
6.2 Final codes in the category *Attackers and Threats* 61
6.3 Final codes in the category *Users' Mitigation Strategies* 70

A.1 The final codebook including all codes and categories 121

Glossary

Likert scale The Likert scale is a research instrument used primarily in questionnaires to scale responses. It is named after its inventor Rensis Likert, who described it first in [Lik32]. 21

skill Voice-activated app for Amazon Alexa that complements the device with additional functionality. Similar to apps for smartphones, skills can be installed by the user from a central marketplace. 6, 7, 19, 24, 34, 39, 60, 61, 64, 66, 67, 69, 70, 78, 81, 82, 86, 90

whisper mode Whisper mode is a function of Amazon Alexa where a user can whisper to Alexa, who will then also whisper the answer back resulting in an much quieter interaction. 43, 73, 75, 87

Acronyms

2FA two-factor authentication. 11

API application programming interface. 34, 37

ATI affinity for technology interaction. 51, 55, 92

ATM automated teller machine. 11, 71, 81

AVS Alexa voice service. 34

CFIP concerns for information privacy. 51, 55, 92

CSI channel state information. 22

DOA direction of arrival. 21, 22

E2E end-to-end. 17, 18, 42

EER equal error rate. 30

ERB ethical review board. 58

HTTPS Hypertext Transfer Protocol Secure. 16, 17

IoT Internet of Things. 1, 2, 5, 7, 19, 20, 34, 38, 43, 45, 66, 69, 88

LED light-emitting diode. 33

MEMS micro-electro-mechanical systems. 23, 31, 32

MITM man-in-the-middle. 13, 34, 86

MOS mean opinion score. 30

MTurk Amazon Mechanical Turk. 21, 24

NFC near-field communication. 70

OTP one-time password. 10, 82, 120

PC personal computer. 78

PIN personal identification number. 9, 11, 12, 21, 35, 36, 38, 39, 44, 64–66, 76, 78, 80, 81, 93

RQ research question. 3, 4, 56, 68, 75

RSH robust sound hash. 21, 22

SMS short message service. 17, 18, 83

STT speech-to-text. 6, 31, 61, 67, 71, 86, 89

TAN transaction authentication number. 83

TLS Transport Layer Security. 17

TTS text-to-speech. 6, 23, 37, 61

TV television. 2, 22, 38, 39, 45, 47, 66–68, 74, 86, 97

VPN virtual private network. 14

VSButton Virtual Security Button. 22

VUI voice user interface. 22, 28, 43, 51

Bibliography

[Abu+17] Ruba Abu-Salma, Martina A. Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. "Obstacles to the Adoption of Secure Communication Tools". In: 2017 IEEE Symposium on Security and Privacy (S&P). San Jose, CA, USA: IEEE, May 2017. ISBN: 978-1-5090-5533-3. DOI: 10.1109/SP.2017.65.

[Abu+18] Ruba Abu-Salma, Elissa M. Redmiles, Blase Ur, and Miranda Wei. "Exploring User Mental Models of End-to-End Encrypted Communication Tools". In: 8th USENIX Workshop on Free and Open Communications on the Internet (FOCI 18). 2018. URL: https://www.usenix.org/conference/foci18/presentation/abu-salma (visited on 10/14/2020).

[AGK20] Simon Anell, Lea Gröber, and Katharina Krombholz. "End User and Expert Perceptions of Threats and Potential Countermeasures". In: 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). Sept. 2020.

[ARS19] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. "More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants". In: 15th Symposium on Usable Privacy and Security (SOUPS 2019). 2019. URL: https://www.usenix.org/conference/soups2019/presentation/abdi (visited on 10/14/2020).

[AS99] Anne Adams and Martina A. Sasse. "Users Are Not the Enemy". In: *Communications of the ACM* 42.12 (Dec. 1, 1999), pp. 40–46. ISSN: 0001-0782. DOI: 10.1145/322796.322806.

[BAF20] Julia Bernd, Ruba Abu-Salma, and Alisa Frik. "Bystanders' Privacy: The Perspectives of Nannies on Smart Home Surveillance". In: 10th USENIX Workshop on Free and Open Communications on the Internet (FOCI 20). 2020. URL: https://www.usenix.org/conference/foci20/presentation/bernd (visited on 10/14/2020).

[BC06] Virginia Braun and Victoria Clarke. "Using Thematic Analysis in Psychology". In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101.

[Blu+18] Logan Blue, Hadi Abdullah, Luis Vargas, and Patrick Traynor. "2MA: Verifying Voice Commands via Two Microphone Authentication". In: *Proceedings of the 2018 on Asia Conference on Computer and Communications Security* (Incheon, Republic of Korea). ASIACCS '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 89–100. ISBN: 978-1-4503-5576-6. DOI: 10.1145/3196494.3196545.

[BR99] Christine Barter and Emma Renold. "The Use of Vignettes in Qualitative Research". In: *Social Research Update* 25.9 (1999), pp. 1–6.

[BSR19] Matt Bishop, Elisabeth Sullivan, and Michelle Ruppel. *Computer Security: Art and Science*. Second edition. Boston: Addison-Wesley, 2019. ISBN: 978-0-321-71233-2.

[Car+16] 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). 2016. ISBN: 978-1-931971-32-4. URL: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini (visited on 10/14/2020).

[Cho19] Eugene Cho. "Hey Google, Can I Ask You Something in Private?" In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–9. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300488.

[Cov05] Lynne Coventry. "Usable Biometrics". In: *Security and Usability: Designing Secure Systems That People Can Use*. Ed. by Lorrie F. Cranor and Simson Garfinkel. O'Reilly, 2005. ISBN: 978-0-596-00827-7.

[Eck18] Claudia Eckert. *IT-Sicherheit: Konzepte, Verfahren, Protokolle*. 10. Auflage. De Gruyter Studium. München: De Gruyter Oldenburg, 2018. ISBN: 978-3-11-055158-7.

[Eib+17] Malin Eiband, Mohamed Khamis, Emanuel von Zezschwitz, Heinrich Hussmann, and Florian Alt. "Understanding Shoulder Surfing in the Wild: Stories from Users and Observers". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. New York, NY, USA: Association for Computing Machinery, May 2, 2017, pp. 4254–4265. ISBN: 978-1-4503-4655-9. DOI: 10.1145/3025453.3025636.

[FAW19] Thomas Franke, Christiane Attig, and Daniel Wessel. "A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale". In: *International Journal of Human–Computer Interaction* 35.6 (Apr. 3, 2019), pp. 456–467. ISSN: 1044-7318. DOI: 10.1080/10447318.2018.1456150.

[FFS17] Huan Feng, Kassem Fawaz, and Kang G. Shin. "Continuous Authentication for Voice Assistants". In: *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. MobiCom '17. Snowbird, Utah, USA: Association for Computing Machinery, 2017, pp. 343–355. DOI: 10.1145/3117811.3117823.

[Fri+19] Alisa Frik, Leysan Nurgalieva, Julia Bernd, Joyce Lee, Florian Schaub, and Serge Egelman. "Privacy and Security Threat Models and Mitigation Strategies of Older Adults". In: 15th Symposium on Usable Privacy and Security (SOUPS 2019). 2019. URL: https://www.usenix.org/conference/soups2019/presentation/frik (visited on 10/14/2020).

[He+18] Weijia He, Maximilian Golla, Roshni Padhi, Jordan Ofek, Markus Dürmuth, Earlence Fernandes, and Blase Ur. "Rethinking Access Control and Authentication for the Home Internet of Things (IoT)". In: 27th USENIX Security Symposium (USENIX Security 18). 2018. ISBN: 978-1-931971-46-1. URL: https://www.usenix.org/conference/usenixsecurity18/presentation/he (visited on 10/14/2020).

[HP18] David Harborth and Sebastian Pape. "German Translation of the Concerns for Information Privacy (CFIP) Construct". In: *SSRN Scholarly Paper* ID 3112207 (Jan. 29, 2018). URL: https://papers.ssrn.com/abstract=3112207 (visited on 10/14/2020).

[IRC15] Iulia Ion, Rob Reeder, and Sunny Consolvo. ""...No One Can Hack My Mind": Comparing Expert and Non-Expert Security Practices". In: 11th Symposium On Usable Privacy and Security (SOUPS 2015). 2015. ISBN: 978-1-931971-24-9. URL: https://www.usenix.org/conference/soups2015/proceedings/presentation/ion (visited on 10/14/2020).

[JAE16] Artur Janicki, Federico Alegre, and Nicholas Evans. "An Assessment of Automatic Speaker Verification Vulnerabilities to Replay Spoofing Attacks". In: *Security and Communication Networks* 9.15 (2016), pp. 3030–3044. ISSN: 1939-0122. DOI: 10.1002/sec.1499.

[Jer+99] Ian Jermyn, Alain Mayer, Fabian Monrose, Michael K. Reiter, and Aviel D. Rubin. "The Design and Analysis of Graphical Passwords". In: *Proceedings of the 8th Conference on USENIX Security Symposium - Volume 8*. SSYM'99. Washington, D.C.: USENIX Association, 1999.

[JGL98] Philip N. Johnson-Laird, Vittorio Girotto, and Paolo Legrenzi. "Mental Models: A Gentle Guide for Outsiders". In: *Sistemi Intelligenti* 9.68 (1998), pp. 33–46.

[Jia+18] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. "Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis". In: *Proceedings of the 32nd International Conference*

on Neural Information Processing Systems. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., Dec. 3, 2018, pp. 4485–4495.

[Kan+15] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. ""My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security". In: 11th Symposium On Usable Privacy and Security (SOUPS 2015). 2015. ISBN: 978-1-931971-24-9. URL: https://www.usenix.org/conference/soups2015/proceedings/presentation/kang (visited on 10/14/2020).

[Kri11] Klaus Krippendorff. "Computing Krippendorff's Alpha-Reliability". In: *Annenberg School for Communication Departmental Papers* (2011). URL: https://repository.upenn.edu/asc_papers/43/ (visited on 10/14/2020).

[Kro+19] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel von Zezschwitz. ""If HTTPS Were Secure, I Wouldn't Need 2FA" - End User and Administrator Mental Models of HTTPS". In: 2019 IEEE Symposium on Security and Privacy (S&P). San Francisco, CA, USA: IEEE, May 2019. ISBN: 978-1-5386-6660-9. DOI: 10.1109/SP.2019.00060.

[KSM14] Elie Khoury, Laurent E. Shafey, and Sébastien Marcel. "Spear: An Open Source Toolbox for Speaker Recognition Based on Bob". In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, May 2014. DOI: 10.1109/ICASSP.2014.6853879.

[Kum+18] 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). 2018. ISBN: 978-1-939133-04-5. URL: https://www.usenix.org/conference/usenixsecurity18/presentation/kumar (visited on 10/14/2020).

[Kwa+19] Il-Youp Kwak, Jun H. Huh, Seung T. Han, Iljoo Kim, and Jiwon Yoon. "Voice Presentation Attack Detection through Text-Converted Voice Command Analysis". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–12. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300828.

[Lei+18] Xinyu Lei, Guan-Hua Tu, Alex X. Liu, Chi-Yu Li, and Tian Xie. "The Insecurity of Home Digital Voice Assistants - Vulnerabilities, Attacks and Countermeasures". In: 2018 IEEE Conference on Communications and Network Security (CNS). Beijing, China: IEEE, May 2018. ISBN: 978-1-5386-4586-4. DOI: 10.1109/CNS.2018.8433167.

[LFH17] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research Methods in Human-Computer Interaction*. Second edition. Cambridge, MA: Morgan Kaufmann Publishers, 2017. ISBN: 0-12-805390-9. URL: http://www.sciencedirect.com/science/book/9780128053904 (visited on 10/14/2020).

[Lik32] Rensis Likert. "A Technique for the Measurement of Attitudes." In: *Archives of Psychology* 22 (1932), pp. 5–55.

[LS20] Karen Levy and Bruce Schneier. "Privacy Threats in Intimate Relationships". In: *Journal of Cybersecurity* 6.1 (Jan. 1, 2020), pp. 1–13. ISSN: 2057-2085. DOI: 10.1093/cybsec/tyaa006.

[LZS18] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. "Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers". In: *Proceedings of the ACM on Human-Computer Interaction* 2 (CSCW Nov. 1, 2018), 102:1–102:31. DOI: 10.1145/3274371.

[Mar+16] Diogo Marques, Ildar Muslukhov, Tiago Guerreiro, Luís Carriço, and Konstantin Beznosov. "Snooping on Mobile Phones: Prevalence and Trends". In: 12th Symposium on Usable Privacy and Security (SOUPS 2016). 2016. ISBN: 978-1-931971-31-7. URL: https://www.usenix.org/conference/soups2016/technical-sessions/presentation/marques (visited on 10/14/2020).

[MCG18] Michael McTear, Zoraida Callejas, and David Griol. *The Conversational Interface: Talking to Smart Devices*. 1st. Springer Publishing Company, Incorporated, 2018. ISBN: 3-319-81411-7. DOI: 10.1007/978-3-319-32967-3.

[Mil94] Benjamin Miller. "Vital Signs of Identity". In: *IEEE Spectrum* 31.2 (Feb. 1994), pp. 22–30. ISSN: 1939-9340. DOI: 10.1109/6.259484.

[MMS19] Richard Mitev, Markus Miettinen, and Ahmad-Reza Sadeghi. "Alexa Lied to Me: Skill-Based Man-in-the-Middle Attacks on Virtual Assistants". In: *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*. Asia CCS '19. New York, NY, USA: Association for Computing Machinery, July 2, 2019, pp. 465–478. ISBN: 978-1-4503-6752-3. DOI: 10.1145/3321705.3329842.

[MSS15] Dibya Mukhopadhyay, Maliheh Shirvanian, and Nitesh Saxena. "All Your Voices Are Belong to Us: Stealing Voices to Fool Humans and Machines". In: *Computer Security – ESORICS 2015*. Ed. by Günther Pernul, Peter Y. A. Ryan, and Edgar Weippl. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 599–621. ISBN: 978-3-319-24177-7. DOI: 10.1007/978-3-319-24177-7_30.

[MW05] Robert C. Miller and Min Wu. “The Memorability and Security of Passwords”. In: *Security and Usability: Designing Secure Systems That People Can Use*. Ed. by Lorrie F. Cranor and Simson Garfinkel. O’Reilly, 2005. ISBN: 978-0-596-00827-7.

[Pea+19] Sarah Pearman, Shikun A. Zhang, Lujo Bauer, and Nicolas Christin. “Why People (Don’t) Use Password Managers Effectively”. In: 15th Symposium on Usable Privacy and Security (SOUPS 2019). 2019. URL: https://www.usenix.org/conference/soups2019/presentation/pearman (visited on 10/14/2020).

[Por+18] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. “Voice Interfaces in Everyday Life”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. New York, NY, USA: Association for Computing Machinery, 2018, 640:1–640:12. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3174214.

[PSC05] Ugo Piazzalunga, Paolo Salvaneschi, and Paolo Coffetti. “The Usability of Security Devices”. In: *Security and Usability: Designing Secure Systems That People Can Use*. Ed. by Lorrie F. Cranor and Simson Garfinkel. O’Reilly, 2005. ISBN: 978-0-596-00827-7.

[Rei+17] Dennis Reineck, Volker Lilienthal, Annika Sehl, and Stephan Weichert. “Das faktorielle Survey. Methodische Grundsätze, Anwendungen und Perspektiven einer innovativen Methode für die Kommunikationswissenschaft”. In: *M&K Medien & Kommunikationswissenschaft* 65.1 (2017), pp. 101–116. ISSN: 1615-634X. DOI: 10.5771/1615-634X-2017-1-101.

[Ren05] Karen Renaud. “Evaluating Authentication Mechanisms”. In: *Security and Usability: Designing Secure Systems That People Can Use*. Ed. by Lorrie F. Cranor and Simson Garfinkel. O’Reilly, 2005. ISBN: 978-0-596-00827-7.

[RHR17] Nirupam Roy, Haitham Hassanieh, and Romit R. Choudhury. “BackDoor: Making Microphones Hear Inaudible Sounds”. In: 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys ’17). Niagara Falls, New York, USA: Association for Computing Machinery, 2017, pp. 2–14. ISBN: 978-1-4503-4928-4. DOI: 10.1145/3081333.3081366.

[Ros06] Ellen A. Rose. “An Examination of the Concern for Information Privacy in the New Zealand Regulatory Context”. In: *Information & Management* 43.3 (Apr. 1, 2006), pp. 322–335. ISSN: 0378-7206. DOI: 10.1016/j.im.2005.08.002.

[Roy+18] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit R. Choudhury. “Inaudible Voice Commands: The Long-Range Attack and Defense”. In: 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18). 2018. ISBN: 978-1-939133-01-4. URL: https://www.usenix.org/conference/nsdi18/presentation/roy (visited on 10/14/2020).

[RRF04] Volker Roth, Kai Richter, and Rene Freidinger. “A PIN-Entry Method Resilient against Shoulder Surfing”. In: *Proceedings of the 11th ACM Conference on Computer and Communications Security*. CCS ’04. New York, NY, USA: Association for Computing Machinery, Oct. 25, 2004, pp. 236–245. ISBN: 978-1-58113-961-7. DOI: 10.1145/1030083.1030116.

[RVR14] Karen Renaud, Melanie Volkamer, and Arne Renkema-Padmos. “Why Doesn’t Jane Protect Her Privacy?” In: *Privacy Enhancing Technologies*. Ed. by Emiliano D. Cristofaro and Steven J. Murdoch. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 244–262. ISBN: 978-3-319-08506-7. DOI: 10.1007/978-3-319-08506-7_13.

[Sam+18] Nithya Sambasivan, Garen Checkley, Amna Batool, Nova Ahmed, David Nemer, Laura S. Gaytán-Lugo, Tara Matthews, Sunny Consolvo, and Elizabeth Churchill. “"Privacy Is Not for Me, It’s for Those Rich Women": Performative Privacy Practices on Mobile Phones by Women in South Asia”. In: 14th Symposium on Usable Privacy and Security (SOUPS 2018). 2018. ISBN: 978-1-939133-10-6. URL: https://www.usenix.org/conference/soups2018/presentation/sambasivan (visited on 10/14/2020).

[SC15] Anselm Strauss and Juliet Corbin. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Sage Publications, Inc., 2015. ISBN: 978-1-4129-9746-1.

[Sch+20] Lea Schönherr, Maximilian Golla, Thorsten Eisenhofer, Jan Wiele, Dorothea Kolossa, and Thorsten Holz. “Unacceptable, Where Is My Privacy? Exploring Accidental Triggers of Smart Speakers”. In: *arXiv preprint arXiv:2008.00508 [cs.CR]* (2020). URL: https://arxiv.org/abs/2008.00508 (visited on 10/14/2020).

[Sha+10] Richard Shay, Saranga Komanduri, Patrick G. Kelley, Pedro G. Leon, Michelle L. Mazurek, Lujo Bauer, Nicolas Christin, and Lorrie F. Cranor. “Encountering Stronger Password Requirements: User Attitudes and Behaviors”. In: *Proceedings of the 6th Symposium on Usable Privacy and Security*. SOUPS ’10. New York, NY, USA: Association for Computing Machinery, July 14, 2010, pp. 1–20. ISBN: 978-1-4503-0264-7. DOI: 10.1145/1837110.1837113.

[SMB96] Jeff H. Smith, Sandra J. Milberg, and Sandra J. Burke. “Information Privacy: Measuring Individuals’ Concerns About Organizational Practices”. In: *MIS Q.* (1996). DOI: 10.2307/249477.

[SSS00] Drik Scheuermann, Scarlet Schwiderski-Grosche, and Bruno Struif. *Usability of Biometrics in Relation to Electronic Signatures*. GMD-Report. GMD-Forschungszentrum Informationstechnik, 2000.

[Sug+20] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. "Light Commands: Laser-Based Audio Injection Attacks on Voice-Controllable Systems". In: 29th USENIX Security Symposium (USENIX Security 20). 2020. ISBN: 978-1-939133-17-5. URL: https://www.usenix.org/conference/usenixsecurity20/presentation/sugawara (visited on 10/14/2020).

[SWH16] Robert C. Streijl, Stefan Winkler, and David S. Hands. "Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives". In: *Multimedia Systems* 22.2 (Mar. 1, 2016), pp. 213–227. ISSN: 1432-1882. DOI: 10.1007/s00530-014-0446-1.

[Vai+15] 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 (WOOT15). 2015. URL: https://www.usenix.org/conference/woot15/workshop-program/presentation/vaidya (visited on 10/14/2020).

[Was10] Rick Wash. "Folk Models of Home Computer Security". In: *Proceedings of the 6th Symposium on Usable Privacy and Security*. SOUPS '10. Redmond, Washington, USA: Association for Computing Machinery, July 14, 2010, pp. 1–16. ISBN: 978-1-4503-0264-7. DOI: 10.1145/1837110.1837125.

[WT99] Alma Whitten and Doug J. Tygar. "Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0." In: USENIX Security Symposium. Vol. 348. 1999. URL: https://www.usenix.org/legacy/publications/library/proceedings/sec99/whitten.html (visited on 10/14/2020).

[Yan+05] Jeff Yan, Adam Blackwell, Ross Anderson, and Alasdair Grant. "The Memorability and Security of Passwords". In: *Security and Usability: Designing Secure Systems That People Can Use*. Ed. by Lorrie F. Cranor and Simson Garfinkel. O'Reilly, 2005. ISBN: 978-0-596-00827-7.

[Yua+18] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, Xiaofeng Wang, and Carl A. Gunter. "CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition". In: 27th USENIX Security Symposium (USENIX Security 18). 2018. ISBN: 978-1-939133-04-5. URL: https://www.usenix.org/conference/usenixsecurity18/presentation/yuan-xuejing (visited on 10/14/2020).

[Zha+17] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. "DolphinAttack: Inaudible Voice Commands". In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. CCS '17. Dallas, Texas, USA: Association for Computing Machinery, 2017, pp. 103–117. ISBN: 978-1-4503-4946-8. DOI: 10.1145/3133956.3134052.

[ZMR17] Eric Zeng, Shrirang Mare, and Franziska Roesner. "End User Security and Privacy Concerns with Smart Homes". In: 13th Symposium on Usable Privacy and Security (SOUPS 2017). 2017. ISBN: 978-1-931971-39-3. URL: https://www.usenix.org/conference/soups2017/technical-sessions/presentation/zeng (visited on 10/14/2020).

[ZR19] Eric Zeng and Franziska Roesner. "Understanding and Improving Security and Privacy in Multi-User Smart Homes: A Design Exploration and In-Home User Study". In: 28th USENIX Security Symposium (USENIX Security 19). 2019. ISBN: 978-1-939133-06-9. URL: https://www.usenix.org/conference/usenixsecurity19/presentation/zeng (visited on 10/14/2020).

[ZTY17] Linghan Zhang, Sheng Tan, and Jie Yang. "Hearing Your Voice Is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication". In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. CCS '17. Dallas, Texas, USA: Association for Computing Machinery, 2017, pp. 57–71. ISBN: 978-1-4503-4946-8. DOI: 10.1145/3133956.3133962.

Online References

[Amaa] Amazon. *Add Personalization to Your Skill | Alexa Skills Kit*. URL: https://developer.amazon.com/en-US/docs/alexa/custom-skills/add-personalization-to-your-skill.html (visited on 10/14/2020).

[Amab] Amazon. *Amazon Alexa Official Site: What Is Alexa?* URL: https://developer.amazon.com/en-US/alexa (visited on 10/14/2020).

[Amac] Amazon. *Amazon.Com Help: What Do the Lights on Your Echo Device Mean?* URL: https://www.amazon.com/gp/help/customer/display.html?nodeId=GKLDRFT7FP4FZE56 (visited on 10/14/2020).

[Amad] Amazon. *Amazon.Com Hilfe: Manage Voice Profiles for Purchases with Alexa*. URL: https://www.amazon.com/gp/help/customer/display.html?nodeId=GKW8N2S4UQ2SAEBL (visited on 10/14/2020).

[Amae] Amazon. *Amazon.Com Hilfe: Require a Voice Code for Purchases with Alexa*. URL: https://www.amazon.com/gp/help/customer/display.html?nodeId=GAA2RYUEDNT5ZSNK (visited on 10/14/2020).

[Amaf] Amazon. *Amazon.Com: Amazon Echo & Alexa Devices: Amazon Devices & Accessories*. URL: https://www.amazon.com/smart-home-devices/b?ie=UTF8&node=9818047011 (visited on 10/14/2020).

[Aud] Audacity. *Audacity Download*. URL: https://www.audacity.de/ (visited on 10/14/2020).

[Aug] August. *Control Your August Smart Lock with Amazon Alexa | August*. URL: https://august.com/pages/alexa (visited on 10/14/2020).

[Cap] Capital One. *Capital One Is on Amazon Echo. Questions? Just Ask Alexa.* URL: https://www.capitalone.com/applications/alexa/ (visited on 10/14/2020).

[Dig] Bundesministerium für Digitalisierung und Wirtschaftsstandort. *PayPass/NFC-Kontaktlos – das kontaktlose Bezahlen.* URL: https://www.oesterreich.gv.at/themen/steuern_und_finanzen/bankgeschaefte/Seite.750281.html (visited on 10/14/2020).

[FAW] Thomas Franke, Christiane Attig, and Daniel Wessel. *ATI Scale.* URL: https://ati-scale.org/ (visited on 10/14/2020).

[Gooa] Google. *Google Assistant, Your Own Personal Google.* URL: https://assistant.google.com/ (visited on 10/14/2020).

[Goob] Google. *Google Formulare: Kostenlos Umfragen Erstellen Und Analysieren.* URL: https://www.google.com/forms/about/ (visited on 10/14/2020).

[Gooc] Google. *Google Home – Intelligenter Lautsprecher Und Home Assistant – Google Store.* URL: https://store.google.com/product/google_home_speaker (visited on 10/14/2020).

[Good] Google. *Sprachausgabe: Lebensechte Sprachsynthese | Cloud Text-to-Speech.* URL: https://cloud.google.com/text-to-speech?hl=de (visited on 10/14/2020).

[honl] heise online. *Girocard kontaktlos: Limit für Zahlungen ohne PIN auf 50 Euro erhöht.* URL: https://www.heise.de/newsticker/meldung/Girocard-kontaktlos-Limit-fuer-Zahlungen-ohne-PIN-auf-50-Euro-erhoeht-4693250.html (visited on 10/14/2020).

[iFL] iFLYTEK. *iFLYTEK Open Platform-China's First Artificial Intelligence Open Platform for Mobile Internet and Intelligent Hardware Developers.* URL: https://global.xfyun.cn/ (visited on 10/14/2020).

[Lat] Literature and Latte. *Scapple | Literature & Latte.* URL: https://www.literatureandlatte.com/scapple/overview (visited on 10/14/2020).

[Lip17] Andrew Liptak. *Amazon's Alexa Started Ordering People Dollhouses after Hearing Its Name on TV.* Jan. 7, 2017. URL: https://www.theverge.com/2017/1/7/14200210/amazon-alexa-tech-news-anchor-order-dollhouse (visited on 10/14/2020).

[Moz20] Mozilla. *Mozilla/DeepSpeech.* Mozilla. Sept. 16, 2020. URL: https://github.com/mozilla/DeepSpeech (visited on 10/14/2020).

[Pov] Daniel Povey. *Kaldi ASR.* URL: http://kaldi-asr.org/ (visited on 10/14/2020).

[Pro20] CMU Festvox Project. *Festvox/Festvox.* Sept. 9, 2020. URL: https://github.com/festvox/festvox (visited on 10/14/2020).

[Til16] Aaron Tilley. *How A Few Words To Apple's Siri Unlocked A Man's Front Door*. Sept. 21, 2016. URL: https://www.forbes.com/sites/aarontilley/2016/09/21/apple-homekit-siri-security/ (visited on 10/14/2020).

[Wie07] TU Wien. *Code of Conduct – Rules to Ensure Good Scientific Practice, Decision by the Chancellor's Office of 23 October 2007*. Oct. 23, 2007. URL: https://www.tuwien.at/index.php?eID=dms&s=4&path=Richtlinien%20und%20Verordnungen/Code%20of%20Conduct%20fuer%20wissenschaftliches%20Arbeiten.pdf (visited on 10/14/2020).

APPENDIX A

Appendix

In the Appendix, we present complementary material relevant to our work.

A.1 Interview Guideline

The guideline we used for our interviews looked as follows, note that italic text indicates actions taken by the interviewer.

Introduction

Greet participant and introduce topic: "Hi, thank you for taking part in this interview." *Present interviewee with consent sheet, explaining purpose of the study.* "In the following, I will ask you some questions where I'm interested in your personal opinions and experiences, so keep in mind there are no wrong answers. If you feel like drawing anything throughout the interview, feel free to use this pen and paper here. Do you have any questions?" *Answer questions of interviewee, if any.* "So let's start with the first question!"

Question: "How long are you using Alexa already?"
Alternative: "When was your first contact with Alexa"
Question: What devices are you using Alexa on?
Question: Where are those devices usually located?
Question: What are some typical tasks you perform with Alexa?
Question: Did you ever use Alexa for online shopping?
If yes: "Did you encounter any issues while doing so?"
If no: "Where there specific reasons for you not to use this feature?"

Scenarios

Lead over to scenarios: "Thank you for your answers so far. Now I would like you to have a look at some scenarios. For this interview, let's assume that all of the following features are implemented in Alexa, even though some of them are not currently available."

"Now I would like you to please take one of the scenario cards, have a look at it and read it aloud." *Let interviewee choose a card and flip it over.*

Question: "Please identify any issues that could arise in such a situation?"
Follow up: "Why do you think that is problematic?"
Question: "Can you identify threats for the user in such a scenario?"
Question: "Who could be the source of such a threat?"
Question: "What would you do to protect yourself?"

Transition to next scenario: "Great, let's continue with the next scenario. However, we can always come back to a previous scenario if you want to add something." *Repeat process for all four scenarios.*

Question: "Now that you have seen all four scenarios, what do you think they have in common?"

Demographics

Conclude scenario part and retrieve demographic data: "Thank you for the collaboration so far. Please take the tablet and fill out the questionnaire there." *Hand tablet to interviewee to complete the questionnaire.*

Finally

Question: "Do you have any final questions or marks you would like to make?"

Thank interviewee for their collaboration and bid farewell: "Thank you again for your participation and have a nice day!"

A.2 Codebook

Table A.1 includes the final codebook we obtained during the thematic analysis of our interview data. Codes are separated by category and reported along the number of overall occurrences. Note that a code could be assigned to more than one category.

Code	Frequency
Attackers and Threats	
Accidents as threat	49
Amazon listening in on conversations	8
Bystanders as threat	51
Criminals as threat	42
Cyberattacks as threat	40
Insiders as threat	39
Malicious skills as threat	2
Pranks as threat	19
Sharing data with Amazon undesirable	95
Biometric Authentication	
Annoyance of false negatives when using biometrics	3
Authentication via voice recognition desirable	50
Risk of false positives when using biometrics	16
Uncertainty about security of voice recognition	16
User wants Alexa in combination with face recognition	17
Voice recognition should distinguish live voice from replays	4
Building Trust	
Build/Lose trust through interaction experience	51
Build trust in security mechanism via trial-and-error	6
Trust from reviews	5
Trust in familiar people	41
Trust in system is transferred from trustworthy entity	39
Knowledge-based Authentication	
Enter voice code via smartphone rather than Alexa	25
High number of voice codes difficult to remember and distinguish	30

User wishes for duress code	2
Voice code protects against unauthorized access	27
Whispering the voice code protects against eavesdropping	3
Optimistic Authentication	
Optimistic authentication does not protect from physical access	1
Optimistic authentication via delayed verification	35
Perceptions of Alexa	
Insufficient mental model	17
Personification of Alexa	15
Uncertainty about security of Alexa ecosystem	22
Perceptions of Authentication	
Properties of authentication method are transferred from other systems	86
Possessions-based Authentication	
Risk of Replay Attacks when using tokens	2
User wishes for Alexa in combination with OTP	25
User wishes for Alexa in combination with token-based authentication	39
Public Sphere of Alexa Interaction	
Openness of voice interaction security/privacy relevant	85
Reconnaissance of Alexa easily possible	4
Requirements of Authentication	
Multiple users use Alexa in parallel	10
Risk Assessment of Alexa Authentication	
Minimal protection by law	8
Users notice acoustic attacks on their Alexa if they are present	3
User wishes for multiple authentication steps	37
Variable security requirements depending on location	42
Variable security requirements depending on presence of user	9
Weighing up risks and effort of authentication	65
Risk-Benefit Analysis of Alexa	
Alexa needs right to exist	117
Refrain from using the system due to security reasons	37

Weighing up use against increased exposure to risk	30
Social Aspects of Alexa Use	
Hierarchy among Alexa users	1
Take time for important actions	6
Using Alexa means being lazy	24
Voice interaction inappropriate in specific social situations	31
Transparency and Agency	
User wishes for agency over transparent processes	86
Users' Mitigation Strategies	
Build trust in security mechanism via trial-and-error	6
Change voice code regularly	11
Move to another room to use Alexa	30
Refrain from using the system due to security reasons	37
Take time for important actions	6
Users notice acoustic attacks on their Alexa if they are present	3
Voice code protects against unauthorized access	27
Voice interaction inappropriate in specific social situations	31
Whispering the voice code protects against eavesdropping	3

Table A.1: The final codebook including all codes and categories

A.3 Participant Drawings

During our interviews, participants could use pen and paper provided by the interviewer and prepare drawings to express themselves. P6 was the only interviewee to exercise their option, the resulting drawing is depicted in Figure A.1, a translated reproduction including a detailed description is depicted in Figure 6.1.

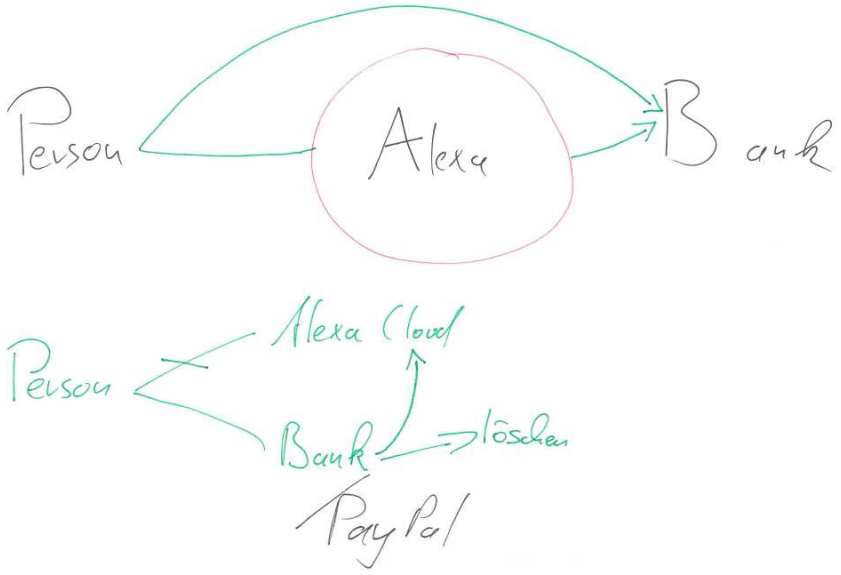

Figure A.1: Original version of Figure 6.1, created by P6