User-Centered Biometric Interfaces

Dissertation

an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von
M.Sc. Medieninformatik

Lukas Benjamin Mecke

München, den 28. September 2023

Erstgutachter:	Prof. Dr. Florian Alt
Zweitgutachter:	Prof. Dr. Max Mühlhäuser
Drittgutachter:	Prof. Dr. Daniel Buschek

Tag der Abgabe: 28. September 2023
Tag der mündlichen Prüfung: 19. Dezember 2023

ABSTRACT

Authentication has become an essential part of our daily lives. Examples include using authentication tokens like keys to enter a building or vehicle, or the use of passwords, PINs, and patterns to access digital accounts and devices. However, such traditional approaches start to reach their limits, as the ever-increasing number of required authentications strains both users' memory and time.

Biometric methods make use of unique patterns in user physiology or behavior for the purpose of authentication and are proposed as a potential solution. They do not require mental effort, cannot be stolen or forgotten, and can operate in the background with no active user engagement required. However, biometrics also come with drawbacks: their underlying machine learning models mostly act as black-boxes to users while at the same time being prone to systemic biases and inconsistent recognition performance. Users get little insight into what constitutes model decisions, let alone control over the authentication mechanism that is to protect their data.

This thesis takes a user-centered approach to both enhance existing interfaces with biometric systems and propose new ones with the aim to facilitate 1) user literacy and 2) agency over the recognition process. We conducted studies to understand user preferences and needs with regard to biometrics and designed biometric interfaces to support users in understanding influencing factors on their authentication and take control over if and when to be recognized. Overall, we explored how biometric interfaces could look like, how they could improve interaction with biometric systems, and if they can contribute to an informed and secure use of biometric authentication.

Zusammenfassung

Authentifizierung ist heute zu einem permanenten Bestandteil unseres täglichen Lebens geworden. Beispiele dafür sind die Verwendung von Authentifizierungstoken wie Schlüsseln, um ein Gebäude zu betreten oder ein Fahrzeug zu öffnen, oder der Gebrauch von Passwörtern, PINs und Mustern für den Zugriff auf Geräte und digitale Konten. Diese traditionellen Ansätze stoßen jedoch allmählich an ihre Grenzen, da die ständig wachsende Zahl der erforderlichen Authentifizierungen zu einer zunehmenden Belastung für Gedächtnis und Zeit der Benutzenden wird.

Biometrische Methoden nutzen einzigartige Muster in der menschlichen Physiologie oder im Verhalten der Nutzenden für die Authentifizierung und können eine mögliche Lösung für diese Herausforderung darstellen. Sie erfordern keine geistige Anstrengung, können nicht gestohlen oder vergessen werden und arbeiten im Hintergrund, ohne dass Benutzende aktiv werden müssen. Biometrische Verfahren haben jedoch auch Nachteile: Die zugrundeliegenden maschinellen Lernmodelle sind meist undurchschaubar, während sie gleichzeitig anfällig für voreingenommene Entscheidungen und inkonsistente Erkennungsleistungen sind. Die Nutzenden erhalten kaum Einblick in die Entscheidungen der Modelle, geschweige denn Kontrolle über den Authentifizierungsmechanismus, der ihre Daten schützen soll.

In dieser Arbeit wird ein nutzer-zentrierter Ansatz verfolgt, um sowohl bestehende Schnittstellen zu biometrischen Systemen zu verbessern als auch neue vorzuschlagen. Die Ziele dieser Arbeit sind es 1) die Kenntnis der Nutzenden zu erweitern und ihnen 2) die Kontrolle über biometrische Modelle zu erleichtern. Wir haben Studien durchgeführt, um die Präferenzen und Bedürfnisse in Bezug auf biometrische Systeme zu verstehen und biometrische Schnittstellen zu entwerfen, die die Nutzenden dabei unterstützen, Einflussfaktoren auf ihre Authentifizierung zu verstehen und die Kontrolle darüber zu übernehmen, ob und wann sie erkannt werden. Insgesamt haben wir untersucht, wie biometrische Schnittstellen aussehen könnten, wie sie die Interaktion mit biometrischen Systemen verbessern können und ob sie zu einer informierten und sicheren Nutzung biometrischer Authentifizierungmechanismen beitragen können.

ACKNOWLEDGMENTS

Throughout the journey of completing this dissertation, I have been fortunate enough to receive support and encouragement from many individuals whose contributions have been invaluable to this work. Each interaction, whether through guidance, inspiration, or camaraderie, has greatly contributed to my research and personal growth. To all those who have offered their time, insights, and unwavering belief in my abilities, please know that your impact is deeply appreciated and will always be cherished, even if your names are not explicitly mentioned here. Thank you all for being an integral part of this journey.

Thank you, **Florian (A.)**, for sparking my interest in research when accepting me as a bachelor's student over ten years ago and for being a positive voice of support, inspiration, and encouragement ever since. You are a big part of why I chose to pursue a Ph.D to begin with. Thank you for supervising my thesis, taking me on a road trip through the USA, and sharing many more unforgettable trips and experiences. Thank you, **Daniel (B.)**, for reviewing my thesis and providing advice and a second opinion whenever I was in doubt. Thank you for sharing your expertise in machine learning, supporting my first steps as an office mate at HM, and for many great experiences like the escape rooms and our various trips. Thank you, **Max** and **Albrecht**, for completing my committee and making my defense possible.

Next, I would like to thank my colleagues at the Usable Security group at UniBW. Thank you, **Sarah (P.)**, for being the best office mate I could have wished for and for sharing this journey with me from the very beginning. Thank you for working together on countless projects (including many in this thesis) and for many cherished memories from our conference trips, to San Francisco, Boston, Edinburgh, and many more. Thank you for sharing your leg space on long flights and for explaining biometrics together. Thank you for not only being a great colleague but also becoming a good friend. Thank you, **Sarah (D.R.)**, for supporting countless of my projects since the very start of my PhD, sharing my passion for 3D printing (and making things in general), and physically realizing many of our ideas. Thank you for the awesome shared trip to Ireland. Thank you, **Verena (D)**, for discussing study designs, statistics, large language models, and so much more (on- and off-topic) during the coffee breaks. Your expertise was invaluable to this thesis. Thank you, **Felix**, for many a great chat, tech support, cool joint projects, and sharing pesto and pasta. Thank you, **Oliver (H.)**, for teaching usable security with me and discussing and exploring new technologies. Thanks to my former colleagues **Ken**, **Mariam**, **Michael (B. & F.)**, **Rivu**, **Ville**, **Yasmeen** and **Yomna** for many thoughtful and inspiring conversations and cool and successful joint projects. Thank you, **Pascal**, for the unique achievement of sharing a room but never physically meeting during CHI in Hamburg. Thanks also to **Doruntina**, **Robin**, **Verena (W.)** and **Victoria** who more recently joined the group but already made a positive impact on my work and live in the office. I am looking forward to getting to know each other better and starting many more interesting projects. Finally, thank you, **Heike**, for supporting me with all my paperwork, organizing our events, and inviting me to a great dinner. Thank you for always being available for questions or just having a chat. Thanks to all of you for your support, and for making work and coming to the office as enjoyable as it was.

Thank you also to the great people beyond our group that I had the pleasure to work with or whom I met throughout this journey. Thanks to **Veronika**, **Peter**, **Erik**, **Thomas (V.)**, and **Andrea (M.)** for providing a great environment at HM in the first year and a half of my PhD. Thanks to **Stefan**, **Alia**, **Jonathan** and **Uwe** for being great partners in the *Scalable Biometrics* project, for hosting me many times when visiting Essen, for working together on many projects in this thesis and beyond, and for sharing drinks and burgers. Thank you, **Hai** and **Florian (L)**, for exploring Large Language Models and generative AI with me. Thank you, **Mohamed**, for supervising my master's thesis and being a great support whenever we met. Thank you, **Matthias (H.)**, for sharing ramen and eggs benedict in Montreal. Thank you, **Heiko**, for delving into the intricacies of eye tracker calibration together. Thank you, **Florian (B.)**, for sharing the office at LMU. Thank you, **Sarah (A.H.)** and **Kai**, for exploring Kairo together. Thank you, **Karola**, for obtaining a Kilt for me in Glasgow and for sharing drinks in Boston. Thank you, **Rainer**, for the technical support on my defense and the makeshift lectern. Thank you, **Franziska**, **Christa**, and **Ulrike**, for supporting all my paperwork at LMU. Thanks to **Bastian**, **Beat**, **Christina**, **Christian**, **Fiona**, **Jakob**, **Linda**, **Thomas (K. & W.)**, **Tobias**, and so many more who I met at LMU, during research visits and as participant or SV during conferences. It was always great to meet you, exchange ideas, and collaborate.

Thank you to the many students I supervised and worked with over the years. Your efforts were a great support and contribution to my research. Thank you, **Ismael**, for your invaluable hardware expertise and efforts in making the magnetic keyboard a reality. Thank you, **Rupert**, for the countless hours you poured into Imitation Game. Many thanks to **Mathias (K.)** for exploring intentional typing control with me, to **Alice** for your work on context-sensitive authentication, and to **Oliver (D.)**, **Andrea (N.)** and **An** for your work on understanding user preferences with regards to biometrics. Thanks to **Assem**, **Celina**, **Daniel (R.)**, **Hanna**, **Jan**, **Jens**, **Matthias (G.)**, **Melina**, **Pia (K.)**, **Rana**, **Simon**, **Vanessa**, and all other students I had the pleasure of working with.

Last but not least, I would like to thank my friends and family. You know who you are. Thank you for checking in on me, for your unconditional and unwavering support, and for being my anchor to keep me motivated throughout this long process. A special thanks to **Pia (L.)** for proofreading, testing my prototypes, giving feedback on my studies and papers, and providing delicious food at my defense.

Thanks for everything.

Publications & Co-Authorship

This thesis is based on my research at LMU Munich, the University of Applied Sciences Munich, and the Research Institute CODE at the University of the Bundeswehr Munich. However, this thesis was improved and made possible through the support and constant feedback of many different people, including my supervisors, colleagues, and students who did their thesis and practical courses with me. I will use the scientific 'we' throughout this thesis to reflect this.

Parts II, III, and IV of this thesis are based on co-authored publications at international, peer-reviewed conferences. Some of them are currently unpublished and/or under submission and will be marked as such. Many projects were supported through practical work conducted by students for their thesis or practical courses that I supervised. Parts I and V were written exclusively for this thesis. Chapter 2 contains some elements of the contributing publications. Collaborations on all included publications are listed in detail below.

> **Lukas Mecke**, Alia Saad, Sarah Prange, Uwe Gruenefeld, Stefan Schneegass, and Florian Alt. 2024. *Do They Understand What They Are Using? — Assessing Perception and Usage of Biometrics*. **arXiv preprint** arXiv:2410.12661 [186]

This project was my idea. I designed the questionnaire with feedback from Florian Alt and conducted both rounds of the survey. The qualitative analysis was done together with Alia Saad, Sarah Prange, and Uwe Gruenefeld. I did the qualitative analysis and led the writing of the paper. Alia and Uwe contributed to the related work and results section. I iterated each part of the paper based on feedback from all co-authors.

> **Lukas Mecke**, Ken Pfeuffer, Sarah Prange, and Florian Alt. 2018. *Open sesame! user perception of physical, biometric, and behavioural authentication concepts to open doors*. In Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia (MUM '18) [183]

This paper is based on a project conducted by Oliver Duerr, Andrea Ngao, and An Ngo Tien during a practical course for their master studies at LMU Munich and was supervised by Ken Pfeuffer, Sarah Prange, and myself. The students designed and conducted the study under our continuous feedback. They provided a first analysis which I iterated on. The writing for the first version was shared between Ken, Sarah, and myself. I led the writing of the final version and presented the paper at MUM 2018.

> **Lukas Mecke**, Sarah Prange, Daniel Buschek, and Florian Alt. 2018. *A Design Space for Security Indicators for Behavioural Biometrics on Mobile Touchscreen Devices*. In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18) [184]

The idea for this project was initially suggested by my supervisor Florian Alt. Sarah Prange and I designed the focus group and conducted and evaluated it together. Both Florian and Daniel Buschek gave feedback on those steps. I led the writing which was shared between Sarah, Daniel Buschek, and myself. Sarah and I presented the poster together at CHI 2018.

> **Lukas Mecke**, Daniel Buschek, Uwe Gruenefeld, and Florian Alt. 2024. *Exploring the Lands Between: A Method for Finding Differences between AI-Decisions and Human Ratings through Generated Samples.* **arXiv preprint** arXiv:2409.12801 [180]

I had the idea for this project, generated the dataset and designed the study with feedback from Daniel Buschek, Uwe Gruenefeld, and Florian Alt. I conducted the study and analyzed the data myself with feedback from Daniel and Uwe. I wrote the paper and iterated it based on feedback from all co-authors.

> Sarah Prange, **Lukas Mecke**, Alice Nguyen, Mohamed Khamis, and Florian Alt. 2020. *Don't Use Fingerprint, it's Raining! How People Use and Perceive Context-Aware Selection of Mobile Authentication.* In Proceedings of the International Conference on Advanced Visual Interfaces (AVI '20) [208]

The concept and idea for this paper were brought up by Alice Nguyen in the context of her master's thesis at LMU Munich. The thesis was supervised by Sarah Prange and myself. Alice conducted the focus group, developed the application and conducted the final study under close supervision by Sarah, Mohamed Khamis and myself. Alice provided a first analysis of the data which I iterated on. Sarah and I shared the writing of the paper with feedback and iterations by all other co-authors. We reflected this by indicating equal contribution to the paper. Florian presented the paper at AVI 2020.

> **Lukas Mecke**, Sarah Delgado Rodriguez, Daniel Buschek, Sarah Prange, and Florian Alt. 2019. *Communicating Device Confidence Level and Upcoming Re-Authentications in Continuous Authentication Systems on Mobile Devices.* In Fifteenth Symposium on Usable Privacy and Security (SOUPS '19) [182]

The idea for this project was developed together with Sarah Delgado Rodriquez for her bachelor's thesis at LMU Munich. I supervised the thesis supported by Sarah Prange and Daniel Buschek. Sarah (D.R.) developed the concept and conducted the focus group and final study under our close supervision. She provided a first analysis of the results which we collaboratively iterated and refined. I led the writing of the paper and iterated it based on feedback from all co-authors. Sarah (D.R.) presented the work at SOUPS 2019.

> **Lukas Mecke**, Daniel Buschek, Mathias Kiermeier, Sarah Prange, and Florian Alt. 2019. *Exploring intentional behaviour modifications for password typing on mobile touchscreen devices.* In Fifteenth Symposium on Usable Privacy and Security (SOUPS '19) [181]

The idea for this project was developed together with Mathias Kiermeier for his bachelor's thesis at LMU Munich. I supervised the thesis supported by Sarah Prange and Daniel Buschek. Mathias developed the Android app and conducted the study under our close supervision. The study design was developed in many iterations by all authors. Mathias provided a first analysis of the data which was redone and extended by me. Daniel conducted the statistical tests and implemented the machine learning parts. I led the writing of the paper supported by iterations and feedback from all co-authors and presented it at SOUPS 2019.

> **Lukas Mecke**, Rupert Oxenius, Sarah Delgado Rodriguez, Daniel Busckek, and Florian Alt. 2024. *Imitation Game: A Mobile Game to Support Players in Learning to Control Features of their Typing Behavior*. **In preparation for publication.**

I had the idea to extend our previous study as a game which was executed by Rupert Oxenius as the project for his bachelor's thesis at LMU Munich. The thesis was supervised by Sarah Delgado Rodriguez and myself with constant feedback from Florian Alt and Daniel Buschek. Rupert developed the first version of Imitation Game and tested it in a small study that is not part of the paper. After his thesis, he continued development on the project as a practical course supervised by Sarah and me. Rupert implemented the game and created the story. I designed all visuals for the game. All authors planned the study design together, which was conducted by Rupert and myself. The data analysis was shared between Sarah and myself. I led the writing of the paper and iterated it based on feedback from all co-authors.

> **Lukas Mecke**, Ismael Prieto Romero, Sarah Delgado Rodriguez, and Florian Alt. 2023. *Exploring the Use of Electromagnets to Influence Key Targeting on Physical Keyboards.* In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23) [185]

The idea for externally influencing typing originated from a discussion between my supervisor Florian Alt, Sarah Delgado Rodriquez, and myself. Ismael Prieto Romero came up with the concrete idea of using electromagnets for this project. He implemented and evaluated the concept in the course of his bachelor's thesis under my supervision and with feedback from Sarah. I iterated the initial analysis provided by Ismael and led the writing of the paper based on a draft provided by him. All co-authors contributed iterations and feedback. The first authorship on the poster was shared between Ismael and me to reflect his strong technical contribution and my contribution in steering the project and writing the paper. I presented the poster together with Sarah at CHI 2023.

TABLE OF CONTENTS

I

Introduction & Background

PART I: INTRODUCTION & BACKGROUND

In this part, we introduce the topic of this thesis, our approach to answering our research questions as well as relevant background information. As such it serves as both motivation and structural overview for the rest of this thesis.

- ❖ **Chapter 1** introduces the motivation for our work and the research questions we derive from it. It outlines our contributions and research approach and gives an overview of the rest of the thesis.
- ❖ **Chapter 2** introduces relevant background information on the topics of authentication and biometrics and highlights the limitations of previous work that motivate our investigations.

1
Introduction

Biometrics leverage unique characteristics in human physiology or behavior for authentication and are increasingly used to protect (mobile) devices. However, they are inherently based on machine learning models and are thus hard to predict and control, in particular for end-users. In this thesis, we propose the use of user-centered biometric interfaces as an approach to foster user literacy about biometrics and empower them to take control over their authentication mechanism.

In this chapter, we outline the motivation for our work and introduce the research questions that guide the rest of this thesis. We summarize our contributions and introduce our research approach to answer these questions. We conclude the chapter with an overview of the thesis.

1.1 Motivation

Many of the devices we interact with on a daily basis store highly sensitive and personal data like images or medical and business information [151]. Beyond the stored data, they also provide access to many essential services like financial applications, messages, or the user's social media presence. It is thus of great importance to protect access to those devices and authentication has become a necessary part of interaction.

While passwords are still widely used to secure access to accounts, they have started to reach their limits. As the amount of required authentication increases, users need to remember and enter more and more passwords, straining users' memory and requiring time.

As an alternative, the use of biometric methods has emerged and gained popularity over the past years with an estimated 80% of mobile devices capable of using them in 2020[1]. Biometrics leverage unique characteristics in human physiology and behavior for authenticating or identifying individuals [190, 285]. Most notably, fingerprint [141, 142] and face recognition [275] are available on most modern smartphones, but also other features like the user's iris, gait, or the way of typing on a keyboard [12, 43, 229] have been proposed. Compared to other authentication approaches, biometrics do not need to be remembered and cannot be lost or stolen [139]. They do not require additional user input for the sake of authentication beyond their presence or normal interaction with the device and can thus facilitate authentication in the background.

However, this approach also takes away user control over authentication. When a user's behavior is captured constantly there is no clear point for them to express their intent to be authenticated. Note, how this is also relevant for users, who do not actively use biometrics, as recognition can be done without the user's knowledge or consent, e.g. through entering text on a website or by being recognized in public space (e.g. using face or gait recognition). Previous work has shown that biometric models can be prone to external factors [29]. The machine learning models underlying biometric recognition act as a black box to the users, making their decisions hard to understand and leading to biases [48, 82, 298]. Their performance can vary greatly between users [294], meaning that performance metrics commonly used in machine learning to describe the quality of a model (e.g. accuracy or equal error rate) may not be relevant and applicable to the individual. As a consequence, users can lose trust and access, or in the worst case be harmed [298].

While experts and designers may be aware of those points, this knowledge does not necessarily map to the users of a security system [4, 15] and incorrect mental models can lead to insecure behavior [284]. Thus, focusing on the user is of great importance [4, 230]. Thinking back to the relevance of the data and applications protected with biometrics today, this illustrates the need for action toward understanding users' current knowledge of biometrics and the individual factors that constitute their performance. The next step has to be communicating this knowledge to form trust and avoid potential errors. Finally, users need both the

[1] `https://www.statista.com/topics/4989/biometric-technologies/`, last accessed October 16, 2024

motivation and opportunity to act on their knowledge to control their biometric systems and make an informed choice about which system to use or whether to use a biometric system in the first place [277].

In this thesis, we investigate which information users are currently missing about their biometric systems to both enhance existing interfaces with biometric systems and propose new ones where none exist so far. We explore how such interfaces could look like and assess how they can improve interaction with biometric systems. The results of this thesis can support researchers and practitioners to design and investigate user-centered biometric interfaces and contribute to a secure and informed use of biometrics.

1.2 Research Approach

In this section, we outline the research questions we derived to guide our research and the types of contributions we made to those research questions. We also give an overview of the methods used in this thesis and the rationale behind their choice.

1.2.1 Research Questions

Based on the illustrated challenges, we derive three steps to enable more secure and informed use of biometric methods: As a first step, we need to *understand user needs* and what design opportunities exist to address them. As a second goal, we propose to support users in gaining a better understanding of biometric methods and, thus, improve their *biometric literacy*. However, we argue, that knowledge alone is not enough and thus propose to offer users options to control biometric models and, thus, gain *agency* over their authentication.

While the first step provides a foundation for our work, we propose to address the other goals by designing *biometric interfaces*, that is, by creating or augmenting points where users and a biometric model come in contact with each other (e.g. during enrollment or authentication). We argue for following a *user-centered* approach when designing those interfaces, taking into account user needs and personalizing information to the user and their context. We summarize our goals as research questions that guide the work in this thesis below:

> RQ1 What are **user needs** and how can they be addressed through the design of biometric interfaces?
>
> RQ2 How can users be supported to acquire **biometric literacy** through biometric interfaces?
>
> RQ3 How can biometric interfaces be leveraged to extend **user agency?**

1.2.2 Research Contributions

Wobbrock and Kientz [289] outlined seven types of contributions to Human-Computer Interaction (HCI) research. Here we give an overview of the four types that our work makes contributions to based on this categorization.

Empirical Research Contributions

In this thesis, we aim to understand user needs with regard to biometric interfaces and design and evaluate approaches to address them. As such, empirical research is at the core of this work and all projects that are part of it make empirical contributions in one way or another. We provide empirical data on the use and perception of biometrics (Chapter 3) as well as participants' preferences for different methods (Chapter 4). Based on a focus group we derive design opportunities for biometric interfaces (Chapter 5). We uncover differences in the rating of similarity between humans and a face recognition model (Chapter 6) and gain empirical insights into users' willingness to change their authentication method based on context (Chapter 7). We collect data on users' interaction with indicators that illustrate the state of a continuous authentication system (Chapter 8) and show that they are able to intentionally modify their typing behavior (Chapter 9). We uncover differences in this ability between a lab setup and use in the wild (Chapter 10) and show, that electromagnets can be used to influence typing as well (Chapter 11).

Artifact Contributions

Where possible we designed prototypes or applications to give participants an impression of how interaction with our solutions would look and feel. As many of our evaluations involved participants interacting with a biometric system in the wild those artifacts mostly took the form of apps. We developed an application that leverages context information to suggest appropriate authentication mechanisms on mobile phones (Chapter 7). With another application, we explored user interaction with indicators that gave insights into a mocked continuous authentication system and warned users of upcoming re-authentications in the wild (Chapter 8). We developed a game that supports players in learning to modify their typing behavior (Chapter 10) and implemented a physical keyboard that does the same by exerting force on a magnet attached to the user's finger (Chapter 11). Overall, the use of artifacts allowed us to gain both more realistic and long-term insights into users' interaction with our designs of biometric interfaces.

Methodological Contributions

While methodological contributions were not a focus of this work we developed some methods to enable our other investigations. Based on a focus group we contribute a design space that provides guidance on factors to consider when building biometric interfaces (Chapter 5). To gain a better understanding of face recognition models we proposed a method for finding and generating challenging samples that can be used to identify potential mismatches in perceived similarity between human raters and a decision-making model (Chapter 6). Finally, we proposed and demonstrated an approach to convert a security study to a game with the goal of having participants use it in the wild (Chapter 10).

Dataset Contributions

Whenever we collected larger datasets in the course of our empirical evaluations we also contributed them for further analysis by other research. We compared ratings of similarity for pairs of face images between human raters and a face recognition model. We provide all comparisons and the respective image pairs as a dataset (Chapter 6). Based on our investigation of users' ability to modify their own typing behavior we contribute a dataset of expected behavioral patterns and participants' actual typing (Chapter 9). We do the same for our game implementation that compares the same task between lab and remote participants (Chapter 10).

1.2.3 Empirical Research Methods

Here we give an overview of the *study paradigms* and *data collection methods* we used throughout this thesis and our rationale behind their use. More details can be found in the respective chapters.

Lab Studies

We used lab studies as a means to gain control over the environment, vary specific variables, and make repeated measurements (Chapter 9). A second motivation was to enable user interaction with artifacts that were not portable and could not be used without the oversight of a researcher (Chapters 4 and 11).

Field Studies

We employed field studies as a means to gain more realistic insights into users' interaction with our prototypes over a longer period of time. We used this for our evaluation of a mechanism to suggest switches to an appropriate authentication mechanism based on context and an interface to provide insights into a (mocked) continuous authentication model (Chapters 7 and 8). As a special case, we developed a game with the goal of being playable outside a study context to support users in controlling their typing behavior (Chapter 10).

Surveys

We employed surveys for their ability to reach a large number of participants for tasks that did not require interaction with a prototype. We employed this method to assess the use and perception of biometric methods (Chapter 3) and to collect human ratings on the similarity of pairs of face images for our comparison against a face recognition model (Chapter 6).

(Expert) Focus Groups

We used focus groups to validate and iterate designs for prototypes for our field studies (Chapters 7 and 8). This was of particular importance for this type of study, as we had no direct oversight over participants using our developed artifact and thus needed a diverse set of feedback before deploying them. As a special case, we used an expert focus group to find design considerations for biometric interfaces that we used to derive a design space (Chapter 5).

Data Collection: Questionnaires, Interviews, and Experience Sampling

Most of our empirical studies were accompanied by questionnaires to gain additional insights and understand the effects we observed in the quantitative results. In our lab studies, we also employed interviews after the main tasks as a more interactive way for participants to share their experiences. In the field studies, we employed experience sampling as a method to capture feedback over a longer time frame.

1.3 Ethical Considerations

In Germany, where this research was conducted, there is no formal requirement for approval from an Institutional Review Board (IRB) for the type of research we conducted in this thesis. However, we always took great care to comply with all guidelines given by our institution and national data protection regulations. In particular, consent was always gathered before collecting any data. Email addresses and other types of contact information were stored separately from study data and only used for communication and compensation of the participants.

1.4 Research Context

This thesis is based on research that I conducted at LMU Munich, the University of Applied Sciences Munich, and the Research Institute CODE at the University of the Bundeswehr Munich between December 2017 and September 2023. During this timeframe, I worked on user-centered biometric interfaces and their evaluation in the wild. My work contributed to the project "*Biometrics++ — Leveraging Behavioral Biometrics Beyond Security to Both Secure and Personalize Interactive Ubiquitous Computing Devices*" funded by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B) as well as to the project "*Designing and Evaluating Scalable Behavioral Biometrics Systems for Pervasive Computing Environments*" funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).

Part I: Introduction & Background		
Part II: Understanding User Needs and Design Opportunities		
Chapter 3 *Understanding Use and Perception of Biometrics*	Chapter 4 *Exploring User Preferences for Biometrics*	Chapter 5 *Design Opportunities for Biometric Interfaces*
Part III: Biometric Interfaces to Support User Literacy		
Chapter 6 *Exploring (personalized) Performance of Face Recognition using Generated Samples*	Chapter 7 *Leveraging Context Cues to Inform Authentication Choice*	Chapter 8 *Communicating the State of Continuous Authentication Systems*
Part IV: Biometric Interfaces to Support User Agency		
Chapter 9 *Exploring Intentional Keystroke Control*	Chapter 10 *Extending Intentional Keystroke Control to the Wild with Imitation Game*	Chapter 11 *Supporting Key Targeting using Electromagnets*
Part V: Discussion & Conclusion		

Table 1.1: Overview of the parts and chapters comprising this thesis.

1.5 Thesis Outline

This thesis consists of 13 chapters organized in five parts of which you are currently reading the first. Parts I and V are framing our work on answering the research questions in Parts II-IV with an introduction and a discussion (see Table 1.1). Here we give a more detailed overview of what to expect in this thesis.

Part I

In this part, we already outlined the motivation for this work and introduced the overarching research questions guiding our work. We also introduced the contributions our work makes and the methods we use. In Chapter 2 we provide further background for this work by introducing relevant concepts like authentication in general and biometrics in particular as well as relevant work from related fields. We conclude this chapter with an overview of the challenges that motivate our work.

Part II

We use this part to answer our first research question and lay the foundation for our further exploration of biometric interfaces by understanding user needs and design opportunities. In Chapters 3 and 4 we explore users' use and preferences with regard to biometrics. We take different approaches for the two chapters, one time focusing on the biometrics participants use and their general perception of biometrics and in the other case comparing different mechanisms to a non-biometric alternative to uncover preferences and advantages of both options. Our investigations reveal some misconceptions and a demand for further information as well as a need for retaining control over authentication. In Chapter 5 we add a design perspective to those findings and explore how interfaces for biometric systems could look like to address those points. Together with related work, this part informed our choice and design of solutions proposed in the next parts and our focus on supporting user *literacy* and *agency* over biometric systems.

Part III

The main focus of this part is our exploration of interfaces to support users in gaining a better understanding of biometric methods and thus improve their *biometric literacy* as captured in our second research question. We followed a breadth-first approach, exploring interfaces for both different biometrics and different interaction scenarios. Our aim behind this approach was to show the possibility and potential of enhancing existing interfaces to support user interaction with biometric systems. In Chapter 6 we propose a method to gain insights into the performance of a biometric model. We apply the approach to a face recognition model and show, that it can not only be used to evaluate a model as a whole but can yield insights into performance for single users. It can thus be used to improve the *enrollment* process of a biometric system by providing users with a better estimation on how the model will work for them specifically. In Chapter 7 we propose to show users information about context factors that could impact fingerprint *authentication*, giving them both more information about the biometric model and nudging them to switch to a more appropriate method. We conclude this part with Chapter 8 where we explore how providing continuous and event-based information about a model's state can help users to better anticipate and cope with *re-authentications*. This process can become necessary when the model is not sufficiently certain that a legitimate user is interacting with it. We also introduce a mechanism for users to take more control over the model by voluntarily re-authenticating when they anticipate such a situation to avoid being interrupted in inconvenient situations.

Part IV

In this part, we explore our second design goal for biometric interfaces: supporting users in gaining *agency* and control over the biometric systems they use (see RQ 3). Part III already introduced some concepts in this direction but in this part, they are at the focus of our work.

In contrast to the previous part we here follow a depth-first approach and explore the single use-case of authentication by typing behavior for our investigation. We took this approach to understand how to design interfaces for biometrics that are designed to run in the background and thus do not normally offer user interaction (except for re-authentication requests) and to compare different options for improving the same interaction. This was not possible in the first part, as all tested approaches shared a common goal but were fundamentally different. As a first step, we explore in Chapter 9 if users can gain control over being recognized by a biometric system using typing behavior. To this end, we developed a visualization to communicate typing features and used it to show that users were able to adjust their typing accordingly. With our goal of supporting users in gaining this type of agency, we built a game based on our study setup in Chapter 10 that was designed to support users in learning to modify their typing on their own and in a playful manner. With Chapter 11 we explore the use of electromagnets to free users from having to actively control their behavior to achieve typing modifications. To summarize, we use this part to show that users can take agency over authentication through typing behavior, how this can be achieved in the wild, and how an approach could look like that requires less user involvement to achieve this goal.

Part V

We conclude this thesis with a discussion of the findings and implications of our work. In Chapter 12 we offer our insights into design considerations for implementing biometric interfaces and reflect on the methods used throughout this thesis. In Chapter 13 we summarize the contributions we made to our research questions and outline potential directions for future research before concluding the thesis with some final remarks.

2
Background & Related Work

In this thesis, we propose the design and implementation of user-centered biometric interfaces. This places our research at the intersection of IT-Security and Human-Computer Interaction (HCI) research, more commonly referred to as usable security [107].

In this chapter, we give an introduction to usable security and authentication before giving a deeper overview of what biometrics are, how they work, in which forms they exist, and what their challenges are. *Related work on biometric interfaces is yet very sparse, so we introduce insights on similar research from other fields. More detailed related work will be presented in the chapters on a per-case basis.*

We conclude this chapter with an overview of open challenges that motivate our work.

2.1 Authentication & Usable Security

There exist many different definitions for authentication, but in this work, we follow the interpretation by Saltzer and Schroeder [227]:

> **Authenticate**: To verify the identity of a person (or other agent external to the protection system) making a request.

In the context of this thesis, this request generally refers to gaining access to a device, account, or area that is protected by an authentication mechanism.

The type of authentication can be further categorized by the type of evidence a person presents to make this request. Based on O'Gorman [199], authentication systems can be broadly divided into three types: Knowledge-Based, Object-Based, and ID-Based. *Knowledge-Based* systems utilize a secret, that the user knows and encompass approaches like passwords, PINS, and patterns. *Object-Based* systems identify a person based on the possession of an object, commonly a token, key, or their smartphone. *ID-Based* authentication uses things that are unique to a person, including their ID card and what is commonly referred to as biometrics.

However, when users interact with authentication systems, they do not do so to authenticate but because they want to access a protected good (e.g. unlocking a phone to read a new message). As such, authentication is generally a secondary task [230, 231] and should also be viewed as such. This motivates the field of *usable security* [107], arguing for designing security systems in a way that users can easily and with low effort use them. This requires a user-centered approach to design [310] and means that interfaces should be designed to support users instead of seeing users as a weakness of security systems [4, 130, 230].

In this thesis, we adapt this philosophy to the design of biometric interfaces informed by user needs and dedicated to supporting them in their interaction with biometric systems. We give a more detailed overview of biometrics next.

2.2 Biometrics

Biometrics are a type of authentication system that uses unique characteristics in human physiology or behavior to identify individuals. More formally, they are defined in ISO 2382-37 [136] as follows:

> **Biometrics**: automated recognition of individuals based on their biological and behavioral characteristics.

Based on Jain et al. [144], any human physiological or behavioral characteristic can be used for this kind of authentication, as long as it fulfills four criteria:

1. *Universality*: each person should have the characteristic
2. *Distinctiveness*: any two persons should be sufficiently different in terms of the characteristic
3. *Permanence*: the characteristic should be sufficiently invariant (with respect to the matching criterion) over a period of time
4. *Collectability*: the characteristic can be measured quantitatively

2.2.1 Types of Biometrics

Biometrics are generally divided in physiological and behavioral methods.

Physiological biometrics leverage unique body features like faces [261, 269, 307], fingerprints [174, 175, 297], hand or palm prints [166, 308] or the human iris [30, 73]. As such, they usually require the user to actively present the respective feature for authentication.

Behavioral biometrics on the other hand are based on learned movements such as typing [42, 152, 214, 262] or touch [85, 263] behavior, or gait [104, 242, 251, 282]. Capturing those features can only be done over time and thus needs users to execute the respective behavior for authentication to be possible. Behavioral biometrics tend to be less stable and are overall less adopted [57], but open up the opportunity for implicit authentication (see Section 2.2.4).

O'Gorman [199] observed, that the distinction in behavioral and physiological methods can sometimes fall short. For example, voice exhibits both physiological features based on the vocal tract and a behavioral element in the way it is used. Actually, all types of behavior have some physiological component, as they are influenced by the human body. As such, O'Gorman proposed a slightly more nuanced distinction by the type of biometric signal presented to the system and speak of *stable* and *alterable* biometric signals. However, this distinction is not commonly used so we use the distinction between physiological and behavioral biometrics for the remainder of this thesis.

We refrain from giving further details on any specific type of biometric method here, but instead introduce them in the respective chapters where they become relevant.

2.2.2 General Functionality

A biometric system generally consists of four components [144] and follows a similar process: 1) A *sensor* captures the respective features presented by the user. The captured data is then 2) preprocessed and salient features need to be *extracted*. The so-created biometric templates are input to 3) a *matching algorithm*, which compares them to other templates stored in 4) a *database*. The matcher then returns a decision on whether or not the user should be authenticated.

Before first using the system, users have to *enroll*. This means that they have to follow the same process with the difference, that no match will be made but instead the newly generated biometric template is stored in the database. Only then can the user be recognized.

2.2.3 Authentication Modes

Authentication in the context of biometrics can be further distinguished into two main modes: verification and identification. Here we follow the definition by Jain et al. [144] who describe the terms as follows:

> *Verification*: In the verification mode, the system validates a person's identity by comparing the captured biometric data with her own biometric template(s) stored in the system database.
>
> *Identification*: In the identification mode, the system recognizes an individual by searching the templates of all the users in the database for a match.

In other words, the two modes differ in the user claiming an identity or not. If no identity is claimed, a match has to be found between all known templates, otherwise, the user has only to be verified against the template associated with the claimed identity. Throughout this thesis, we use the term authentication as an umbrella term for those two options when the exact form is not relevant in the current context.

2.2.4 Explicit and Implicit Authentication

Most traditional authentication mechanisms like passwords or tokens require an explicit user action, e.g. entering the secret or presenting the identifier. Similarly, many physiological biometric features enable *explicit authentication*, e.g. fingerprints or iris recognition. In contrast to explicit authentication, the term *implicit authentication*[1] describes the process by which a user is authenticated without requiring explicit interaction. In implicit authentication systems, the initial explicit authentication step to gain access to a device is replaced or complemented by a continuous evaluation of the user's identity that is reflected in a score (or device confidence level (DCL)). Similar to a fallback in explicit authentication systems (e.g. the use of a PIN when the fingerprint scan fails), an explicit so-called *re-authentication* is required in case the model can not verify the user's identity.

[1] also called transparent or continuous authentication (e.g., [63])

Some methods suggested for implicit authentication rely on the user's context [24, 125, 188, 218], but the method is most prevalently used with behavioral features. Examples include mechanisms that authenticate users based on gait recognition [76], continuous eye-tracking [192], or the users' tap or app-execution behavior [39, 71, 226, 247].

2.2.5 Error Types and Performance Measures

When making an authentication decision, biometric models can make two types of errors. They can wrongly authenticate a non-legitimate user (*false positive*) or wrongly reject the legitimate user (*false negative*). To describe a model, those errors are collected over all samples the system was tested on to yield the *False Acceptance Rate (FAR)* and *False Rejection Rate (FRR)* for false positive and false negative decisions respectively. A high FAR is a security problem because it means more non-legitimate users could wrongly bypass the authentication. A high FRR on the other hand poses a usability problem as the legitimate user may have to spend multiple tries to gain access or has to re-authenticate often.

Biometric models are generally machine learning models and are driven by a decision function on which a threshold is applied. This threshold can be varied to be more or less strict and thus favor one or the other type of error. Thus, designers of biometric models have to make a trade-off between usability and security of their model.

To get a comparable measure for the performance of a model, an established measure is the *Equal Error Rate (EER)* which is described as the point where FRR and FAR are equal. Other popular performance metrics are Accuracy, Precision, Recall, F1 score, and ROC curves [206]. We will not go into detail about all of them but remark, that they all describe the ability of the model to make correct decisions on a global level, i.e. based on the presented set of training data.

2.3 Commercial Biometric Systems

Commercial biometric systems are widely available these days. Here we introduce some of the available sensors and services offering biometric authentication. Note, that this section is intended to give an overview and the commercial solutions listed here are by no means an exhaustive list of available products.

2.3.1 Inbuilt Biometric Systems

The main source of distribution and use for biometrics are built-in solutions for modern mobile phones and laptops. Both Android and Apple devices offer fingerprint readers and face recognition in the form of Apple Face-ID[2] and Androids Face Unlock[3] feature.

They can also include capabilities for behavioral biometrics like the use of gait recognition in the Smart Lock feature[4] offered on Android Devices. Windows offers support for both face and fingerprint recognition in the form of Windows Hello[5].

Those types of biometrics are likely the ones users interact with the most, as they come shipped with devices they use in their daily lives and do not require additional setup.

2.3.2 External Biometric Sensors and Security Solutions

To use biometrics in contexts where they are not shipped with a system, companies offer both sensors and full-fledged security solutions. Exapmples are USB devices like the Kensington VeriMark[6] that can be plugged into a laptop to enable fingerprint authentication. Fujitsu offers a similar solution with the PalmSecure[7] that can be attached to a laptop to enable biometric recognition based on a palm vein scan. The company IrisID offers both standalone scanners and solutions for iris identification[8].

In addition to extending the capabilities of (mobile) devices, biometric sensors are often offered as systems for access control. This includes Solutions such as the aforementioned IrisID systems or the Ekey-Uno[9] that extends smart lock systems with the functionality for unlocking with fingerprints.

Finally, many systems offer multiple authentication mechanisms and biometrics in a package. An example is the eufy Smart Lock E130[10] offering capabilities for unlocking with a

[2] Apple Face-ID: `https://support.apple.com/en-us/102381`, last accessed October 16, 2024

[3] Android Face Unlock: `https://support.google.com/pixelphone/answer/9517039`, last accessed October 16, 2024

[4] Google Smart Lock: `https://support.google.com/android/answer/9075927`, last accessed October 16, 2024

[5] Windows Hello: `https://learn.microsoft.com/en-us/windows-hardware/design/device-experiences/windows-hello`, last accessed October 16, 2024

[6] Kensington VeriMark: `https://www.kensington.com/de-de/software/verimark-setup/`, last accessed October 16, 2024

[7] Fujitsu PalmSecure: `https://www.fujitsu.com/de/services/security/offerings/biometrics/`, last accessed October 16, 2024

[8] IrisID: `https://www.irisid.com/`, last accessed October 16, 2024

[9] ekey-UNO: `https://www.ekey-uno.net/`, last accessed October 16, 2024

[10] eufy Smart Lock E130: `https://us.eufy.com/products/t8510`, last accessed October 16, 2024

smartphone, key, fingerprint or a knowledge-based secret. The Tenon Smart Lock[11] offers both face and fingerprint recognition to unlock doors.

Overall, commercial solutions in this category offer users the option to post-hoc enable biometric recognition for their devices or integrate them as a part of their smart home solutions.

2.3.3 Biometrics as a Service

Finally, some types of biometrics do not require additional hardware and are thus commercially offered as a service. Examples include the Microsoft Azure AI Face service[12] offering business customers models for face detection and face recognition. Similarly, Clearview.ai[13] is offering face recognition for law enforcement.

Some solutions also leverage behavioral biometric features like TypingDNA[14] which provide typist verification based on keystroke dynamics. BehavioSec[15] combines different behavioral traits like keystrokes, device movement, mouse movement, and touch features for user verification and fraud prevention.

In general, there is a large pool of services, that offer some kind of fraud detection and worker monitoring to ensure that services are not accessed by unauthorized third parties. Here, combinations of biometric features and other usage data is leveraged, but it often remains unclear, how exactly they are used. Examples include NetHone[16] using behavioral biometrics as part of their account protection solution and Castle[17] leveraging diverse user interaction features to prevent account abuse and automatically triggering responses. BioCatch[18] makes use of behavioral features to prevent online fraud and PuriLock[19] leverages keyboard and mouse interactions to verify current users and prevent compromised sessions.

Overall, those solutions are rather targeted at businesses and larger companies and often do not give clear insights into what constitutes the models they use. The wide variety of such solutions also underlines the importance that the use of biometrics has for businesses these days to protect their networks, detect intrusion, and prevent fraud.

[11] Tenon Smart Lock: https://www.aptenontech.com/products/automatic-electronic-doorbell-face-recognition-smart-lock/, last accessed October 16, 2024

[12] Azure AI Face service: https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/overview-identity, last accessed October 16, 2024

[13] Clearview.ai: https://www.clearview.ai/, last accessed October 16, 2024

[14] TypingDNA: https://www.typingdna.com/, last accessed October 16, 2024

[15] LexisNext BehaioSec: https://risk.lexisnexis.com/products/behaviosec, last accessed October 16, 2024

[16] NetHone: https://nethone.com/solutions/user-lifecycle-fraud-prevention, last accessed October 16, 2024

[17] Castle: https://castle.io/, last accessed October 16, 2024

[18] BioCatch: https://www.biocatch.com/, last accessed October 16, 2024

[19] PuriLock: https://plurilock.com/products/defend/, last accessed October 16, 2024

2.4 Challenges of Biometric Authentication

While biometric models can offer a fast and convenient way of authentication, they also come with a range of possible drawbacks. Jain et al. [140] propose a list of five concerns for biometric authentication that systems should address: Performance, Bias and Fairness, Security, Explainability and Interpretability, and Privacy. Here we give an overview of those challenges.

2.4.1 Performance

While the accuracy of modern-day biometric recognition systems often exceeds human-level performance [140], they very much rely on the quality of input they receive. As such, noise to this input can greatly deteriorate their performance. Chugh et al. [56] found fingerprint performance impacted by the quality of samples with factors like motion blur, or changes to the fingerprint due to humidity, all of which decreased recognition performance. Similarly, external factors like light conditions or pose variations can have an effect on face recognition [116]. Bhagavatula et al. [29] found those effects also mirrored in the usability of those biometrics in a lab study where participants would use them under varying conditions. Sieger et al. [250] found that voice or speaker recognition is not usable in crowded places.

In addition to external factors, the users themselves seem to provoke differing biometric system performance as well. Yager and Dunstone [294] found that system performance can vary greatly depending on the user and Cabrera et al. [48] showed such effects for specific user groups.

Finally, scalability can become an issue. In particular for recognition models that rely on a one-to-many comparison, more users make the task more difficult. As an example, Pfeuffer et al. [203] found the performance of their system to identify users in VR to deteriorate with increasing numbers of users.

Those challenges both highlight the need for further improvements to models' performance in border cases as well as for biometric interfaces to communicate those cases to the end-users of biometric methods.

2.4.2 Bias and Fairness

Biometric models are often prone to (demographic) biases, that is, they perform better for certain groups and worse for others. This effect is mainly researched in the field of face recognition, where multiple studies found models to be biased [48, 116, 164, 264, 298]. While biases vary strongly between different models, related work uncovered both biases in face recognition systems towards people of color [48, 298] as well as women [164]. Another known bias is fingerprint models performing worse for infants [89, 138].

Possible reasons for such biases can be biased datasets (i.e., certain groups are underrepresented in the data) and algorithmic optimizations, that lead to the model focusing on overall performance rather than optimizing for corner-cases [140].

As this problem is as of now still open and unsolved, biometric interfaces can help mitigate its effects and communicate the existence of such biases to the users before they decide to use a biometric system. In Chapter 6 we introduce an introspective method that may be used for uncovering such effects on a personal level to allow for better anticipating model performance.

2.4.3 Security

Jain et al. [140] categorize potential threats to biometric models by the part of the recognition process they attack. They speak of *presentation attacks* targeting the sensing module, *adversarial attacks* targeting the feature extraction and template theft and subsequent *template reconstruction attacks* aimed at the template database.

Presentation attacks can have both the goal of evading recognition (e.g. through damaging a fingerprint [301]) or gaining access to a protected device (e.g. using an artificial finger [13]). The equivalent to such presentation attacks for behavioral biometrics is called a *mimicry attack*. This involves a human (or artificial input [193, 302]) trying to emulate the legitimate users' input. This was successfully done for e.g. keystroke dynamics [158, 159, 160, 265]. Similarly, Yampolskiy and Govindaraju [296] observed that gait-based authentication could be tricked by an impersonator imitating the walk of the registered user.

Adversarial attacks entail the process of so-called adversarial examples [16, 114, 127] by adding small and human-imperceptibly perturbations that can lead to misclassification. Such attacks have been shown to be successful in the context of faces [74] but can also be used to protect users' privacy in video conferencing [245].

When an attacker gains access to the database, they can reconstruct the original input from the biometric templates. This has been shown possible for several biometrics, e.g. face recognition [172], fingerprints [50], and iris recognition[7].

There are various countermeasures to each of those types of attacks, that we will not go into detail here. However, this shows that similar to other types of authentication, attacks are a real possibility and have shown to be successful numerous times.

2.4.4 Explainability and Interpretability

Due to their use of deep machine learning models, the decisions of biometric models can be hard to interpret. However, it is important both for developers and users of biometrics to understand the models' behavior and potentially improve it. The most common approach for this is visual highlighting of relevant input areas. Stylianou et al. [257] propose a method that highlights regions that contribute to pairwise similarity (e.g. of two faces). Yin et al. [299] trained a model in such a way that its features correspond to face areas, enabling saliency maps that correspond to meaningful facial features. Engelsma et al. [88] highlight extracted feature points for fingerprint recognition. Shi and Jain [248] propose to embed faces as a probability distribution instead of single points or feature vectors. In this model, the variance of the distribution corresponds to the uncertainty of the corresponding features and can thus also be used for visualization. Finally, Terhörst et al. [264] took a different approach and analyzed the impact of face features on performance. Some features were predictive of decreased performance and could thus give labeled explanations of model biases.

All of those approaches were designed for model inspection by developers and explorations of their use for end-users are as of now largely missing. The focus is as of now mainly on the use of visual inspection using heatmaps. In Section 2.5.1 we give some more details on explainable artificial intelligence independent of the biometrics use case and how its methods could be extended to facilitate more user-centered biometric interfaces.

2.4.5 Privacy

In a study by Elliott et al. [86] participants voiced their concerns when using biometrics including which applications and who would have access to their data. Such concerns are not uncommon and unfortunately often warranted (e.g. [268]). In particular behavioral data is very sensitive, as it can reveal information on health [223] or allow for building movement profiles. In the EU, the General Data Protection Regulation GDPR[20] requires data protection by design and defines biometric information as personal data, requiring special protection.

Using sensitive data to build biometric models in a privacy-preserving manner is a challenging task. Aggarwal et al. [6] suggest the use of federal learning for face recognition models. In this approach, models are trained locally and fused afterward to avoid amassing a central dataset of private (biometric) data. Other approaches are based on Homomorphic Encryption (e.g. [145]), proposing to train biometric models on encrypted data.

However, in the end, privacy is also always a question of user acceptance. As such it is vital to inform them about the collected data and the uses and risks associated with using a biometric model.

[20] GDPR: `https://gdpr-info.eu/`, last accessed October 16, 2024

2.5 Learning from other Fields

Research on biometric interfaces is sparse and often limited to very specialized use cases (e.g., enabling mimicry attacks [159, 265] or locking approaches for implicit authentication [157]). Thus, our work actively borrows concepts and methods from other fields. Here we give a general overview of other research areas that we leveraged as inspiration to inform biometric interfaces and their evaluation. However, as our work relates to widely different fields, we highlight more specific background and related work in the respective chapters of this thesis.

2.5.1 Explainable AI

Explainable Artificial Intelligence (xAI) is a field concerned with allowing introspection into complex (machine learning) models and making them understandable. With biometrics being a special case of machine learning models, findings and techniques of explainable AI are also relevant here. Models can either be conceptualized to be interpretable (transparent-box) by design [118, 299] or *post-hoc* techniques can be used to gain insights into a given black-box (i.e. non-transparent) model.

Such post-hoc techniques include text and local explanations, visualizations, explanations by example or simplification, and the use of feature relevance [14]. Some of those techniques also have the potential to explain model decisions and contributing factors in user-centered interfaces. For example, textual explanations [27] can give insights into the reasoning process of a model. Similarly, explanations or visualizations of feature relevance [217, 264, 299] can give direct insights into contributing factors, as long as the respective features are meaningful. Visualizations [200, 241] are the most used technique in exploring decisions of biometrics so far (see Section 2.4.4), though they are mostly targeted at developers and experts and it is unclear if they can be interpreted by end-users. Finally, explanations by example [33] can be a viable method of illustrating corner cases in an easy to comprehend way for end-users.

As such, explainable AI proposes techniques that can help uncover and communicate weaknesses and influencing factors in machine learning models. However, they follow a model-focused approach where the focus is on analyzing models for further development to uncover weaknesses in general. In this thesis, we instead put (end-)users and their interaction with a biometric system at the center of our investigation. Nonetheless, we both use insights gained through explainable AI methods as well as post-hoc techniques like explanations and examples to communicate those findings to the end-users in our proposed interfaces.

2.5.2 Interfaces in Usable Security

When looking at other areas of usable security, in particular on knowledge-based authentication we observe two main directions of interfaces. On the one hand, there are interfaces designed to mitigate different kinds of attack vectors where researchers propose extensions or completely novel interactions for that purpose. On the other hand, there are interfaces that use nudging techniques to convince users to choose more secure secrets. Here, we give a short overview of some of those approaches and illustrate, what we can learn for the design of biometric interfaces.

Mitigating Observation and Reconstruction Attacks

One driver of interface design in usable security has been the mitigation of common attack vectors. Such threats can be, for example, observation attacks, where an attacker observes the user entering their secret (and can then use that to gain access), or reconstruction attacks, where an attacker can reconstruct the secret from some kind of residue. Some possible attack vectors are shoulder surfing [84], thermal attacks [2], or video-based observations. De Luca et al. [72] proposed and evaluated authentication at the back of the device to make it invisible to an attacker and Khamis et al. [155] suggested a combination of different modalities for entering a secret to make observation more difficult. Similarly Von Zezschwitz et al. [278] suggested an interface facilitating swiping gestures with directional cues that disappeared on touch to make observation harder. To protect patterns from being leaked through smudge traces on the device, von Zezschwitz et al. [281] designed a pattern interface in a way that would leave less interpretable smudges. A similar approach was suggested for graphical passwords by transforming the underlying image [237]. To counteract attacks on the fallback authentication mechanism Tiefenau et al. [267] suggest augmenting the unlock screen by adding information about the reason for the switch (e.g. too many failed attempts at authentication by fingerprint).

Those examples show, that a focus on different threats can be a valuable approach to designing new interfaces. However, the main focus of this related work was on protecting the secrets used for authentication. For biometrics, those secrets are generally always in plain sight. For example, the users' faces or movements can easily be recorded by cameras, and interacting with everyday objects leaves fingerprints on them. As such, it is hard to hide those features, and attack mitigation is often a task for the biometric systems themselves (e.g. by detecting replay attacks).

Still, we make an attempt at designing interfaces with attack vectors in mind in Part IV of this thesis, where we explore if users can actively change their typing and how they can be supported in doing so (e.g. to evade recognition). Apart from that, we follow a broader interpretation for designing biometric interfaces against threats and focus on non-adversarial impact factors on general performance in the form of demographic and external factors.

Nudging Users to Secure Choices

Besides creating interfaces to mitigate attacks, another big application for interfaces in usable security is convincing users to make good privacy- and security choices.

An underlying observation for this type of interface is, that users tend to choose similar secrets. Examples are picking keys in a row like “1234” as a password, starting a lock pattern in the bottom right corner [280] or choosing salient points in images for graphical passwords [37]. This can then inform the design of nudges as an effective tool to convince users to make more secure choices. Von Zezschwitz et al. [280] used different background images to nudge users to start their patterns in different locations. Seitz et al. [240] used the decoy effect to influence password choice. While they found no effect of their decoy, presenting a comparison alone was enough to nudge users to choose more secure passwords. Similarly, the use of password strength indicators has been shown to be effective in nudging users to choose stronger options [165, 272]. Seitz and Hussmann [239] designed a game and found it effective in increasing users’ awareness of password strength.

However, nudges can not only be used to improve the choice of secrets. Busse et al. [46] evaluated the effectiveness of incentives to foster the adoption of two-factor authentication for games, finding, that they can be an effective nudge. De Luca et al. [70] augmented a keyboard with visual cues, nudging users to avoid unsafe decisions (e.g. entering a password on an unprotected website). Tiefenau et al. [266] made privacy settings graspable in the form of a “privacy hat” and could nudge users to more actively engage in controlling them.

While the use of nudges in usable security is technically speaking also a design choice aimed at threat mitigation (e.g. guessing attacks), they also motivate a new and interesting perspective for our work. The design of nudging interfaces generally follows a two-step process of first uncovering a kind of unsafe or otherwise undesirable behavior and then nudging users through explanations or other approaches toward better options. We use a similar approach in this thesis by understanding user needs and designing biometric interfaces to support their literacy and nudging them to secure choices while preserving user agency.

When using biometric systems, users do not have agency over the input and thus are technically unable to choose a more secure variant. However, similar to knowledge-based nudges, introspection into the underlying mechanisms can be used to inform user decision-making. Knowing how biometrics perform for the individual (e.g. leveraging their demographic background and activity profile) can help users in choosing an appropriate model or correct setting if they are available.

2.5.3 Evaluation of Security Mechanisms

Evaluating security mechanisms is generally a challenging task as authentication is not the users’ primary task and thus gaining reliable insights into their use of authentication mechanisms can be difficult. A recent literature review on the methods used in usable security research found that interviews, experiments, and questionnaires were the most prevalent methods [80]. Both interviews and questionnaires are mostly suited to assess participants' perceptions. However, they often rely on self-reporting which does not necessarily align with actual choices (see e.g. the privacy paradox [108]). To assess interactions with an authentication mechanism experiments are needed.

When designing usable security experiments, many aspects have to be considered [80]. As security is a secondary task, evaluating it as a primary task in a study can lead to overly optimistic results with low ecological validity. Similarly, a lab setting and study framing by itself can already lead to participants behaving differently and caring more for security[83], in particular as they may come in contact with a mechanism or the threat it is designed against more often than they would in reality [107]. As an example, participants in the lab study by Von Zezschwitz et al. [278] unanimously stated they would use the introduced authentication approach in their daily lives. However, a subsequent field study found this not to be the case. Conducting studies in the field can thus paint a more valid picture of the actual use of security mechanisms. However, such studies are also less controlled and many external factors can lead to decreased internal validity.

Finally, many security studies inherently have to involve the risk that the tested mechanism is designed against. The most ecologically valid approach to do so would be to observe the user interacting with the mechanism while this risk naturally occurs (e.g. an attacker is trying to observe their PIN). However, this is generally not practical and can put users and their data and devices at risk. The common solution is to simulate risks and attacks. However, this comes with the aforementioned limitations. As an example, De Luca et al. [72] tested the effectiveness of back-of-device authentication against observations by recording participants' interactions and post-hoc trying to reconstruct the inputs.

When evaluating biometric interfaces, we are confronted with all of those challenges and every decision in the end is a trade-off between different goals. We used surveys and interviews to understand user needs and contextualize our experimental findings. Otherwise, we always tried to allow users to interact with our proposed solutions in as realistic a way as possible and thus opted for field studies when we could. Chapters 9 and 10 describe our approach of first studying an effect (here ability to modify typing behavior) under controlled settings in the lab before implementing a version that was suitable for an in-the-wild study and comparing both approaches.

2.6 Summary: Challenges for Biometric Interfaces

Here we give a short summary and overview of the challenges for biometric interfaces and how they motivate our work.

In many fields of usable security, there exists a plethora of user interfaces, designed to enable new ways of authentication, help users make secure choices, or protect them from risks. For biometrics, such work is still largely missing. Existing work is mainly model-centric, introducing new biometric methods, understanding their biases, and giving insights to experts and developers. This calls for and motivates the *user-centered approach* of this thesis.

Biometrics rely on complex pattern matching and machine learning models, making them hard to understand and predict. However, related work has shown, that biometrics have several shortcomings, including demographic biases and context-dependent performance. Current ways of reporting the performance of biometric models focus on global metrics like accuracy and thus do not account for individual factors. However, for the individual, such effects can be highly important and call for interfaces that communicate such insights to foster *user literacy* and informed use of biometrics.

Techniques like nudging are an effective and often-used tool in usable security research to encourage secure and privacy-preserving user choices while preserving their agency. For biometrics, this is more difficult, as many biometric features cannot be easily changed and thus user agency over the performance of their system is limited. Continuous biometrics additionally introduce the concept of constant authentication which can enhance security but also take away user agency over if and when to be recognized. This calls for approaches to return *agency* back to the users and motivates our investigation of active control of typing features and nudges to mitigate shortcomings of biometrics.

II

UNDERSTANDING USER NEEDS AND DESIGN OPPORTUNITIES

PART II: UNDERSTANDING USER NEEDS AND DESIGN OPPORTUNITIES

In the previous part, we gave an introduction and overview of this thesis. Here we report on our work to gain a better understanding of how biometrics are used as well as what user preferences and concerns are when interacting with them. We conclude this part by exploring considerations and opportunities for the design of biometric interfaces. This work serves as a foundation to inform our further investigations on biometric interfaces presented in Parts III and IV.

- ❖ **Chapter 3** reports on the results of two surveys that were conducted towards the beginning and the end of this thesis to capture the *use and perception* of biometrics and potential advances and changes that happened while this thesis was written.
- ❖ **Chapter 4** introduces a lab study we did to understand user *preferences* when interacting with different biometric mechanisms in comparison to traditional approaches.
- ❖ **Chapter 5** proposes a *design space* derived from an expert focus group that can be leveraged to inform the design of biometric interfaces.

3 Understanding Use and Perception of Biometrics

This chapter is based on the following publication:
Lukas Mecke, Alia Saad, Sarah Prange, Uwe Gruenefeld, Stefan Schneegass, and Florian Alt. 2024. *Do They Understand What They Are Using? — Assessing Perception and Usage of Biometrics.* arXiv preprint arXiv:2410.12661 [186]

The aim of this thesis is the design of biometric interfaces to support users in their interaction with biometric authentication models. However, when designing for users, the first step is understanding their current interactions, knowledge base, and needs. This chapter is dedicated to this goal.

Most modern smartphones offer some kind of biometric authentication method like fingerprint [142], face recognition [275] (for example, "Face ID" on iPhones), or gait recognition [254] as part of Google's "Smart Lock" feature on Android[1]. However we know, that

[1] `https://support.google.com/android/answer/9075927`, last accessed October 16, 2024

those biometrics are inherently based on machine learning models, making them prone to external factors [29] and biases [82]. As a consequence, their behavior can be hard to predict, – even for experts [228]. This leads us to the assumption, that this is even harder for end-users and that they may often not fully understand the biometric systems they use. As a consequence, they may also be unfamiliar with their strengths and weaknesses, and the potential consequences of attacks.

Other fields of authentication have already shown how better understanding their users can help to improve the interaction. Collecting commonly used passwords led to password policies and approaches to aid users in understanding the strength of their passwords (e.g. [239]). Similarly, knowledge of how users chose graphical patterns allowed for approaches to nudge users to a more secure choice (e.g. [37]). The aim of this chapter is to create a similar knowledge base to support and inform the design of biometric interfaces.

However, the perception of biometrics may also evolve with the availability and adoption of new technology [99], so sampling a single point in time may not be sufficient. To address this, we propose to compare the perception and use of biometrics for two points in time.

We thus designed and conducted an online survey in two rounds (first round, 2019: N=57, second round, 2023: N=47) assessing participants' knowledge about biometrics, their current use of this technology, and their perception of security and usability. We explicitly covered both physiological and behavioral biometrics to be able to make a comparison.

We found, that most participants indicated being unable to define or explain biometrics. However, our open-text answers revealed that many participants were able to name correct examples and have some basic understanding while in-depth knowledge was often lacking. More participants actively used biometrics for authentication in the second round of our survey. Contrary to our expectation we did not observe other clear effects between the rounds. Behavioral biometrics seemed overall less known in both rounds and were also consistently rated worse than physiological mechanisms.

> In this chapter we contribute 1) an online survey conducted in two rounds assessing users' perception and use of biometric methods. We 2) identify common themes and misconceptions, and 3) discuss how our insights can be used to improve biometric interfaces and foster future informed use of biometric methods.

3.1 Background and Research Approach

Here we give a short overview of previous related work on the use and perception of biometrics before discussing our learnings and deriving the research questions guiding this work.

3.1.1 Understanding Biometrics Perception

Large parts of the literature about biometrics focus on technical aspects, such as data collection (e.g., [204, 258]), feature extraction (e.g. [106]) or classification (e.g. [259, 296]). Here we give an overview of work investigating user perception of biometrics.

Furnell and Evangelatos [103] used Likert scale questions to understand users' awareness and usage of biometrics, finding that participants preferred methods they had previously heard of and considered easy to use. In a study by Elliott et al. [86] participants voiced their concerns when using biometrics including cleanliness of the devices, safety, and which applications and who would have access to their data. Bhagavatula et al. [29] compared the usability of fingerprint, face recognition, and PIN under different conditions. They found fingerprint to be the overall preferred method with mixed results for face recognition (e.g., because it was unusable in dark environments). A survey on the perception of facial recognition [135] found that despite 90% of the participants being familiar with the technology, only 5% claimed adequate knowledge to build a solid opinion on its usage and its implications. In a large-scale survey (N=10,000), Franks and Smith [98] found that 76% of the respondents used biometric technology, mainly fingerprint and facial recognition. Saad et al. [225] investigated the impact of the Covid-19 pandemic on device usage and authentication in an online survey. They found that the pandemic countermeasures (e.g., sensitization measures, wearing masks) negatively affected biometric-based authentication approaches such as fingerprint and face recognition. In a usability questionnaire on both physiological and behavioral methods, Alhussain et al. [8] showed that 87.3% of participants believed that biometrics (particularly fingerprints) help to protect critical information on their phones. Karatzouni et al. [150] conducted a focus group and found that participants showed interest in adopting biometrics to enhance privacy while also having concerns about constantly being recorded. In an online survey by Rasnayaka and Sim [213], security awareness levels reflected users' willingness to adopt biometric-dependent continuous authentication and Buckley and Nurse [36] found that context is fundamental with regard to acceptability, despite the general observation that users find familiar biometrics most convenient. Sieger et al. [250] found that voice or speaker recognition is not suitable in crowded places. Similarly, a survey by Ellavarason et al. [85] showed that users were concerned with external factors affecting identification performance (for example, surrounding noise on voice recognition). In general, the fingerprint was often chosen as the most secure biometric identification approach [29, 36].

3.1.2 Implications of Related Work

Related work often focused on technical aspects of biometrics (for example, with the aim to improve the underlying models), specific methods (e.g., only face recognition[135]), or specific contexts (e.g., [29, 225]). Furthermore, related work often used predefined questions – leaving participants less room to express their own experiences – and assessed knowledge of biometrics after giving a definition (e.g., [47]) instead of exploring their initial association.

For this chapter, we extend previous work by taking a more holistic approach and including both behavioral and physiological biometrics without a specific scenario. We allow for more open (and unprimed) responses – in particular when assessing knowledge – and investigate use in daily life and changes over time. To our knowledge, no previous work attempted to compare their results over a larger time span to observe changes.

3.1.3 Research Questions

Overall, we aim to assess which biometrics participants know and use, and if they understand them. As a second step, we try to gain insights into users' perceptions of the usability and security of biometrics and uncover potential misconceptions they might have.

Our work is thus guided by the following research questions:

RQ1 **Literacy**: Do participants know what biometrics are and can they explain how they work?

RQ2 **Perception & Usage**: What is participants' personal view about biometrics and where do they see their value, both in their daily life and in general?

RQ3 **Usability & Security**: How do participants perceive usability and security aspects of biometric methods and how do they think they can be improved?

3.2 Survey

Here, we introduce our study design, the structure of the survey, our recruitment strategy, and the data analysis approach.

3.2.1 Study Design

Our study design follows a mixed-methods approach with two independent variables. To extend previous work, we distinguish between TYPES of biometrics and explore our research questions for both *physiological* and *behavioral* methods. To uncover potential changes over the past years we compare quantitative results between ROUNDS and employ a two-step coding approach (see Section 3.2.4).

3.2.2 Survey Structure

To address our research questions (Section 3.1.3), we designed an online survey that we repeated in two rounds to explore changes over time. The survey comprised eight parts as described below. Refer to Appendix A for the full list of questions.

A **Preface**: Participants were informed about the study and consented to the data collection. To test awareness and understanding, it was important to ensure that participants did not look up terms or return to change their answers once they were given a definition. We asked participants to follow those guidelines and disabled returning to previous questions in the survey.

B **Demographics**: This part included questions about the participant's age, gender, and occupation. Participants were also asked to estimate their technical knowledge on a 5-point Likert scale.

C **Biometric Methods**: In this part, we asked participants if they were familiar with the concept of biometric methods and – in case they were – to explain the concept and how it works. If they were unfamiliar with the concept, we asked them to answer with their thoughts instead.

D **Briefing**: At this point, we gave participants the following definition of biometrics and a short explanation[2]:

> **Biometrics**: *automated recognition of individuals based on their biological and behavioral characteristics* (ISO 2382-37) [136].
> In other words, a biometric system uses unique characteristics in human physiology or behavior to accurately identify individuals.

We then asked participants for biometric methods they knew, methods they used in their daily lives, and other application areas of biometrics they knew or could think of. Finally, we asked them to think of other characteristics that could be used for biometric identification.

E **Interlude**: In this part, we explained the term authentication and gave examples for biometrics:

> One of the main application areas for biometrics nowadays is **authentication**. That means that a user can verify their identity to, for example, access an account or device.
> Common examples for **physiological** biometrics include fingerprint recognition, face recognition, and iris scans.
> Common examples for **behavioral** biometrics include gait recognition (walking patterns), keystroke dynamics (typing behavior), and interaction behavior (e.g. credit card usage surveillance to prevent fraud).

We then asked for the participants' smartphone OS and authentication scheme. We moved this from part B to avoid bias towards smartphone authentication in parts C-D.

[2] Note that we did not expect participants' answers in part C to exactly match this definition, but rather wanted to ensure a common ground of knowledge for the rest of the survey.

F **Biometric Perception**: In this part, we asked participants to answer perception questions about using biometrics on a 5-point Likert scale.

G **Performance & Security**: Here, we asked for participants' confidence in biometrics, the impact changes in their appearance or behavior might have, and their fallback strategy if the biometric system did not work. We further asked for the hack's consequences for them, their approach to attacking a biometric system (to explore their understanding of possible attack vectors), and their ideas to improve biometrics.

H **Conclusion**: Finally, we gave participants the option to comment on their previous answers, leave a general comment, and leave their email to participate in the raffle.

Parts C, F, and G were repeated for physiological and behavioral biometrics with part C being counterbalanced. We did so to avoid bias in the awareness and understanding questions based on previously answering the questions for the other biometric group. We did not counterbalance the later parts of the survey, as definitions and examples for both groups were given in parts D and E, respectively.

Overall, we tried to avoid biases where possible and used counterbalancing where this was not the case (e.g. we counterbalanced the order of questions about physiological and behavioral biometrics). On the other hand, we wanted to ensure that all participants were on the same page regarding the terms used. Thus, we structured the survey to provide definitions and additional information only after we had asked for previous experiences and knowledge and before other questions where the knowledge was relevant/needed.

3.2.3 Participants & Recruitment

In both rounds, we followed the same recruiting strategy and advertised the study via social networks and university mailing lists. To not prime participants prior to survey participation, we kept the invitation (title: "Survey about biometric perception") and introduction to the survey on an abstract level. We recruited N=57 participants in the first round (December

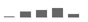
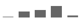
	Round 1		Round 2		
Gender	33	(58%)	26	(55%)	Female
	24	(42%)	15	(32%)	Male
	0	(0%)	6	(13%)	Not stated
Age	29.2		27.4		Mean
	18-66		18-64		Range
Technical Knowledge	3.3		3.3		Mean

Table 3.1: Demographics of the participants of the first (N=57) and second (N=47) round of our survey. Technical knowledge was assessed on a 5-point Likert scale.

2019) and N=47 in the second round (January 2023). The survey was conducted in English and participants in each round could voluntarily participate in a raffle for three 30 € online shopping vouchers. In both samples, participants were mostly students from non-technical fields with a slight bias towards female participants. They had a mean age between 27 and 29 years and medium technical knowledge. Table 3.1 provides an overview.

3.2.4 Data Analysis

Four authors analyzed the responses to the open questions. We started with the responses given in the first round, following the approach for thematic analysis by Braun and Clarke [34] – an approach for inductive theme generation. After an initial phase of familiarization with the provided statements, we independently applied open coding to the statements of the first round[3]. In a review meeting, we discussed and iteratively refined the codes. We then constructed an online affinity diagram [122] of these open codes and organized them into groups, which were in a next step further refined into themes using an online whiteboard[4].

As a result of the analysis, we derived a codebook containing the first round's themes, groups, and codes. The same authors continued the analysis with a deductive approach by independently applying the codebook to the statements of the second survey round. Our rationale behind this two-step approach was to find differences between the rounds based on codes disappearing or new codes emerging in the second round. We reviewed the coding in a final meeting (see Appendix A for the final codebook). Due to the exploratory nature of our study, we refrain from reporting inter-rater agreement scores [177]. Any disagreements were resolved through discussion.

3.2.5 Limitations

We used two independent samples which may have induced underlying differences in the groups that did not result from the temporal distance. To minimize this effect, we exactly replicated our recruitment strategy and compared the demographics from both rounds; finding them to be very similar (see Section 3.2.3). Our sample was self-selected and biased towards young female students and thus, our results may not apply to the general population. Security behaviors as stated in our survey might differ from participants' real-world behavior. Lastly, experimenter bias may have impacted our results. To address this, four researchers were involved in the analysis of open-ended questions (see Section 3.2.4). As such, we believe that this would not influence the resulting discussion and practical implications.

[3] In particular, we did not pre-assume participants' statements to exactly match the ISO definition given in Section 3.2.2 (part D), but followed an open coding approach based on the collected answers.

[4] Miro: https://miro.com, last accessed October 16, 2024

	Round 1			Round 2			
	Yes	No	NA	Yes	No	NA	
Familiar	12 (21%)	45 (79%)	–	7 (15%)	40 (85%)	–	PB
	4 (7%)	53 (93%)	–	5 (11%)	42 (89%)	–	BB
Access	32 (56%)	11 (19%)	14	30 (64%)	11 (23%)	6	PB
	27 (47%)	6 (11%)	24	24 (51%)	3 (6%)	20	BB
Change	32 (56%)	10 (18%)	15	31 (66%)	8 (17%)	8	PB
	32 (56%)	5 (9%)	20	31 (66%)	2 (4%)	14	BB

Table 3.2: Participants' answers to the questions if they were *familiar* with the concept of physiological/behavioral biometrics (PB/BB), if they believed someone could *access* a device protected by PB/BB, and if *changes* in their physiology/behavior would impact a PB/BB system.

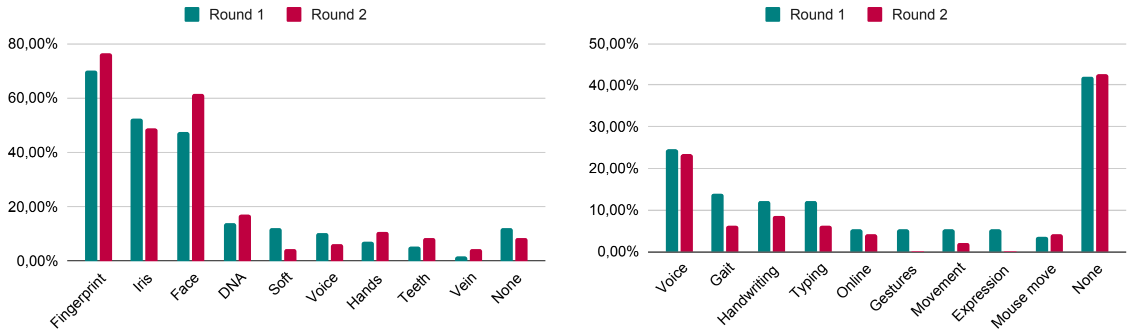
3.3 Results

In the following, we present our quantitative and qualitative findings. We structure this section based on the themes identified in our thematic analysis (see Appendix A for the codebook) and add quantitative results from our survey where they thematically fit. We indicate the number of participants mentioning specific themes to provide a descriptive overview of our data. However, we cannot assume that participants not mentioning a specific aspect is equivalent to them not knowing the answer. Thus, we only conduct statistical tests on our quantitative results. We cite participants from both samples with their IDs as assigned by our survey tool and indicate the respective survey round (e.g. P12$_{R1}$ would refer to participant 12 who was part of the first round of our survey). We distinguish between types of biometrics by using PB and BB in subscript for physiological and behavioral biometrics respectively.

3.3.1 Definition and Function of Biometrics

We asked participants to define and explain, with their existing knowledge, what biometrics are and how such methods work. Participants were encouraged to guess if they were unfamiliar with biometrics. We prefaced this open question with a binary choice, where only a minority of participants in both rounds indicated they were familiar with either physiological (familiar$_{R1}$ = 21%, familiar$_{R2}$ = 15%) or behavioral (familiar$_{R1}$ = 7%, familiar$_{R2}$ = 11%) biometrics (Table 3.2). Using Fisher's exact test, we found no effects of the type of biometrics or the round on reported familiarity.

While we did not expect participants to exactly replicate a technical definition (such as given in Section 3.2.2, part D), we saw that many of them mentioned features related to biometrics, either remaining rather generic (e.g. naming just "physiological features") or giving concrete examples like face geometry or fingerprints ($N_{R1} = 45$ and $N_{R2} = 19$). Some participants also explicitly mentioned where they had heard of biometrics (P181$_{R1}$: "*I only know the word biometric from ID suitable photographs*", P98$_{R1}$:"*Like in Mission: Impossible – Rogue*



(a) Known *physiological* biometrics

(b) Known *behavioral* biometrics

Figure 3.1: Known biometrics as mentioned by the participants. We excluded mentions of unrelated methods as well as biometrics that were mentioned by less than 3 participants across both rounds.

Nation where they measure how you walk”). Fewer participants included a correct or related verb (e.g. recognize or authenticate) in their definition ($N_{R1} = 32$ and $N_{R2} = 15$) or only mentioned an example related to biometrics ($N_{R1} = 24$ and $N_{R2} = 12$). At the same time, a large number of participants showed some missing knowledge ($N_{R1} = 48$ and $N_{R2} = 34$) by either explicitly stating to have no idea ($N_{R1} = 29$ and $N_{R2} = 4$) or referring to other concepts. Those often revolved around related topics like medicine and health (e.g. P130$_{R1}$: “*The status of ones health in numbers*”), body functions (e.g. P129$_{R1}$: “*It could be about the way we perceive the locations of our extremities*”) or influences on behavior (e.g. P124$_{R1}$: “*trying to derive how people behave from their physical features*”, P280: “*How and why we behave as we do*”). Almost all participants left at least one of the questions about defining and explaining biometrics empty ($N_{R1} = 53$ and $N_{R2} = 42$). Finally, some participants explicitly expressed confusion concerning the terms physiological and behavioral biometrics ($N_{R1} = 2$ and $N_{R2} = 9$). Interestingly, one participant expressed that the term physiological biometric seems to be incorrect, saying “*face, fingerprints, and such for me are anatomical characteristics, not physiological*” (P297$_{R2}$).

After giving participants a definition of biometrics in the survey, we asked them to name all biometric methods they knew (see Figure 3.1). The most mentioned physiological approaches were fingerprint ($N_{R1} = 40$, $N_{R2} = 36$), iris or retina scans ($N_{R1} = 30$, $N_{R2} = 23$), and face recognition ($N_{R1} = 27$, $N_{R2} = 29$). For behavioral biometric methods, the most common mention was voice recognition ($N_{R1} = 14$, $N_{R2} = 11$), followed by gait ($N_{R1} = 8$, $N_{R2} = 3$), handwriting or signatures ($N_{R1} = 7$, $N_{R2} = 4$), and typing or keystroke dynamics ($N_{R1} = 7$, $N_{R2} = 3$). Overall, participants knew more physiological methods, which is also reflected in the high number of participants who indicated being unable to name a single behavioral method ($N_{R1} = 24$ and $N_{R2} = 20$). Many participants mentioned features that are associated with biometrics but not commonly used in the consumer market. Examples are features used in a forensic context (e.g. DNA or teeth) and for profiling and tracking purposes (e.g. movement, mouse movement, online behavior). Some participants mentioned

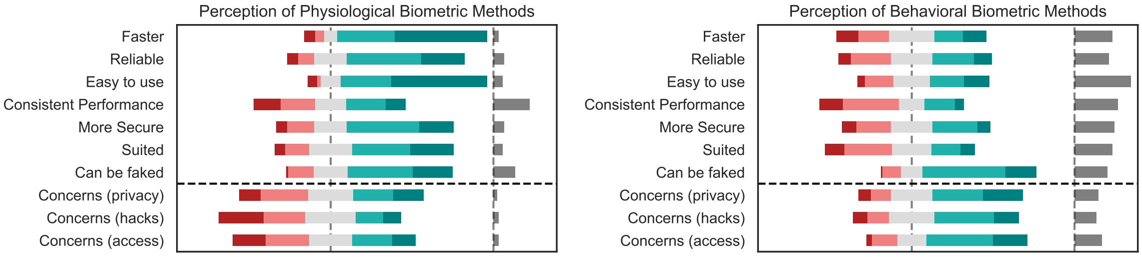


Figure 3.2: Participants' ratings on the Likert statements combined for both rounds of our online survey. Participants that did not give a rating are indicated in gray. See Appendix A for the full questions.

features like height or eye color (also called soft biometrics [68]) that have some biometric value but are not commonly used alone but rather in conjunction with other biometrics. Notably, voice was mentioned both as a behavioral and physiological trait. While voice recognition is often considered to be a behavioral biometric method the distinction is not completely clear cut and voice does have a clear physiological component to it [144].

3.3.2 Perception and Usage of Biometrics

In the survey, participants mentioned several aspects related to the usage of biometrics. Many participants gave concrete usage examples of biometrics ($N_{R1} = 29$, $N_{R2} = 27$), such as fingerprint, face recognition, ID cards, and signatures, among others. Moreover, they mentioned specific devices and use cases in which biometrics are utilized ($N_{R1} = 19$, $N_{R2} = 23$), primarily mentioning mobile devices and computers. A few participants provided reasons as to why they use biometrics ($N_{R1} = 4$, $N_{R2} = 7$): for example, they stated that biometrics are easy, fast, safe, and less error-prone. One participant stated to "*use the facial recognition and fingerprint scanner on [their] phone and tablet, to unlock [their] devices more easily and avoid the danger of other people seeing [their] PIN code or password*" (P324$_{R2}$). In contrast, a few participants stated directly that they do not use biometrics and use, for example, passwords instead ($N_{R1} = 1$, $N_{R2} = 3$). Some more participants gave reasons against using biometrics ($N_{R1} = 14$, $N_{R2} = 7$), mentioning not having a reason for using them, having concerns about privacy and recovering from data getting compromised or pointing out issues with the recognizer/classifier. Some participants just did not like the thought of using biometrics with P20$_{R1}$ finding face recognition "*creepy and insecure*" and P317$_{R2}$ stating to "*feel more comfortable not using it*".

We also asked participants about their mobile device authentication scheme (see Table 3.3 for full results). The most common combination used was fingerprint authentication ($N_{R1} = 30$, $N_{R2} = 23$) with a PIN as a fallback ($N_{R1} = 28$, $N_{R2} = 25$). While face recognition was not used among our participants in the first round of the survey, 12 participants (26%) used this scheme for authentication in the second round.

	Round 1		Round 2		
Operating System	38	(67%)	28	(60%)	Android
	17	(30%)	18	(38%)	iOS
	2	(4%)	1	(2%)	Other
Unlock	30	(53%)	23	(49%)	Fingerprint
	9	(16%)	5	(11%)	PIN
	8	(14%)	2	(4%)	Pattern
	8	(14%)	3	(6%)	Slide/None
	2	(4%)	2	(4%)	Password
	0	(0%)	12	(26%)	Face Recognition
Fallback	28	(51%)	25	(53%)	PIN
	15	(26%)	7	(15%)	None
	7	(12%)	6	(13%)	Pattern
	4	(7%)	9	(19%)	Password
	2	(4%)	0	(0%)	Other

Table 3.3: Operating system and authentication schemes used by the participants of the first and second round of our survey.

Finally, we also asked participants to rate a collection of Likert statements on the usability and security of biometrics on a scale from 1 (strongly disagree) to 5 (strongly agree). Figure 3.2 shows the results. We compared ratings across rounds with a Mann-Whitey U test, finding significant differences only for participants' perceptions of the consistency of biometrics. Participants in the second round disagreed with the performance of biometrics being equal for all users while they were neutral in the first session (Z=4092.00, p=.014).

To find potential differences in the perception of physiological and behavioral biometrics, we conducted Wilcoxon tests excluding answers where neither type of biometrics was rated. Physiological biometrics were rated significantly faster compared to Pin/Password than behavioral biometrics (Mdn$_{PB}$ = 5, Mdn$_{BB}$ = 3, Z=5.60, p<.001). Similarly, they were rated more reliable (Mdn$_{PB}$=4, Mdn$_{BB}$=3, Z = 3.80, p<.001) and easier to use (Mdn$_{PB}$ = 5, Mdn$_{BB}$=3, Z=4.96, p<.001). Participants rated physiological biometrics as significantly more secure compared to PIN/Password than behavioral biometrics (Mdn$_{PB}$ = 4, Mdn$_{BB}$ = 3, Z=2.41, p<.014) and found them better suited to protect their personal data (Mdn$_{PB}$ = 4, Mdn$_{BB}$=3, Z=4.06, p<.001). In contrast, concerns about privacy (Mdn$_{PB}$ = 3, Mdn$_{BB}$ = 4, Z=-3.14, p=.001), about being hacked (Mdn$_{PB}$ = 3, Mdn$_{BB}$ = 4, Z=-4.45, p<.001) and about loosing access (Mdn$_{PB}$ = 3, Mdn$_{BB}$ = 4, Z=-4.45, p<.001) were all rated significantly higher for behavioral biometrics.

3.3.3 Attacks and Challenges

The majority of participants in both rounds of our survey indicated their belief that someone else could access their device when using physiological (access_{R1} = 56%, access_{R2} = 64%) or behavioral (access_{R1} = 47%, access_{R2} = 51%) biometrics (see Table 3.2). Notably, a large group of participants indicated they did not know if this was possible – particularly for behavioral methods (43% and 42% in the two rounds respectively). We saw very similar results regarding the effect of changes on biometric methods. A majority of participants believed that changes in their physiology or behavior would have an impact on the respective biometric methods (impact_{R1} = 56%, impact_{R2} = 66%). We did not find effects of biometric type or survey round on both of those measures when using a Fisher's exact test.

In their open-text responses, almost all participants mentioned aspects related to attacking biometrics and current challenges. Many participants mentioned non-malicious reasons why biometrics could fail. They gave generic reasons ($N_{R1} = 18$, $N_{R2} = 23$), such as changed hardware or a longer time span between authentication attempts and physical reasons ($N_{R1} = 20$, $N_{R2} = 24$), such as cosmetics, haircuts, or injuries. For example P297_{R2} named "*diseases of skin [and] injuries*" as potential reasons for fingerprint not working, adding they had personal experience with non-detection from "*very dry skin with deep cuts from outdoor work*". Finally, participants mentioned behavioral reasons ($N_{R1} = 2$, $N_{R2} = 7$), such as an impact resulting from mood or a purposefully changed behavior (e.g. P16_{R1}: "*I type much faster when I argue with my girlfriend. This might bias the system*"). Moreover, they saw different attack vectors for biometrics. They mentioned software-based attacks ($N_{R1} = 13$, $N_{R2} = 10$), including hacking or circumventing of biometrics. They described the application of force ($N_{R1} = 7$, $N_{R2} = 9$), including destroyed hardware, removed body parts, or an attack during sleep (e.g. P314_{R2}: "*When sleeping/unconscious most of the physiological biometrics can be used without my consent*"). Participants further mentioned imitation/replay attacks ($N_{R1} = 36$, $N_{R2} = 36$), including deepfakes and mimicry attacks (e.g. P328_{R2}: "*trying to build an imprint of a fingerprint or (more elaborate): building a face-mask*"), and gave some other attack vectors ($N_{R1} = 11$, $N_{R2} = 5$), including social engineering or outsourcing attacks (e.g. P21_{R1}: "*I would pay someone to do it for me*"). Furthermore, many participants expressed perceived weaknesses of biometrics that exist from their perspectives ($N_{R1} = 29$, $N_{R2} = 20$), such as the possibility of faking them or the collection of their identity by governments or companies. A few participants stated that they do not believe that biometrics can be attacked or it would not have an impact ($N_{R1} = 9$, $N_{R2} = 4$). For example, as an answer to whether physiological changes could impact biometrics, P98_{R1} answered: "*Not for fingerprints*".

3.3.4 Consequences of an Attack

Most participants explained how they would deal with a successful attack on their biometric authentication and voiced potential damage resulting from such an attack. In case of a successful attack, participants highlighted different actions they would take. Most participants stated that they would fall back to another authentication method ($N_{R1} = 50$, $N_{R2} = 43$), while several participants stated they would try to control the damage ($N_{R1} = 24$, $N_{R2} = 21$), by, for example, informing contacts, recovering data or resetting their device. A different strategy was seeking support ($N_{R1} = 24$, $N_{R2} = 10$), by contacting the police or their provider (e.g. P67$_{R1}$: "*I would research the right institution to call in this case or go to the police*"). Concerning the potential damage resulting from a successful attack, participants mention that the attacker could misuse (e.g. P124$_{R1}$: "*I would be unable to prove that this person is not me*") or simply access their data ($N_{R1} = 17$, $N_{R2} = 18$) or that they would lose their data ($N_{R1} = 17$, $N_{R2} = 11$). A few participants stated that an attack would have no impact or they did not intend to react to it ($N_{R1} = 9$, $N_{R2} = 4$).

3.3.5 Future Suggestions & Improvements

Finally, participants listed different suggestions that could improve biometric systems in the future. Many participants mention novel forms of biometrics ($N_{R1} = 37$, $N_{R2} = 34$), including body and facial movements. P343$_{R2}$ suggested using "*genetics through noninvasive tissue sampling (high level of security should sequencing in real time become possible)*". Moreover, participants stated new applications for which biometrics can be used in the future ($N_{R1} = 7$, $N_{R2} = 5$), including payments (e.g. P19$_{R1}$:"*Maybe payment with DNA identification*"), forensics, public contexts, and sports. Finally, many participants gave general improvements for biometrics ($N_{R1} = 27$, $N_{R2} = 30$), including tweaking thresholds, updating models with more data, increasing the precision of sensors, and use hard to replicate methods. Finally, another often mentioned improvement was "*using more than one type of [physiological or] behavioral biometric system*" (P304$_{R2}$).

3.4 Discussion

3.4.1 Do they understand what they are using?

Many participants indicated to be unfamiliar with biometric methods, i.e. they were unable to define or explain the concept. Even though only a few participants indicated knowing what biometrics are, a far larger number were able to give examples, name features used, or correctly associate them with authentication. Thus, asking open questions in addition to the self-assessment allowed us to see that many participants seemed to have at least some knowledge about biometrics and answers often included important aspects of the definition we used. A related study found many participants to be aware of face recognition [135], but only a few had deeper knowledge. In our study, we did not focus on face recognition but observed a similar effect for biometrics in general.

However, some participants had only an abstract association (e.g. from having heard the term in the context of their passport). We also received many answers that were not related to biometric authentication but referred to, for example, body functions, health, or influences on behavior. This can mean one of two things: either participants were indeed not familiar with biometrics or they just had a different association with the terms used.

Unfamiliarity can be problematic in several ways. On the one hand, understanding what features are collected and how they are used is essential for informed consent to use those approaches. On the other hand misconceptions and missing knowledge can lead to users needlessly abandoning biometrics or not adopting them in the first place.

Regarding the term itself, we believe that many participants may have made a connection from biometrics to *bio*logy and *metrics* and thus assumed a connection to e.g. body functions rather than security. One participant also explicitly preferred the term anatomical over physiological biometrics. Potentially, renaming physiological biometrics to e.g. appearance-based identification could aid in clarifying their function.

3.4.2 How do they cope with problems?

While many participants had very sophisticated ideas about how biometric systems could be attacked, surprisingly few participants mentioned attacking the fallback method. However, every modern biometric system uses a fallback like a PIN to enable access in case the biometric factor fails. Many participants stated they believe biometrics to be more secure than using those traditional methods but practically, this currently cannot be the case. This is something that biometric interfaces should clearly communicate to users during enrollment.

At the same time about half of our participants thought that someone else could access their device if protected by a biometric system; showing a contradicting tendency. Many participants indicated they would switch from biometrics to other authentication methods in case they experienced issues with changes (e.g. in their appearance) or were to be attacked. Others had no idea how to cope with problems and suggested outsourcing their solution, e.g. to family members or service providers.

Overall, this means we observed two opposite tendencies, with some participants believing that biometrics could not be attacked and being very confident in their security while others mistrusted the technology and had no clear plan for how to handle issues should they come up. As always, the truth lies somewhere in between. Education may help users to gain a more reflected impression of biometrics, avoid a false sense of security, and prepare them for potential issues.

3.4.3 Do they know behavioral biometrics?

In contrast to the majority of related work, we actively distinguished physiological and behavioral biometrics to understand how perception and understanding differed between the two. While many participants could name physiological biometrics, the only behavioral method mentioned by more than 20% of the participants in both rounds was voice recognition. Similarly, more participants did not know if changes would affect behavioral systems and if an attacker could gain access to such a system. Participants had significantly more concerns about behavioral biometrics and rated them slower, less reliable, and harder to use than their physiological counterparts.

All those aspects imply that knowledge of behavioral methods is less prevalent among our participants, leading to increased uncertainty. This intuitively makes sense, as behavioral methods are normally designed to be transparent and facilitate authentication without explicit action or even knowledge of the user. Yet, this is also a risk, as behavioral traits can be used for (unwanted) profiling and tracking purposes (e.g. recognizing users on a website by their mouse movements [210] or in public places by their gait [276]) even though behavioral biometrics are not (yet) widely used.

3.4.4 Comparison Between the Two Survey Rounds

We designed our study to find potential differences in the use and perception of biometrics over time. To achieve this we openly coded the first round and used the resulting codebook on the answers from the second round to be able to find newly emerging themes and codes. Given the increasing adoption of biometrics in users' daily lives over the recent years we would have expected to see a change in users' perception and knowledge between our rounds. However, this was not the case: no codes emerged or disappeared between rounds (see Appendix A).

Our quantitative results show that the usage of biometrics as the primary authentication mechanism on smartphones increased from 53% in the first round of our survey to 75% in the second round, i.e. three out of four of our participants indicated using either fingerprint or face recognition. Participants indicated significantly less agreement with the performance of biometrics being equal for all users, which may hint at an increased awareness of biases in those models. However, this is speculation so far and would need further confirmation in a future study. Apart from that – and again contrary to our expectation – the increased adoption did not seem to have led to significant differences in perception. For instance, we did not observe changes in perceived ease of use, reliability, or general concerns. While it is unclear to which degree participants need a detailed understanding of biometric methods in order to confidently and securely employ them in their daily lives we argue that a better understanding of the risks and potential shortcomings could ultimately foster safe use and trust. Continued research on this topic will be needed. We further believe in the approach of comparing data over a longer time frame, though it may be worthwhile to explore other methods (e.g. using more closed questions or conducting interviews) as well to better understand such effects.

3.5 Implications

In this chapter, we presented our investigation of user perception of physiological and behavioral biometrics. Our results show, that most participants actively use biometric methods in their daily life and prefer them over traditional methods. At the same time, most participants lacked in-depth knowledge about how biometrics work and showed uncertainties with regard to handling potential problems caused by an attack or changes in their physiology or behavior. Between the two rounds of our survey, we saw a strong increase in the adoption of biometric methods and it is very plausible that this trend will continue with improving algorithms and sensors.

This chapter highlights a *need for clear communication* when it comes to the functionality of biometrics and the security they can offer. It further shows a demand to *illustrate and mitigate* the effects of changes and external factors on biometric systems.

In Part III we propose biometric interfaces to address these needs. Chapter 6 introduces a method to illustrate the personal performance of face recognition with generative samples. Chapters 7 and 8 describe interfaces to communicate the internal state of a biometric system as well as the impact of context factors on it. They also offer mitigation strategies to avoid interruptions and switch authentication mechanisms. In Chapter 12 we discuss how we propose to handle communicating terminology.

4

Exploring User Preferences for Biometrics

This chapter is based on the following publication:
Lukas Mecke, Ken Pfeuffer, Sarah Prange, and Florian Alt. 2018. *Open sesame! user perception of physical, biometric, and behavioural authentication concepts to open doors*. In Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia (MUM '18) [183]

In the previous chapter, we explored the use and perception of biometrics through two surveys. Our focus there was to understand how well users comprehend the biometrics they use and how they perceive their usability, security, and reliability. In this chapter, we follow a similar goal but aim to understand what aspects *of the interaction* users value between different types of authentication. This knowledge is valuable to inform the design of biometric interfaces respecting user preferences.

As a use case, we focus on access control through doors. This is a common task that users are familiar with and encounter on a daily basis. It also allows for exploring a scenario

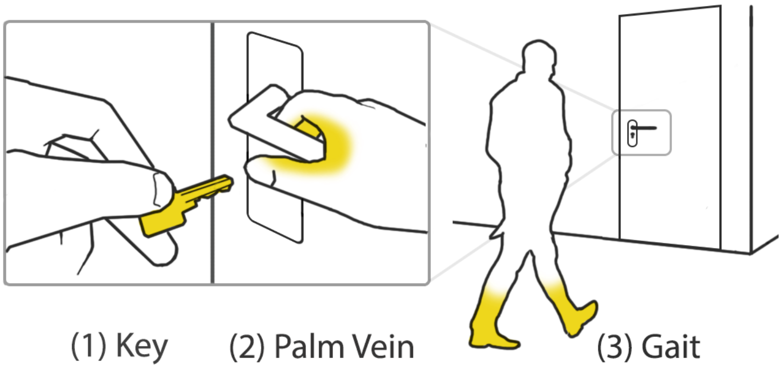


Figure 4.1: In this work we investigate user perception of different authentication mechanisms at doors. Namely, those are (1) a key, (2) a (mock) palm vein scanner, and (3) (mock) gait-based recognition.

beyond digital devices where the value that is protected by the authentication is immediately graspable for users (e.g. access to their homes). While biometric authentication systems have been investigated for electronic devices (e.g. [43, 226]), applications for analog devices are still mostly unexplored. However, both physiological and behavioral biometrics offer the potential to improve interactions here, as the process of unlocking a door could be done seamlessly [105] and - in contrast to most existing approaches - without requiring to either memorize and enter a PIN or carry a physical token like a key or smartphone.

For our comparison, we chose a setup including both a physiological and a behavioral biometric method with a physical key as a baseline condition representing the status quo. To avoid potential impacts of external factors or inconsistent performance [29, 48, 294] of the biometrics between the participants we opted for mocking the interaction with those systems.

In this chapter, we thus conducted a Wizard-of-Oz study, assessing the *user perception* of different authentication mechanisms for unlocking a door. We investigated (1) a physical key as the baseline, (2) a mock palm vein scanner representing a physiological biometric system, and (3) mock gait-based authentication as its behavioral counterpart (see Fig. 4.1). We asked participants to rate the mechanisms and indicate their perception of the usability and security of the compared authentication mechanisms.

We found, that users liked the concept of seamless authentication using biometrics, but still appreciated the control they gained from possessing a physical key. Participants had security concerns for gait recognition and found recovery from authentication failures cumbersome.

> In this chapter we contribute 1) a lab study comparing authentication at doors with different authentication methods and 2) insights into user preferences and concerns as well as advantages of the different methods.

4.1 Background and Research Approach

Here we give some background on palm vein scanners and gait-based authentication before discussing considerations for authentication at doors and deriving research questions to guide our work.

4.1.1 Palm Vein Scanners

In 1968 the first patent for a palm print identification system was granted to N. Altman [10]. The modern palm vein scanner takes an infrared image of the palm to detect vein patterns that are matched to a saved template. Romanowski et al. investigated the acceptability and ease-of-use of a palm vein scanner in 2016 [221]. In their study, 75% of the 55 participants found the technology to be non-intrusive, and 77% did not experience any delays during authentication. The company Fujitsu, as a creator of mass-market palm vein scanners, announced in 2018 that they will replace passwords and smartcards for 80,000 employees in their Japanese headquarters in favor of their palm vein scanner PalmSecure [101]. With these efforts showing high potential, we study palm vein biometric authentication for the purpose of accessing doors.

4.1.2 Gait-Based Authentication

Initially, gait-based recognition became a subject of psychology research in 1977. Cutting and Kozlowski [64] noticed that a person could recognize familiar others simply by an abstract display of the movements made while walking. Visual gait motion data can be processed with pattern recognition methods and matched with registered data [20].

A different approach was explored by Xu et al. [293] who created a gait recognition system for smartwatches, namely “Gait-Watch”, that identifies the user’s distinct way of moving. Sprager and Juric [254] give an overview of this method of recognizing gait from inertial measurements. This unobtrusive form of gait recognition without the use of visual motion capture has also found practical applications for mobile phones e.g. in Google’s “Smart Lock” feature on Android[1].

4.1.3 Considerations for Authentication at Doors

As a baseline for our study, we chose to use a physical key as the currently most used unlocking mechanism. Functionally, this also is very similar to using a token and results should thus transfer between those two mechanisms.

[1] `https://support.google.com/android/answer/9075927`, last accessed October 16, 2024

We wanted to use a physiological mechanism, that avoids additional interaction (e.g. standing still in front of a camera to capture iris information [229]) and is not affected by other context factors like noise or light. This left us with using fingerprints or palm-vein scans as mechanisms that can be integrated into the interaction with a door. Both approaches are as of now not directly integrated into the interaction with a door handle (i.e. they require actively presenting the feature to a sensor). However, to communicate our concept of seamless integration we decided to mock such an interaction and chose the palm vein scanner as the more plausible option (assuming a scanner would be integrated into the handle).

While other proximity-based mechanisms (e.g., NFC technologies) require the user to be at close distance to the door, a functional *gait-based system* would authenticate users by their natural way of approaching the door. Behavioral authentication by such motion is often based on probabilistic measures of walking over time, which requires a larger area. However, in principle, it allows for a completely implicit, i.e. effortless, access through doors.

4.1.4 Research Questions

We derived the following research questions to guide our research:

RQ1 **Preferences**: What type of authentication do participants prefer for authentication at doors?

RQ2 **Perception**: How do participants perceive the interaction with and the security and usability of the different authentication methods?

4.2 Evaluation

The focus of our study was the evaluation of *user perception* and preferences of the interaction with biometric authentication systems at doors. We tested three different mechanisms to unlock a door, using mock-ups and a physical door barrier controllable by the experimenter. Here we give an overview of our study design and the physical setup we built, as well as the procedure of our study and the participants.

4.2.1 Study Design

We designed the study as a within-subject Wizard-of-Oz lab study with a single independent variable. We varied the UNLOCK MECHANISM on three levels, having participants unlock the door using a *physical key*, a *palm vein scanner* integrated into the door handle, and *gait-based authentication* using a Kinect. Both biometric mechanisms were non-functional (i.e. mock-ups). The order in which participants experienced the authentication mechanisms was counterbalanced.

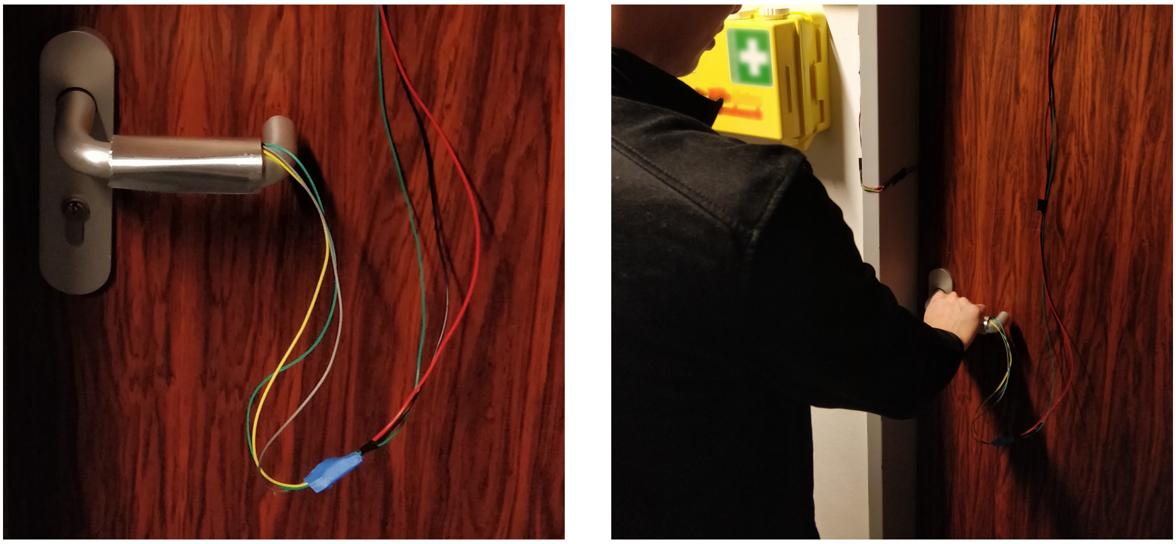


Figure 4.2: Mock-up of a palm vein scanner, made of a thin sheet of metal with some cushioning. It gripped the door handle and was connected to the door lock by visible wires to support the illusion of a running system (left). Participants were asked to grip it to authenticate (right).

4.2.2 Apparatus

An important aspect of Wizard-of-Oz studies is to offer a system that is as believable as possible in mimicking a real system. To support the impression of a functional system, we added a number of technical enhancements: foil and wires at the door handle mocking the palm vein scanner (Figure 4.2), a feedback screen showing the users' skeleton tracked by a Kinect sensor (Figure 4.3, LEDs indicating success of authentication) and a mechanical door lock controllable by the experimenter (Figure 4.4). We give details below.

Door setup

For our study, we used a door with a regular key lock between two rooms. We marked a path and a starting position for the authentication process on the floor with blue tape (see Fig. 4.5). Participants were asked to walk along this path and unlock the door while walking, using one of the three conditions.

To make participants believe that their actions were unlocking the door we remotely unlocked the door by lifting the mechanical blockade using a wifi connection, as soon as a fully implemented system would have recognized the user. We controlled the door lock mechanism using an ESP32 running Arduino software[2]. We offered feedback for the biometric conditions in the form of a green LED turning on after successful authentication and a blinking red LED accompanied by a long beep otherwise (Figure. 4.4, right).

[2] ESP32: https://www.espressif.com/en/products/hardware/esp32/overview, last accessed October 16, 2024

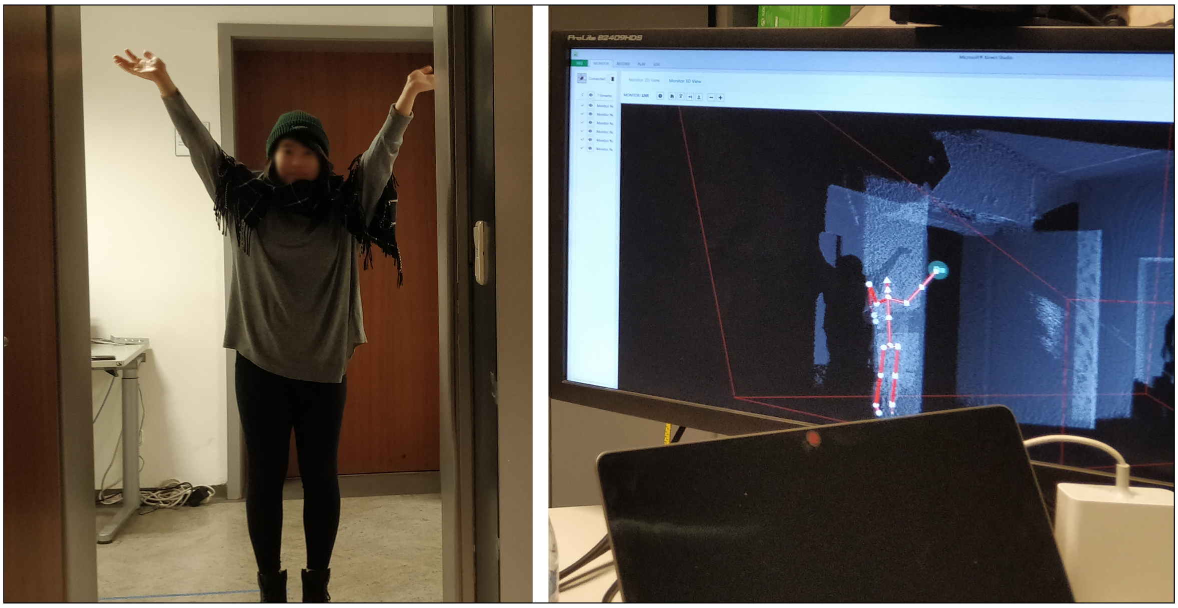


Figure 4.3: Our setup for the mock gait-based authentication used a Kinect to display the body structure of detected humans in the area to create the impression of a running authentication system.

Designing the Authentcation

We took special care to make the interaction with our mocked biometric systems as believable as possible. Here we give details on the implementation of all authentication options.

Physical Key. To authenticate, participants had to insert the key into the keyhole, rotate it twice, and press the door handle.

Palm Vein Scanner. We mocked the palm vein scanner as a metallic surface embracing the door handle that participants had to touch to “unlock” the door. It was connected to our door-lock by visible wires to support the impression of being functional (compare Fig. 4.2).

Gait-based Authentication. We placed a Microsoft Kinect[3] in the middle of the experiment room to create the impression of capturing the participants’ walking behavior between the starting position and the locked door. We placed an additional monitor in the experiment room, which displayed the skeleton data captured by the Kinect in real-time (see Fig. 4.3). The captured data was not used for authentication but had the sole purpose of giving the users the impression that the system could indeed capture their walking behavior.

[3] Microsoft Kinect: `https://developer.microsoft.com/en-us/windows/kinect`, last accessed October 16, 2024

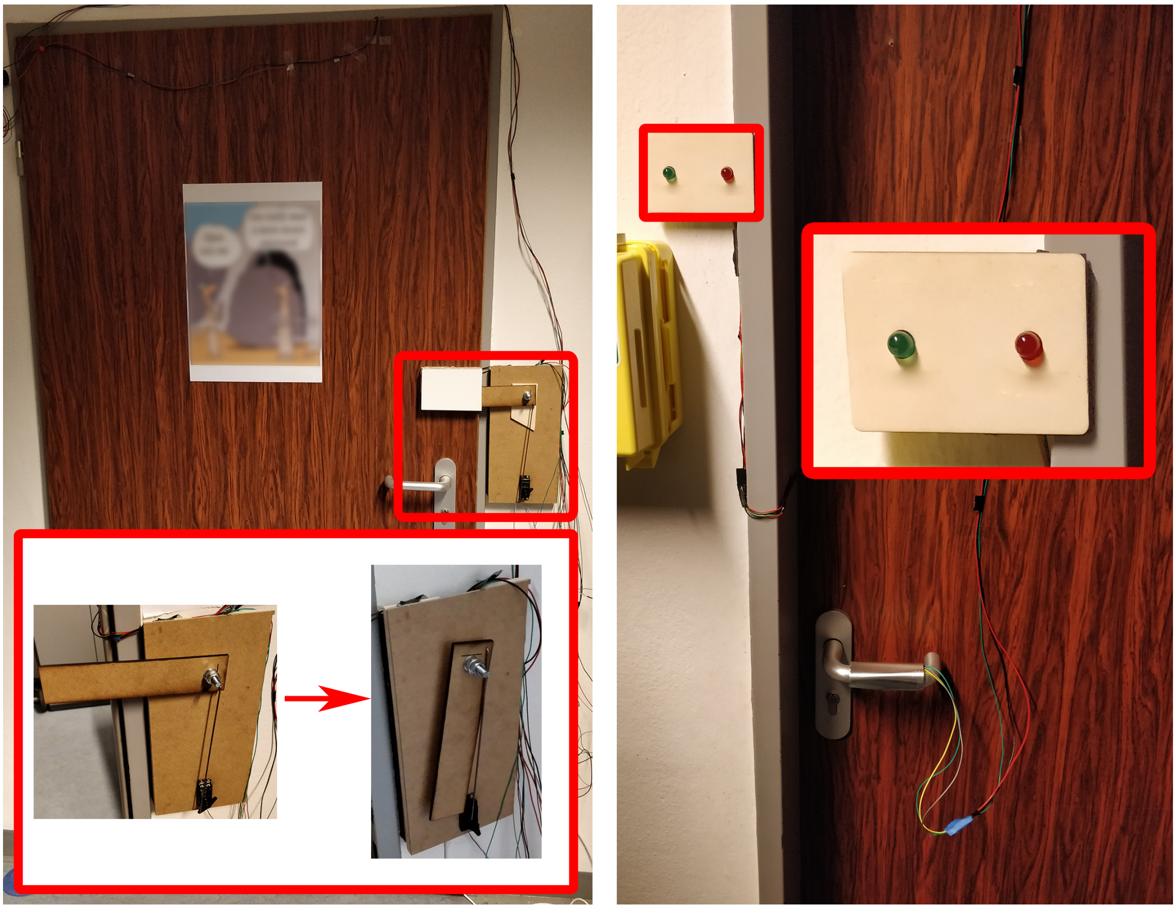


Figure 4.4: Left: The back view of the door. The mechanical door lock was controlled over a wireless network. Right: The front view of the door. Additional feedback regarding the success of an authentication attempt was provided by colored lights at the front side. The green LED on the left indicates success, and the red LED on the right failure.

4.2.3 Procedure

As participants arrived at the lab, we first introduced the purpose of the study. We then had them fill in a demographic questionnaire. After that, participants were asked to use the different door-unlocking mechanisms. The order was counterbalanced. Each mechanism was tested three times. To allow for experiencing situations in which the system failed to authenticate the user in the biometric conditions, we caused one attempt to be unsuccessful with the occurrence again being counterbalanced. Prior to the biometric conditions, participants were required to register themselves by “training the system” (i.e., participants had to use the system a few times prior to the actual study to make the system “capture their data”).

After three successful authentications, participants were interviewed and asked to rate Likert statements (1: strongly disagree, 5: strongly agree) about perceived usability and security. We repeated the questionnaire for all tested unlocking mechanisms.

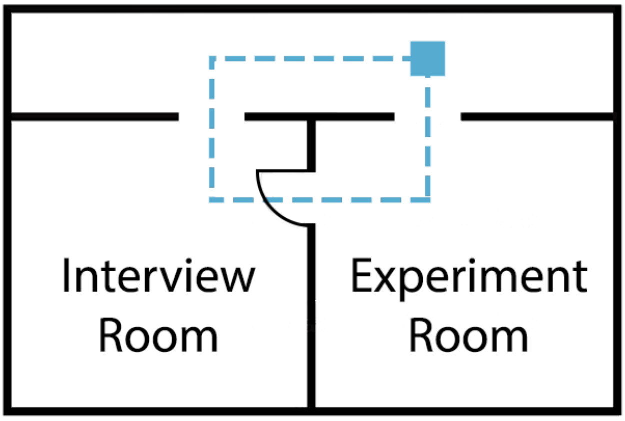


Figure 4.5: The floor plan we applied in our study setup: Participants were asked to walk along the dotted line from the starting position (rectangle in the top right). Participants had to open the door between the experimenters' room and the interview room using either 1) a key, 2) a (mock) palm vein scanner, or 3) (mock) gait-based recognition.

An additional semi-structured open interview concluded the study session. We asked the participants to compare the three authentication systems and if they saw any dangers or benefits when using biometric techniques to open a door. Afterward, we asked the participants to explain their ranking of the authentication systems, how each system could be improved, if a combination should be considered, if they would use it in a daily context, and if the system(s) felt secure. In the last part of the interview, we asked participants how they would handle different situations. We asked what they would do in case of a power blackout (for palm vein scan and gait), if they lost their physical key, if they suffered from a broken arm (palm vein scan) and if they had additional luggage, which would alter their walking behavior. After the last question, we revealed that it was a Wizard-of-Oz study.

4.2.4 Participants

We recruited 15 participants (Mdn age = 23, 14 male, 1 female) for our study. Eleven Participants were students with about half of them being enrolled in IT-related degree programs. From our demographics questionnaire, we found data privacy being an important concern (Mn = 3.47) among the participants.

4.3 Results

Here we report on the results of our study, in particular participants' ranking of authentication mechanisms, their ratings on the Likert statements, and their open Feedback from the interviews.

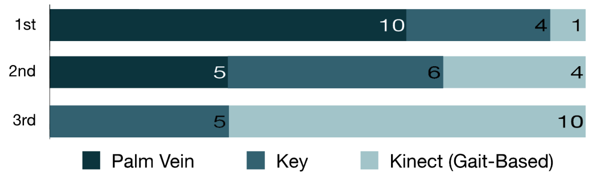


Figure 4.6: Participants' ranking of the three authentication methods. The palm vein scanner performed best with 10 votes for 1st place. Gait recognition was ranked worst.

4.3.1 Ranking

Figure 4.6 shows the rating of authentication mechanisms in our study. Participants mostly preferred the palm vein scanner (10 out of 15 participants ranked it as their most preferred authentication method). Four participants preferred the physical key. The least preferred method was gait-based authentication.

4.3.2 Likert Ratings

The ratings of Likert statements were overall in line with the ranking (See Figure 4.7). For the statistical analysis, we used a Friedman test and a Wilcoxon signed-rank test. For the post-hoc multiple comparisons between conditions, we applied Bonferroni corrections.

Participants reported all methods to be rather *easy to use* (Mn_{palm} = 4.6, Mn_{key} = 3.6, Mn_{gait} = 3.6), with no statistically significant differences found (χ^2(2)=4.2, p=.12). Similarly, the *difficult to use* question received low ratings (Mn_{key} = 1.4, Mn_{gait} = 1.33, Mn_{palm} = 1.2) with no significant differences (χ^2(2)=2.3, p=.31).

As expected, users had more *knowledge* about the key than the other methods (χ^2(2)=26.8, p=.001). The key was already known by all participants (Mn_{key} = 5), while the other methods were rarely known (Mn_{gait} = 1.6, Mn_{palm} = 1.47). Hence, users had significantly more knowledge about the key than the palm vein scanner (Z=-3.5, p=.001) and the gait-based method (Z=-3.4, p=.002).

Regarding *comfort* (χ^2(2)=26.8, p=.001), users perceived the palm vein scan as the most comfortable method (Mn_{palm} = 4.2, Mn_{gait} = 3.6, Mn_{key} = 2.6), which is supported by a statistically significant difference between the key and the palm vein scanner (Z=-2.83, p=.005).

For the category *speed* (χ^2(2)=9.8, p=.007), the palm vein scanner was perceived as the fastest method (Mn = 4). In comparison, participants perceived the gait-based authentication (Mn = 3.4) and the key (Mn = 2.4) slower. A statistically significant difference was found between the vein scanner and the key (Z=-2.83, p=.005).

All methods received low scores (Mn_{key} = 1.93, kinect mn= 1.47, Mn_{palm} = 1.13) for being *cumbersome* (χ^2(2)=9.8, p=.007). In addition, the analysis reported a significant difference

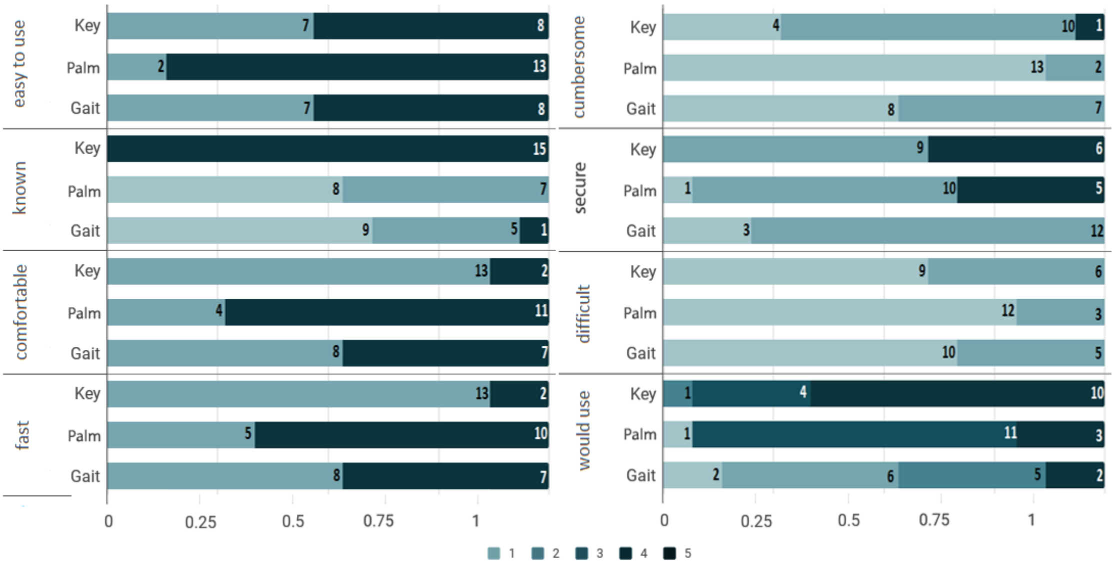


Figure 4.7: Participants' answers to our Likert statements on a scale from 1 (strongly disagree) to 5 (strongly agree).

between the palm vein scanner and the key (Z=-2.67, p=.008), indicating that users found the key slightly more cumbersome.

The results on *security* indicate that users perceive the gait-based method as less secure (Mn = 1.8), whereas the other options were rated as moderately secure (Mn_{palm} = 2.93, Mn_{key} = 3.2). The analysis of security ($\chi^2(2)$=8.1, p=.018) revealed that users found the key significantly more secure than gait-based access (Z=-2.4, p=.016).

In real life scenarios ($\chi^2(2)$=17.4, p=.001), users *would use* the key and palm vein scanner (Mn_{key} = 4.6, Mn_{palm} = 4) rather than the gait-based authentication (Mn = 2.6). This is supported by the statistical analysis as users rated the gait method significantly lower than the key (Z=-3.4, p=.001) and the vein scanner (Z=-2.8, p=.004).

4.3.3 User Feedback

We concluded the study sessions with semi-structured open interviews. The answers generally align with the quantitative data results. Here we give an overview.

Comfort of Use & Reliability

The hand vein scanner was believed to be fast, comfortable, and – similarly to the key – moderately secure. The key was also considered moderately fast, but slower than the biometric conditions. Some considered it to be the most cumbersome, feeling burdened by the mechanical task of unlocking the door. The gait-based authentication had mixed opinions in terms of being comfortable, but it was perceived fast when the authentication worked. However, participants expected the door to open by itself, when they were approaching it (like an automatic sliding door). Participants expressed concerns about the need to always walk in the same fashion to authenticate via their gait. Some felt it was draining to forcefully walk the same way. Others found the method very comfortable to use. Having additional luggage with them did not seem to be a serious concern for the participants. They stated they would put it to the side and authenticate as usual.

Possession & Control

Participants mentioned that the physical property of a key gave them a feeling of security, as well as enabled the option to duplicate and borrow it from others. Using biometric authentication as the sole way of entering a room on the other hand seemed to be intriguing since they would not have to worry about forgetting or losing a physical key. When asked what would happen if they were not able to authenticate with biometrics, most participants suggested calling the company that provided the door lock.

Setup Effort

Our biometric authentication mechanisms (i.e., conditions (2) and (3)) required a setup process (mocking the process of training a biometric system). The palm vein scanner was familiar to the participants since most of them compared it to fingerprint scanners used on phones. Hence, the registration process was likewise familiar. Gait-based recognition felt more cumbersome to set up and the participants were worried about false positives and false negatives. The registration for biometrics was considered inconvenient by one participant.

Perception of Security

Participants stated the key to be reliable and secure, though a physical token can be lost or forgotten. In contrast, participants appreciated that biometrics cannot be lost but were also worried about exposed data. The gait-based authentication was criticized for being too inconsistent and insecure. Participants were worried about imitators. Some were also concerned about the security of our unlocking mechanisms in general, as locks can be "picked" or technology can be "hacked".

Fallback Solutions

At the end of the interview, we asked, what options participants would consider if the door could not be unlocked. For the key, every participant had an idea of what to do as they could call a lock and key service or use a spare key. In comparison, not everyone could name a backup plan for the biometric techniques. Only some reported they would call a support hotline of the manufacturer of the authentication system. Three participants would allow the manufacturer to remotely open the door if they were locked out. In terms of combinations of the systems, participants suggested that a physical key could be used as a backup option or that more than two systems could be used in sequence to enhance security.

Wizard-of-Oz

When we asked the participants if they had noticed anything strange, two of them stated that they were unsure about the system properly working. However, the participants denied that this had any effect on their answers to the questionnaire. This was the final question we asked before revealing that it was a Wizard-of-Oz study. We observed that the participants were focused on the acoustic and visual feedback and did not try to open the door when no success signal was given for the biometric mechanisms.

4.4 Discussion

We conducted our study in a Wizard-of-Oz setting to assess how users would perceive the usability of three unlocking methods that represent physical, biometric, and behavioral authentication. Our focus is on real-world physical door access, which is underexplored in the literature but important considering the number of doors people access every day. Thus, the main contribution is a better understanding of which authentication users prefer, and why. In particular, we summarize our findings on user perception in the following key points:

4.4.1 Users Prefer Biometrics but Keep the Key

The biometric hand vein scanner was the premiere choice for most participants, as it is faster, more comfortable, and easier to use compared to the other authentication methods. However, the key was rated higher than the hand vein scanner with regards to which technology participants would actually use. This might be explained by keys offering a moderately secure, fast, and comfortable authentication, while also being affordable and known to everyone. Participants knew how to react if the key was lost and a fallback was needed. In general, the participants valued the possession of a physical object and the option of sharing it. This is not possible for the tested biometric authentication systems.

4.4.2 Recovery Effort Hampers Gait-Based Authentication

Both, the hand vein scanner and the gait-based condition, were criticized for being inconsistent. We assume that the forced failed authentication in our study design led to this observation. Notably, the use of gait was perceived as most inconsistent. If authentication fails, the act of returning to the starting position to walk again compared to the repeated scan of the veins takes a lot more time and effort. It could be helpful to consider alternatives such as a key, when the gait-based authentication fails to work on the first try, as repeating the measurement disrupts a seamless experience. We propose to further investigate the effect of forced fail conditions and error recovery efforts as influencing factors on user perception.

4.4.3 Imitation Concerns of Gait

Gait-based authentication was perceived as faster and more comfortable to use than the key but was still ranked last. The reason for this might be the concern of imitators and changes in walking behavior. Participants were worried, that attackers could mimic their gait to unlock the door, confirming observations of Yampolskiy and Govindaraju [296]. Also, changes in behavior, such as being injured, could reduce the chance of success dramatically. We propose to further investigate the actual risk from such impersonator attacks as well as adequate fallback options for changes in the user's walking behavior.

4.4.4 Limitations

Our study comes with some limitations. First, we had a relatively small sample of 15 participants. While this amount resulted in statistically significant results on user perception, repeating the study with more participants should provide more reliable data. Further, it is possible that different ages and backgrounds may have an impact on the opinion about authentication systems, demanding further study.

In addition, while we can asses participants' general attitude towards the tested authentication systems, this is but an approximation to how they would react to actual implementations. We carefully crafted our study setup to foster the impression of a real system and increase believability. More studies are needed to cover the range between research prototypes and, in the future, novel authentication methods to gain more confidence in how door-unlocking mechanisms should be designed.

4.5 Implications

Our Wizard-of-Oz study showed that users are willing to consider biometric mechanisms for seamless authentication at doors for their ease of use. However, they still preferred a physical key for actual use for the agency it gave them and their knowledge on how to cope with potential issues. Participants were concerned about changes in physiology or behavior impacting their biometric systems and the opportunity for attackers to gain access. The use of a visual indicator for the success or failure of the biometric system was effective.

Overall, this chapter reinforces the *concerns about changes* impacting biometrics we found in Chapter 3. It also highlights that *ease of use* is an important aspect for considering biometrics though, *being familiar* with the authentication and *having agency* was more important to participants for practical use. We found the *recovery effort* from a failed authentication attempt an important aspect that should also be considered for biometric interfaces.

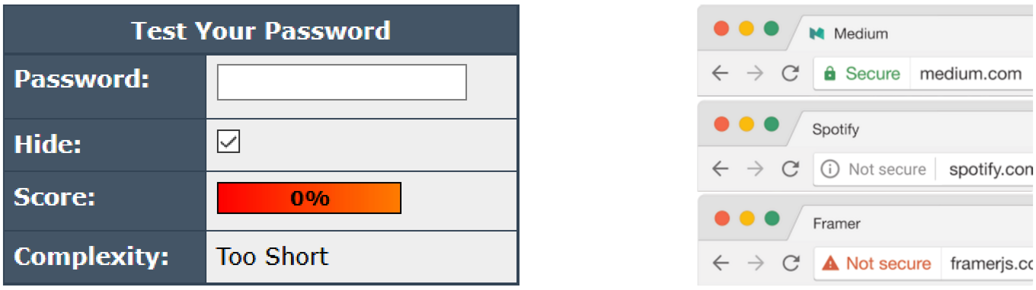
In the context of designing biometric interfaces, those findings stress the importance of supporting user agency over biometric methods (see Part IV). This chapter also shows, that the choice for an authentication mechanism is complex and takes many factors like security, ease of use, recovery effort, and familiarity into account. Biometric interfaces thus should support users in gauging those factors to allow them to make an informed decision for or against a biometric mechanism.

5 Design Opportunities for Biometric Interfaces

This chapter is based on the following publication:
Lukas Mecke, Sarah Prange, Daniel Buschek, and Florian Alt. 2018. *A Design Space for Security Indicators for Behavioural Biometrics on Mobile Touchscreen Devices.* In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18) [184]

In the previous chapters, we assessed user needs and preferences towards biometrics and their interaction with them. In this chapter, we switch the perspective to explore a design aspect of creating biometric interfaces. Related work on how to design biometric interfaces is as of now still largely missing, which motivated us to take inspiration from existing work on security indicators for passwords (see Fig. 5.1).

Similar to the aim of this thesis such indicators support user literacy and agency by giving them feedback on the strength of their authentication secrets, giving context on potential weaknesses, and supporting users in taking agency and improving their security through



Test Your Password	
Password:	
Hide:	☑
Score:	0%
Complexity:	Too Short

(a) Indicator of password strength using visual and textual feedback[a]

(b) Indication of connection security proposed for the Chrome web browser by Felt et al. [92]

[a] https://passwordmeter.com/, last accessed October 16, 2024

Figure 5.1: Examples for classical security indicator approaches. The design of a security indicator for behavioral biometrics would possibly have to differ from that. In contrast to a password, the security of behavioral biometrics depends on the individual person.

meaningful additions to their passwords. Designing for biometrics introduces new challenges: personal and external factors can impact their performance, the features used may not be alterable by the user and the decision process is not a binary matchmaking but is based on complex machine-learning and pattern-matching algorithms. Designers of biometric interfaces thus have to consider many factors, making an investigation and formalization of design considerations a worthwhile endeavor.

To do so, we conducted a focus group with HCI and usable security experts to uncover design challenges when developing interfaces for biometric systems. To make the topic more graspable we introduced security indicators for passwords as context to explore differences and necessary extensions. We also prompted participants to specifically think about indicators for behavioral biometrics to capture the continuous aspect they introduce to authentication.

Based on this focus group we derived a design space for security indicators for behavioral biometrics with the aim to support the design of indicators that facilitate users' decision-making, awareness, and understanding, as well as increase the transparency of biometric systems. We illustrate with three examples, how our design space can be used to generate new ideas for biometric interfaces and consider relevant aspects. While the focus of this work was more narrow (i.e. limited to behavioral biometrics and security indicators), we later discuss if and how this design space can be extended to biometric interfaces in general.

> In this chapter, we contribute 1) a design space for security indicators for behavioral biometrics on mobile touch devices, which we derived from a focus group with experts and the literature. Further, we 2) provide a set of examples of how this design space could be applied in future work for the development of biometric interfaces.

5.1 Background and Research Approach

Here we give an overview of security indicators used for passwords that we used as inspiration for our work. We then introduce the derived approach for finding a design space for such indicators for (behavioral) biometrics.

5.1.1 Security Indicators for Passwords

Related work on visual indicators of password strength shows that users have misconceptions about what constitutes a strong password [239, 271]. The same trend was shown for behavioral biometrics by Ballard et al. [23], using handwriting recognition. Here, forgery was both more successful and harder to detect than users had expected.

Password meters address this by assessing and displaying a password's resilience against attacks (Fig. 5.1a). They can convince users to choose stronger passwords [165, 272]. Giving additional information and detailed, potentially sensitive feedback about the current strength can help users improve their passwords [270]. Related work also showed that user awareness of password strength can be increased [239]. On the other hand, due to inconsistencies in current password strength estimations, more transparency might be needed to reestablish users' trust in security indicators [69].

Existing work on security indication mainly covers passwords and websites [92] (Fig. 5.1b). To our knowledge, similar investigations for biometrics are still missing.

5.1.2 Deriving the Design Space

In contrast to passwords, the security of biometrics depends on the individual person; the same settings may lead to different security levels for different users. Thus, given the potential impact and issues, adapting the design of security indicators to biometrics is both relevant and challenging.

To identify a design space for (behavioral) biometrics we conducted a focus group with eight experts from, but not limited to, the fields of password meters, machine learning, user behavior prediction, and context-aware technology. Participants were introduced to the concepts of security indicators and behavioral biometrics. Subsequently, they were asked to think of how a security indicator for behavioral biometrics would have to differ from classical approaches and what possible benefits they could have both for users and providers. Based on those results participants were asked to come up with concrete ideas and cluster those, filling missing design dimensions as needed.

Our focus group discussions revealed several design dimensions. We post-hoc clustered those dimensions, taking into account the related work, resulting in an additional layer of abstraction with three categories.

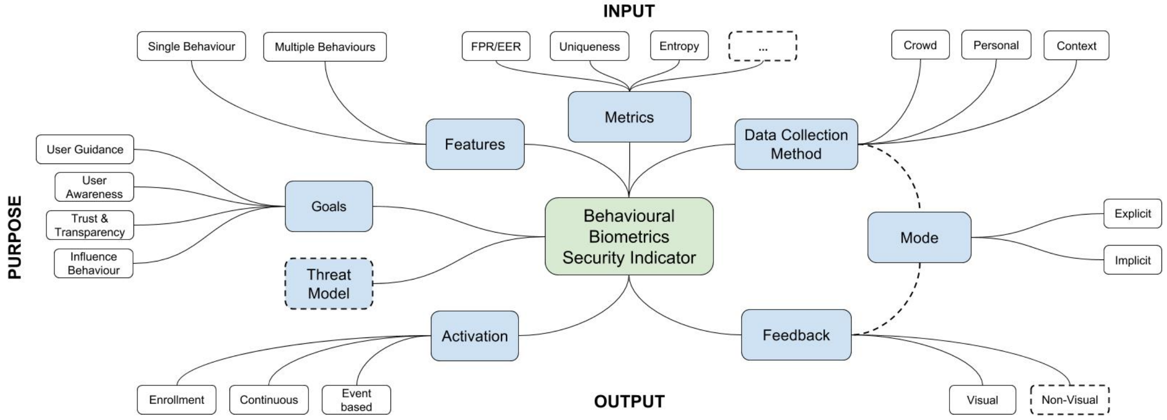


Figure 5.2: Our proposed design space consists of eight main dimensions. We classify dimensions into three categories: purpose-, input- and output-related. Dimensions added based on the literature review are indicated with dashed borders.

5.2 Design Space

Here, we give an overview of the resulting design space. Categories and dimensions are depicted in Figure 5.2 and described in detail below:

5.2.1 Purpose

The category that should be considered first is the purpose of the indicator in question. This includes two dimensions:

Goals: Potential goals designers might try to achieve include, but are not limited to:

1. *User Guidance:* By providing (personalized) security information, indicators may guide a user, for example, when choosing (a combination of) biometrics to select a more secure/unique behavioral feature.

2. *User Awareness*: By communicating levels of security and the system's internal reasoning, indicators may aim to increase user awareness (e.g., risks and benefits of the current settings w.r.t. behavioral biometrics).

3. *Trust & Transparency*: Taking the previous goal a step further, indicators may be designed to increase transparency and trust in the system [69] (i.e. by explaining how and why security assessments are made). This might in turn improve acceptance of the underlying behavioural biometric system.

4. *Influence Behaviour*: Similar to password meters, a behavioral biometrics indicator might aim to improve security by altering the user's behavior in a way that generates more unique, and hence secure, user behavior.

Threat Model: When designing a security indicator we need to think about the threat model to which the indicated security is related. A comprehensive list of possible attacks on biometric systems can be found in [219]. A related aspect, for example, is the distinction between user verification and identification – different threats exist for both of them and indicator designs might differ as well.

5.2.2 Input

Several aspects of the underlying biometric system may be considered as input dimensions to inform the design of a corresponding indicator.

Features: The first choice is which biometric(s) should actually be used. This can be either a *single behavior* or a combination of multiple behavioral traits (e.g., typing speed and pressure). In the latter case, further decisions have to be made with respect to how those traits should be combined (*feature fusion*). This is relevant as indicators might be designed to inform the user about the currently considered combination of behaviors.

Metrics: There are several metrics to estimate the security of a biometric approach, such as system error rates (e.g., false positive rate, equal error rate), "uniqueness" of behavior, and its entropy (similar to entropy as a measure of password strength). Beyond these metrics from our focus group, Rudolph and Schwarz [224] provide an extensive list of indicator metrics.

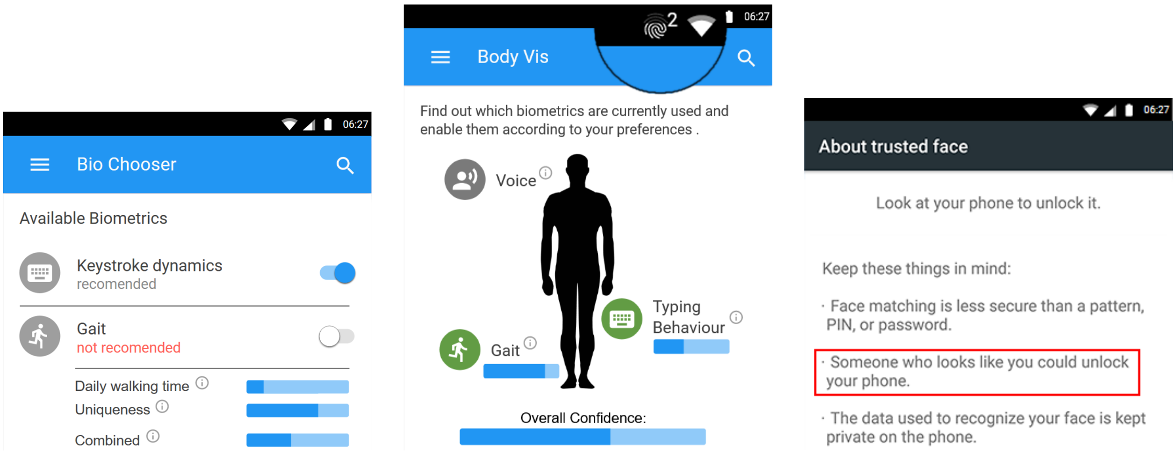
Data Collection Method: Any metric to indicate the security of a behavioral biometric system requires data on which to operate. This data can be either provided by the respective *user*, acquired from the user's *context*, or collected from a *crowd*. The first option is likely the most common one (e.g., data from enrollment or past use). However, the use of crowd data can enable instant feedback without a "cold start", and context information enables adaptive estimation.

5.2.3 Output

Based on the input and purpose of the indicator there are several ways to design the output.

Feedback: Similar to password strength, *visual* feedback can be used to represent the assessed security (Fig. 5.1). One possible option is textual feedback given in the form of scores ("90%"), assessments ("strong"), or metaphors ("One in ten strangers might get access to your data using this behavior for authentication"). Other representations might be diagrams or be abstract. Additionally, feedback might be given in a *non-visual* way, e.g., auditory or haptic.

Activation: There are several points in time when a security indicator might appear. We distinguish *enrollment* (i.e. only once at the beginning), *continuous* (may also be periodical), and *event based* (i.e., as a reaction to context changes, e.g. upon launching an app).



Available Biometrics

Keystroke dynamics
recomended

Gait
not recomended

Daily walking time
Uniqueness
Combined

Find out which biometrics are currently used and enable them according to your preferences .

Look at your phone to unlock it.

Keep these things in mind:

- Face matching is less secure than a pattern, PIN, or password.
- Someone who looks like you could unlock your phone.
- The data used to recognize your face is kept private on the phone.

(a) "Bio Chooser": A possible design for a security indicator to support users when deciding which behavioral biometrics to use, avoiding unnecessary collection of data.

(b) "Body Vis": A design giving a continuous indication of the currently used biometrics and the system's confidence (both in the status bar and in detail).

(c) An example for a state-of-the-art (static) biometric enrollment screen (Android Trusted Face) that might be improved by adding dynamic security indication.

Figure 5.3: Examples of designs for biometric interfaces informed through our design space (5.3a and 5.3b) and state of the art (5.3c).

Mode: We distinguish *implicit* and *explicit* modes for two parts of the design space: 1) Data collection can happen either implicitly (e.g., background logging) or explicitly (e.g., enrollment procedure); 2) the indicator itself can be either implicit (e.g., an informative icon) or explicit (e.g., demanding a user action). In the case of an icon, further considerations might be needed to ensure that users notice and understand [92, 291] the information. Different modes may be chosen for data sources and feedback.

5.3 Using the Design Space

We now illustrate the use of the design space with a set of examples inspired by ideas from the focus group. They cover different design choices along the identified dimensions.

5.3.1 Example 1: “Bio Chooser” – Decision Support System

This indicator supports setup and *enrollment* of a multi-biometric system. Given multiple available biometrics, it indicates how each of them affects security, based on the individual user’s behavior (e.g., *personal* data from past usage or an enrollment sample). It could also include behavior frequencies to “weight” the usefulness of certain biometrics (e.g., keystroke biometrics are more useful if the user types a lot), as well as common *contexts* (e.g., gait recognition might be less useful if user commutes via train). The indicator could *visually* present security implications like expected error rates. In this way, it aims to increase *awareness* and *guide* users' choice. (compare Figure 5.3a).

5.3.2 Example 2: "Crowd Radar" – Local Crowd-based Indicator

This indicator compares (local) *crowd* data with the user's own behavior. It indicates the *uniqueness* of the user's behavior in the vicinity and context. *Visual feedback* is presented as text: "x people in your vicinity have very similar behaviour". Beyond *awareness*, this could be extended towards *guidance*, for example, with recommendations on which (combinations of) biometrics to activate in this (crowd-)context.

5.3.3 Example 3: "Body Vis" – Visualizing Activated Biometrics

This indicator (Fig. 5.3b) *continuously* displays the parts of the body that are currently tracked as a *personal data source* for continuous implicit authentication. This aims to facilitate *awareness* and *transparency*. The indicator might map confidence to color or brightness to prepare the user for possible explicit (re-)authentications.

5.4 Discussion

Based on our focus group and literature research, we defined a comprehensive design space. Here we discuss opportunities for using and extending it.

5.4.1 Extending the Design Space

Indicating only security may not be enough. From our investigation in the previous chapters we know, that usability also has a strong impact on user preferences regarding the use of biometrics and thus propose to extend the design space to account for that. For example, an indicator might display the amount of explicit authentication time saved with certain biometrics settings, estimate the number of necessary re-authentications, or consider user preferences (e.g., speech input vs. typing). Overall, such indicators could support users in finding individually suitable usability/security trade-offs.

Similarly, the focus of the design space could be extended to consider physiological biometrics as well. Note, how this would mainly add more options to the levels of the specific dimensions (e.g. the addition of other features and metrics) but the established dimensions remain valid and applicable.

In its current form, the design space is focused on indicators, i.e. “passive” information presentation. For interfaces that require or should support user interaction, the shape of this interaction has to be considered as an additional factor.

5.4.2 Research Questions for Biometric Indicators

This work lays the foundation for future investigation of indicator designs regarding specific research questions. These might include, for example, questions from password meters, such as: How can we nudge users to choose more secure settings? Do indicators support understanding (i.e. can users better judge the security of behavioral biometrics systems after using indicators)? Do users understand how attackers could try to gain access to their data? Does the content of the security indicator facilitate new threat models?

5.4.3 Concrete Example: Integration into Android Smart Lock

Google’s face recognition system for Android devices displays a static text message to inform users about its security and related issues (e.g., ”Someone who looks like you could unlock your phone”, compare Fig. 5.3c). This is an example of a concrete integration opportunity: We could replace the static text with a dynamic security indicator designed by considering our space. For example, this indicator might compare newly registered users with existing ones in the database to indicate how likely an unintended or malicious unlock from a stranger actually is or predict the impact of changes in the user’s appearance on the performance of the system. Note, how this is also an example of extending our design space beyond its initial focus on behavioral features.

5.5 Implications

In this chapter, we formalized design considerations for security indicators for behavioral biometrics in the form of a design space. Despite the narrower focus of the expert focus group, this design space provides dimensions that are also useful to support the design of biometric interfaces in general. As such it provides a complementary design perspective to our exploration of user needs in the previous chapters.

We used this design space as inspiration for the design of multiple biometric interfaces throughout this thesis. Examples include the design of indicators to give insights into the state of a continuous system and warn users of upcoming re-authentications in Chapter 8 or our approach to present users with context information to allow them to choose an appropriate authentication mechanism in Chapter 7. Our investigation of how to communicate more meaningful measures of performance of a face recognition model in Chapter 6 was directly inspired by the example for improving the enrollment for face recognition given in this chapter.

III

Biometric Interfaces to Support User Literacy

PART III: BIOMETRIC INTERFACES TO SUPPORT USER LITERACY

Based on the previous part and the literature, we found that users had concerns about the impact of changes on their biometric systems and wished for control. In this part, we explore solutions to the first challenge and propose designs for biometric interfaces that give users insights into their biometric models and help them to anticipate their performance, the impact of contextual factors, and their behavior in general. As such, chapters in this part contribute to the goal of supporting users in better understanding their biometric systems and gaining *literacy*. We follow a breadth-first approach and explore interfaces for both different biometrics and different interactions.

- ❖ **Chapter 6** introduces a method we developed to assess the performance of decision-making models by comparing their output to human ratings. We show how this method can be used to gain insights into the performance of a face recognition model and propose how it could be used to give users a better understanding of how the system will perform for them personally.
- ❖ **Chapter 7** reports on a field study in which we assessed users' perception of a system leveraging context information to suggest an appropriate authentication method.
- ❖ **Chapter 8** proposes the use of indicators to give users insight into the current state of a continuous authentication system and help them anticipate upcoming re-authentication requests.

6

Exploring (personalized) Performance of Face Recognition using Generated Samples

This chapter is based on the following publication:
Lukas Mecke, Daniel Buschek, Uwe Gruenefeld, and Florian Alt. 2024. *Exploring the Lands Between: A Method for Finding Differences between AI-Decisions and Human Ratings through Generated Samples*. arXiv preprint arXiv:2409.12801 [180]

Enrollment to a biometric model is the first point when users get in touch with it and also where they make a decision to use a system or not. As such, this step presents a unique opportunity to support users in being able to make this decision deliberately and on an informed basis. In the first part of this thesis, we uncovered, that users consider many factors for such a decision, including the ease of use, being familiar with the type of authentication, and having agency. Another point brought up multiple times was the concern that changes in physiology or behavior would impact a biometric system, or more generally the wish to understand and anticipate the performance of a biometric model.

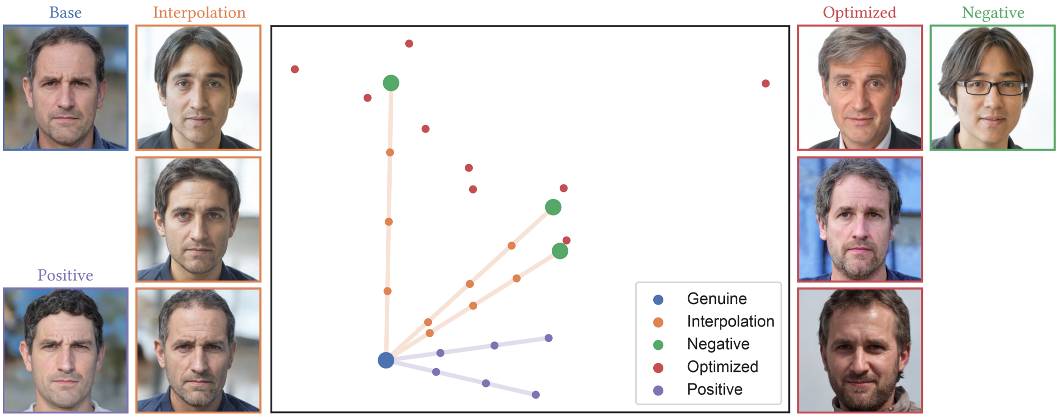


Figure 6.1: We propose to strategically sample alterations (positive, negative, interpolation, and optimized) of a base image from the latent space of a generative model (middle: projection of those samples on two dimensions of this latent space) to gain insights into another model that makes decisions on this data, here: face recognition. This involves comparing its ratings of those samples (i.e. recognition scores) to ratings by human raters. We explore this approach for the use case of examining face recognition models by strategically sampling images from the latent space of StyleGAN2 [154].

When looking at related work for biometrics, but also Artificial Intelligence (AI) decision-making models in general, we can see that such concerns may well be justified. Across many users and decisions, some sensor-based decision systems can enact hidden biases learned from the training data [48, 264]. For example, some facial recognition systems have been reported to work worse for people of colour [48, 298]. System performance can vary greatly depending on the user [294] or specific user groups [48]. Thus, overall performance metrics, like precision or recall of a system, may not be relevant and applicable to the individual. Additionally, such methods of testing decision-making models can fall short, as they rely on real-world data with clear-cut expected decisions (i.e. it is clear if two images were taken of the same person or belong to different identities). However, the model's performance and potential weaknesses may become more apparent when exploring the space between, where decisions are more difficult. This is also where decisions become more relevant for users, as they can give insights into error cases and help users and developers to understand what type of changes caused them.

Following this idea, this chapter presents a method to explore the decision boundaries of (black-box) decision-making systems using artificially generated inputs. We propose three main areas of inputs: samples that are expected to lead to a positive decision, samples that are expected to lead to a negative decision, and samples that explore the space in between where the decision may be unclear. Generating artificial samples has two main advantages: no real-world samples are needed (beyond training the generative model), and we can explore cases that would not be possible in real life to better understand the model (e.g., a voice that is a mix of two speakers). By presenting these samples both to human raters and decision models, it is then possible to find areas of agreement and disagreement. With respect to biometric interfaces, this method can provide more fine grained insights into a model that can be presented to users. Later in this chapter, we also illustrate a possible extension to our method to illustrate the personalized performance of biometrics and discuss what an interface using this information could look like.

To assess the value of our method to gain insight into biometric models, we evaluate it by exploring face recognition with color images as input (e.g., face unlock with a phone camera). This is well suited for our approach (humans themselves are excellent at recognizing faces [79]) and of high societal relevance and impact[1] even beyond this thesis. We generate meaningful alterations (see Figure 6.1) for 40 base images using StyleGAN2 [154] and collected a dataset of perceived similarity and identity of the presented image pairs in an online comparison task with 100 participants.

We found interesting mismatches between the analyzed face recognition model and human raters. The model rated images of children more similar and semantic changes (like the addition of glasses) less similar than humans would. Our optimized samples were successful in fooling the decision model while mostly being perceived as different by humans. Latent distance and perceptual distance were good predictors for perceived identity except for those optimized samples.

We conclude with a discussion of how our approach could be used to explore the performance of AI models in different contexts and for different user groups and how it can be used for biometric interfaces.

> In this chapter we 1) propose a method using generated samples to understand decision-making systems by presenting them with selected samples and comparing the provided decisions to human ratings; we 2) applied our method to explore the performance of a face recognition model, 3) provide the generated dataset, and 4) contribute an initial analysis of this data and a discussion of how the method can be used as input for biometric interfaces and other applications.

6.1 Background and Research Approach

In this section, we give background on approaches for generating artificial content, previous uses of such approaches as well as the topic of adversarial samples that follow a similar idea to our proposed method. We conclude with an overview of our approach.

In the last years, approaches for artificial sample generation have gained public awareness through methods like ChatGPT for text production or Stable Diffusion [222] and DALL-E [212] for generating images from text prompts. Beyond those, there exists a plethora of other approaches for generating content, including autoencoders [162, 205, 273],

[1] ACM statement on the topic: `https://www.acm.org/media-center/2022/february/tpc-tech-brief-facial-recognition`, last accessed October 16, 2024

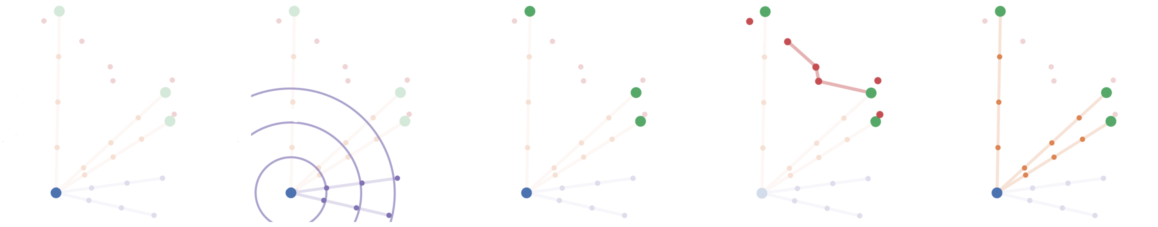
Normalizing Flows [161, 216], and Generative Adversarial Networks [87, 113, 148, 154](GANs), to name just a few. The main principle behind all those models is to learn a representation of the distribution of their training data (latent space) that can then be used to generate new and altered samples. In particular for GANs, there exist many approaches showing how their latent space can be used to generate semantic edits (for example, making a person older) [137, 246, 283], find meaningful dimensions [123], and mixing samples both on a style and content level [55, 154]. Beyond artistic purposes, they can also be used to generate synthetic data for training and evaluating machine learning models [59, 153] and finding biases [19, 75]. A related approach is generating or finding so-called adversarial examples [16, 114, 127]. Those are characterized as small changes to the input of a neural network that are not (or only with difficulty) perceptible by humans but cause the model to flip its decision or predict a different class.

With our approach, we utilize the power of generative models to generate challenging samples for a decision-making model. We conceptually follow a similar approach as is used for adversarial samples: we propose to generate inputs to a decision-making model that lead to unexpected results. However, we are explicitly interested in cases where changes are perceptible, but their impact does not align with human expectations. Thus, we introduce human raters as a comparison to the model's decision to better understand where perception aligns and where the model acts unexpectedly.

In the next section, we first describe our method in detail before exploring its utility by applying it to a face recognition model. We conclude with a discussion of our insights and directions for using our methods for biometric interfaces. We also discuss prerequisites for using the method in other contexts and for different target groups.

6.2 Leveraging generated Samples to Explore Decision-making Models

We suggest comparing the outputs of a decision-making model on strategically sampled inputs to answers by human raters to better understand the model. Note, how both the human and the generated samples are needed. Without human ratings, we don't gain insights into mismatches in perception and by using only real-world data we cannot gain access to the space between clear-cut decisions where we expect those mismatches to be found. Here, we use a generative model to produce samples inspired by the outputs of a classical classification task: true and false positives as well as their negative counterparts. In this section, we give more details on the steps of this process. We illustrate and explore our method for the concrete case of face recognition as an example of a decision-making model. Any empirical claims are limited to this use case. However, we discuss extensions and applications of our method in Section 6.5.



(a) Genuine (b) Positive (c) Negative (d) Optimized (e) Interpolation

Figure 6.2: Illustration of how our proposed samples are generated (simplified illustration of a latent space). Positive samples are generated by sampling points close the the genuine sample. Negatives are random other points in the latent space. Optimized samples are generated starting at a negative sample and using an optimization function to find samples that are more similar to the genuine sample. Interpolation samples are found as steps on the latent path between the genuine sample and negative samples.

6.2.1 Generator

The core component of our approach is a generative model to provide samples that can serve as input to the model we wish to test. For the case of evaluating face recognition, we propose the use of Generative Adversarial Networks (GANs) to generate such samples as they have been shown very capable of generating realistic face images [153, 154] and their continuous latent space can be leveraged for targeted manipulations [246, 283]. They can thus be used to produce alterations of a given starting point as well as samples *between* existing real-world data points for which no ground-truth "true" labeling exists. Note, that while we suggest the use of GANs for face images, other generative approaches (see Section 6.1) are possible as long as they can produce targeted manipulations. In some cases, no model may be required, for example, if the decision-making model solely relies on numerical inputs.

6.2.2 Samples

We now illustrate the sets of samples we propose to generate and the rationale behind choosing them. Each sample is always generated in relation to a *base* (i.e. the starting point in the latent space) that will later be used for the comparison (see Section 6.2.3). For the example of face recognition, this would be the face image to test the decision model on. Figure 6.1 illustrates examples for each of the sample types and Figure 6.2 shows in more detail how they relate to the base image. Note, that the generated samples are independent of the decision-making model to be tested. As such, our approach is also suited to explore black-box models, as long as they can be queried.

Genuine Samples

Genuine samples (Figure 6.2a) are identical to the base image and, thus, what would be called a true positive in traditional classification tasks, i.e. they preserve the identity of the base image. Note, that this is the only possible true positive as it is not necessarily clear to which identity a sample should be attributed. This is the case because, in contrast to the real world, we can continuously sample images from the latent space between two identities. We include this sample both as a baseline for participants to calibrate their rating of similarity against and as a check to make sure participants pay attention while rating the image pairs.

Positive Samples

We propose positive samples (Figure 6.2b) as slight modifications to the base. We find those by sampling a (random) direction in the latent space and taking a small step away from the base in that direction. The assumption behind this approach is that, given a locally stable latent space, this should produce minor alterations to the input and preserve the model's decision on the base (for example, identity for our case of face recognition).

Negative Samples

We choose random samples from the latent space as negative samples (Figure 6.2c), i.e. samples that we expect to be attributed to a different identity as the base. This is based on the assumption that the latent space is big enough that randomly generating a sample similar to the base is unlikely. For practical reasons, we propose the use of other base samples as negative samples. This way bases can serve as both genuine and negative samples, and their function is only defined relative to their respective base.

Optimized Samples

We propose to use each negative sample as a starting point for a black box optimization algorithm to find samples that maximize the decision-making model's target function (for example, similarity). In contrast to the other samples, this step requires either access to the decision-making model itself or a similar model. In our results (see Section 6.4.3) we explore if two different models can be used for this (i.e. one to be explored, one for generation). Given the different starting points, we assume that the generated samples (Figure 6.2d) should represent local maxima in the latent space (instead of finding the original base sample) and may thus very well not be similar to a human observer even though they are similar according to the optimization function. As such optimized samples fulfill the role of potential false positives.

Interpolation Samples

To better understand where the model's decision between two (base) samples changes, we introduce interpolation samples (Figure 6.2e). Those are generated by following the latent vector between the base and each negative sample. Generating candidates for false negatives through optimization would require human ratings as a target function. As those are not available at generation time, interpolation samples are our best attempt at provoking this type of misclassification.

6.2.3 (Human) Raters

The final component of our approach is having the generated samples rated by humans as a baseline to compare against the model's decisions (remember, that for the samples we propose, no ground truth is available; so this step is necessary). This is based on the premise, that the decision-making model is supposed to decide similarly to a human or at least in a way that aligns with human intuition. To reflect this we propose that human raters both judge the target function of the model as well as the decision they would expect. For the example of a face recognition model, this maps to a similarity score for the presented faces and the decision whether two images show the same person. Depending on the use case it may be possible to have the samples rated by a different model instead of humans if only differences between models are explored with no focus on whether they adhere to human perception.

6.3 Experiments

We now illustrate how we implemented the approach described above for the example of face recognition. We explain how we generated the samples and implemented the comparison task for human raters. Note, that latent spaces allow for a vast amount of potential comparisons and our choice can only capture a fraction of them. The choices of both samples and parameters presented here reflect our best attempt at striking a balance between many potential comparisons, sufficiently many samples to observe trends, and a number of comparisons that can realistically be made by participants.

6.3.1 Dataset Generation

We generate all samples using StyleGAN2 [154]. Base images (and consequently negative and genuine images) are sampled as random seeds from the latent space. The number of samples scales with the number of base images, as e.g. interpolation samples are generated between all available bases. Thus, we decided on a batched approach with 4 base images per batch. For each base, we generate one genuine sample (i.e. a copy). We choose two random directions and generate three positive samples (at distances 0.2, 0.4, and 0.6) for each of them (6 total). As negative samples, we include the remaining three base images. Furthermore, we generate interpolation samples at 25%, 50%, and 75% of the distance between the base and each negative sample (9 total). Finally, we use the Covariance Matrix Adaptation Evolution Strategy (short CMA-ES)[2] as a black box optimizer to find optimized samples. This approach empirically finds a gradient by sampling points around a given starting location and optimizing a given score function. We encoded images as a 512-dimensional vector of their latent representation and used a Python face recognition algorithm[3] as the optimization function. We ran 100 iterations with a truncation factor of 0.3 (i.e. we stopped optimization when reaching this distance score) using $\sigma = 1$. We ran one optimization starting from each of the negative samples and chose the first results that achieved a recognition distance below 0.3, 0.4, and 0.5 respectively[4] as optimized samples (9 in total). In case the optimization did not reach this score, the best sample was chosen. For reference: the suggested default recognition distance for deciding on identity in the used Python library is 0.6, so all of those samples would be accepted. Overall, this approach yielded 112 samples in each batch (28 samples for each base image).

6.3.2 Procedure

To collect human ratings on our samples, we designed an online survey. First, participants would be informed about the purpose of the study and had to consent to their data being collected and analyzed. Next, we collected basic demographic data before participants got to the main task. Here, participants were repeatedly presented with an image pair where one was always a base image and the other was one of the samples described in Section 6.3.1. Each participant rated 112 image pairs in randomized order. For each image pair, we asked participants for their perception of the similarity of the two faces and a binary decision if they believed they showed the same person (see Figure 6.3 for an example of such a choice). We concluded by asking participants for their strategy in rating both similarity and identity.

6.3.3 Participants and Recruitment

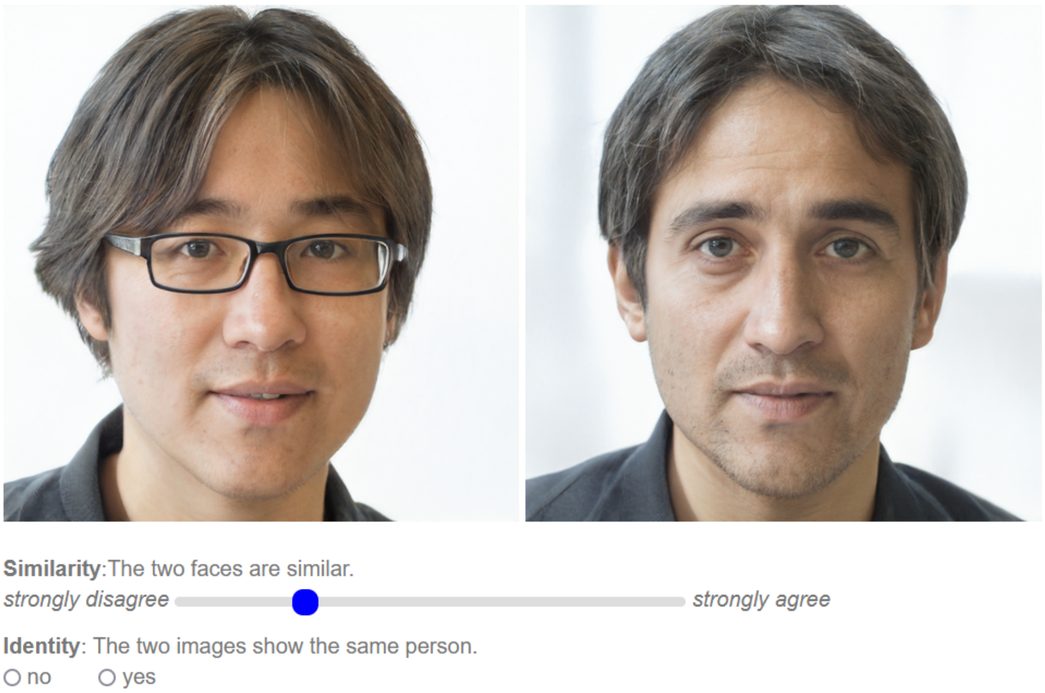
We recruited 100 participants (50 female, 48 male, 1 non-binary, 1 no answer) with a mean age of 27.6 (SD=7.8, range: 19-61) using Prolific[5]. People from all continents contributed to our dataset. The study took about 20 minutes and was compensated with £2.25. The study was approved by our institute's ethics commission under Nr. EK-MIS-2023-204.

[2] CMA-ES: `https://pypi.org/project/cma/`, last accessed October 16, 2024

[3] Face Recognition: `https://pypi.org/project/face-recognition/`, last accessed October 16, 2024

[4] Due to an error in the implementation the images chosen for distances 0.4 and 0.5 included samples with worse ratings

[5] Prolific: `https://prolific.com`, last accessed October 16, 2024



Similarity:The two faces are similar.

strongly disagree *strongly agree*

Identity: The two images show the same person.

○ no ○ yes

Figure 6.3: Screenshot of the main task in our online study. Participants were presented with two images (a base image on the left and a sample generated with our approach on the right) and were asked to rate their similarity and if the images showed the same person.

6.4 Results

In this section we describe the collected dataset, verify the effectiveness of our sampling methods, and demonstrate how our approach can be used to gain insights into the decision-making model.

6.4.1 Dataset Overview

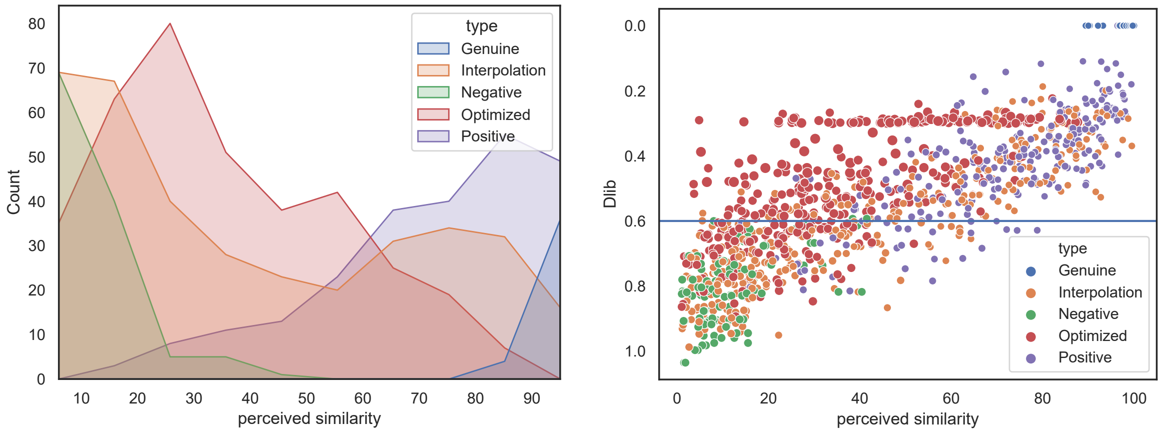
Our dataset consists of 1,120 image pairs (10 batches of 112 image pairs each) that were rated by 10 participants each, resulting in a total of 11,200 ratings of similarity and identity. In addition, we post-hoc calculated distance scores for four common state-of-the-art face recognition algorithms using the DeepFace library by Serengil and Ozpinar [243], as well as latent distance (based on the distance of embeddings in the StyleGAN2 latent space) and perceptual distance (lpips) [306]. For the sake of brevity, we only compare against Dlib face distance when making comparisons to a face recognition model (unless otherwise stated). An overview of our dataset grouped by type of sample is given in Table 6.1.

	perceived		**face recognition models**				**distance metrics**	
Sample type	sim ↑	*id* ↑	Dlib ↓	VGG ↓	FN ↓	OF ↓	lpips ↓	latent ↓
Genuine	97.75	*0.98*	0.00 *(1.00)*	0.00 *(1.00)*	0.00 *(1.00)*	0.00 *(1.00)*	0.00	0.00
Interpolation	39.97	*0.25*	0.61 *(0.43)*	0.92 *(0.40)*	1.12 *(0.32)*	0.85 *(0.17)*	0.52	67.87
– dist 25%	73.30	*0.59*	0.42 *(0.88)*	0.76 *(0.66)*	0.90 *(0.64)*	0.74 *(0.33)*	0.37	33.94
– dist 50%	31.35	*0.13*	0.67 *(0.35)*	0.96 *(0.33)*	1.18 *(0.23)*	0.86 *(0.10)*	0.56	67.87
– dist 75%	15.26	*0.02*	0.76 *(0.07)*	1.05 *(0.22)*	1.29 *(0.07)*	0.94 *(0.07)*	0.64	101.81
Negative	11.34	*0.02*	0.80 *(0.02)*	1.08 *(0.18)*	1.33 *(0.02)*	0.93 *(0.07)*	0.67	135.75
Optimized	35.36	*0.16*	0.50 *(0.68)*	0.89 *(0.48)*	1.08 *(0.41)*	0.86 *(0.12)*	0.60	227.16
– dist 0.3	51.67	*0.31*	0.31 *(0.99)*	0.76 *(0.68)*	0.92 *(0.70)*	0.80 *(0.22)*	0.55	279.99
– dist 0.4	32.49	*0.11*	0.53 *(0.80)*	0.88 *(0.53)*	1.07 *(0.46)*	0.85 *(0.11)*	0.61	217.45
– dist 0.5	21.91	*0.05*	0.66 *(0.25)*	1.01 *(0.22)*	1.26 *(0.07)*	0.95 *(0.03)*	0.64	184.05
Positive	72.32	*0.57*	0.41 *(0.86)*	0.70 *(0.67)*	0.87 *(0.65)*	0.68 *(0.40)*	0.34	38.11
– dist 0.2	88.46	*0.84*	0.29 *(0.99)*	0.55 *(0.79)*	0.68 *(0.81)*	0.57 *(0.61)*	0.24	19.06
– dist 0.4	70.97	*0.53*	0.43 *(0.88)*	0.72 *(0.69)*	0.89 *(0.65)*	0.69 *(0.36)*	0.36	38.11
– dist 0.5	57.52	*0.34*	0.52 *(0.72)*	0.84 *(0.54)*	1.03 *(0.50)*	0.78 *(0.24)*	0.43	57.17
all	44.42	*0.29*	0.53 *(0.58)*	0.85 *(0.48)*	1.04 *(0.41)*	0.80 *(0.22)*	0.51	117.54

Table 6.1: Descriptive overview of our dataset. Arrows indicate the direction of images being perceived as more similar. In brackets, we indicate the acceptance rates (0: all samples were rated as different identities, 1: all samples were rated as the same identity) of human raters and face recognition models (sim: rated similarity, VGG: VGG-Face, FN: FaceNet512, OF: OpenFace) based on their default recognition thresholds (Dlib: 0.6, VGG: 0.86, FN: 1.04, OF: 0.55).

6.4.2 Human Ratings with respect to Sample Types

As a first step, we validate the success of our approach. An overview of the distribution of similarity ratings by sample type is given in Figure 6.4a. As expected, genuine samples were rated as both similar (97.75) and the same person (0.98). The rating of interpolation samples strongly depended on the interpolation distance. Samples at 25% of the distance away from the genuine samples were rated highly similar and often still as the same person. Similarity decreased with more distance as did the identity rating, until at 75% they were similar to the ratings for negative samples (similarity: 11.34, id: 0.02). As designed, optimized samples were often rated as fairly similar (35.36) but not as having the same identity (0.16). Finally, positive samples were rated as similar (88.46) and the same person (0.84) for close distances as intended. However, this effect strongly decreased for larger distances.



(a) Distribution of ratings of perceived similarity for the different proposed sample types.

(b) Perceived similarity compared to the rating of a face recognition model (Dlib, line: default decision threshold) by type.

Figure 6.4: Distribution of perceived similarity based on the type of sample (left) and in comparison to the ratings of a face recognition model (right). Marker size indicates latent distance.

6.4.3 Comparing (Human Ratings to) Face Recognition Models and Distance Metrics

Figure 6.4b shows how the perceived similarity of our human raters was related to the rating by a face recognition algorithm, showing that both ratings are generally correlated. Points farther away in the latent space were generally rated less similar, which is in line with related work [249]. However, our optimized samples break both trends, being rated less similar by humans but more similar by the algorithm while also farther away in latent space. We generated the optimized samples with Dlib distance as the target function. Identification results

	perceived	face recognition models				distance metrics	
Sample type	sim	Dlib	VGG	FN	OF	lpips	latent
Interpolation	0.901**	-0.796**	-0.528**	-0.668**	-0.345**	-0.738**	-0.719**
Negative	0.637**		-0.334**	-0.21*			
Optimized	0.846**	-0.589**	-0.313**	-0.38**	-0.204**	-0.494**	0.347**
Positive	0.838**	-0.787**	-0.556**	-0.614**	-0.389**	-0.71**	-0.634**
all	0.907**	-0.765**	-0.632**	-0.712**	-0.512**	-0.814**	-0.453**

*: p < .05, **: p < .001, empty cells: not significant

Table 6.2: Correlation of different distance metrics with human-rated identity (sim: rated similarity, VGG: VGG-Face, FN: FaceNet512, OF: OpenFace). We omit genuine samples as distance scores for them were mostly constant (see Table 6.1) and correlations thus are invalid.

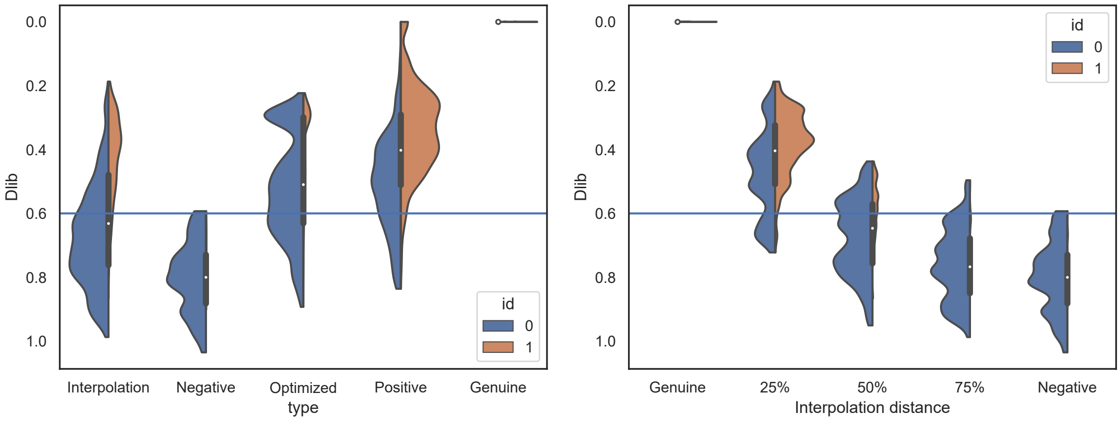


Figure 6.5: Rated identity with respect to the different types of samples and the associated face recognition score (left). Influence of the interpolation on rated identity (right). The blue line indicates the default distance for the face recognition model to accept a face.

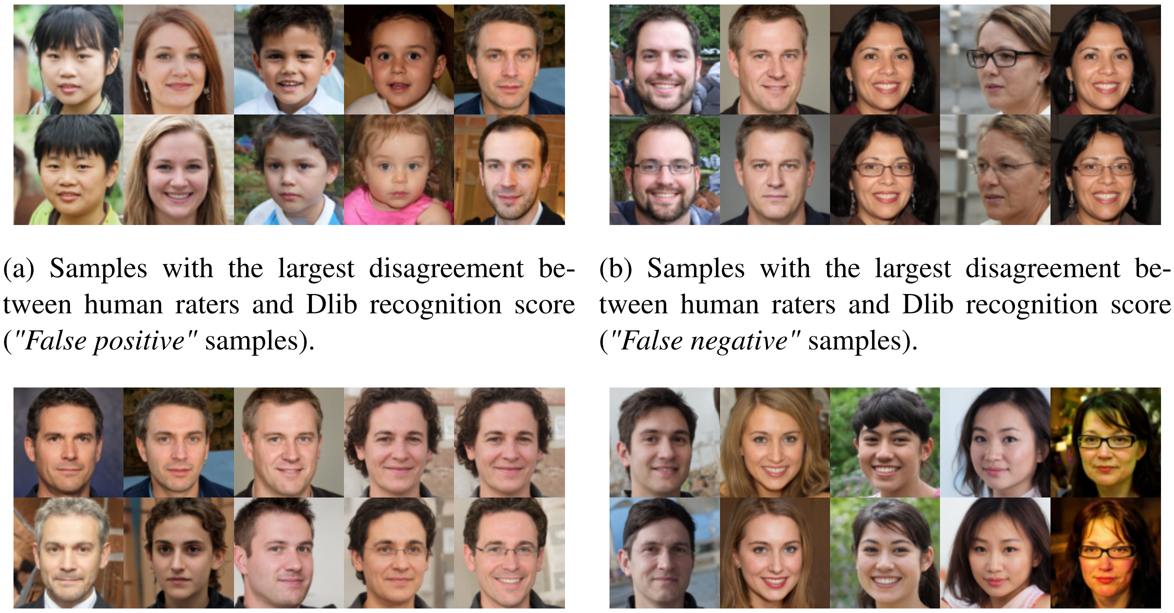
(see Table 6.1) show, that they were also most effective at being recognized by this very approach. However, the acceptance rate of Dlib was overall higher than the other models so we can draw no conclusions if generating optimized samples with a different model than the one to be tested is an effective approach in general.

We conducted a correlation analysis of perceived human identity (see Table 6.2) to better understand those effects. We find all measures to correlate most with rated identity for interpolation and positive samples. Correlation for optimized samples was overall way weaker, showing that they fulfilled their purpose to resemble samples whose identification by a face recognition algorithm is not well aligned with human perception. OpenFace showed a weaker correlation than the others and also identified fewer samples as the same person (see Table 6.1), hinting at a stricter model overall. Latent and perceptual distance (lpips) [306] were surprisingly well aligned with identity ratings by our participants.

Figure 6.5 gives more detailed insights into the distribution of samples rated as either the same or a different person. In an ideal case, we would expect to only see identical samples (orange) above the model's threshold and samples rated as different persons (blue) below. Both genuine and negative samples follow this trend while optimized and positive samples have large areas where their ratings show a mismatch between model and humans. For interpolation samples, the mismatches are mainly concentrated on the first interpolation step.

6.4.4 Finding Disagreement

One of the goals of our approach was to find samples that lead to a mismatch between human raters and a decision-making model or are generally challenging to decide. To find such cases we calculated a disagreement score between human-rated identity and the (inverted) Dlib distance score. We do not use an absolute value, as both directions are interesting and map



(a) Samples with the largest disagreement between human raters and Dlib recognition score (*"False positive"* samples).

(b) Samples with the largest disagreement between human raters and Dlib recognition score (*"False negative"* samples).

(c) Samples with the largest disagreement between participants about *perceived similarity*

(d) Samples with the largest disagreement between participants about *perceived identity*

Figure 6.6: Samples from our dataset with the biggest disagreement between model and participants (6.6a, 6.6b) and between participants themselves (6.6c, 6.6d). The top row in each figure contains base images, bottom row contains generated alterations.

to likely candidates for false positives (face recognition rating similarity higher) and false negatives (humans rating similarity higher). To find disagreements between participants we calculate the standard deviation of their ratings of similarity and identity. Figure 6.6 shows the samples with the largest disagreement with respect to the described measures. We observe, that the candidates for false positives (Figure 6.6a) are mainly children and were generated by optimized samples, hinting at a weakness in the assessed decision-making model to correctly judge the similarity of children. Potential false negatives (Figure 6.6b) were mainly generated through negative and interpolation samples. They have in common, that they differ in meaningful ways like age, pose, or accessories.

6.4.5 Determining Similarity

We asked participants for strategies to determine the similarity and identity of the given image pairs. Most participants stated that they compared facial features and decided based on faces looking similar or following their intuition. The main features participants looked at were eyes and eye color, hair (color), the nose, and the mouth area. When judging identity, participants also particularly focused on identifying details, for example, P3: “I was looking for particular facial features (for example, dimples, teeth, wrinkles, nose shape etc.)”. They also tried to ignore parts of the image that did not contribute to identity: “If the majority of faci[a]l features were identical and only the hair or clothe[s] changed I assumed that the images showed the same person.”(P5).

In addition to the feedback from our participants, we also investigated influencing factors on perceived similarity computationally. We used a facial feature prediction model on all generated samples to generate a vector of 37 features. We calculated the distance of those vectors between each sample and its respective base and trained a random forest regressor to predict the collected perceived similarity from those vectors achieving a R^2 score of 0.45. We calculated permutation importance for this model to find the features with the greatest impact on the decision. The most impactful features were related to age (bags under eyes), face shape (oval face, high cheekbones), and styles (hair and makeup).

6.5 Discussion

In this chapter, we proposed a method to explore decision-making models and applied it to gain insights into a face recognition model. Here, we reflect on our insights from this test, the limitations of our approach, and further target groups and applications to explore. We also illustrate, how our method can be used for biometric interfaces to allow users to make a more informed decision when using a face recognition model.

6.5.1 Experimental Findings and Insights

We found that human perception overall aligned with the outputs of the tested face recognition model. We found a strong correlation between human identity ratings and perceptual distance (lpips) [306], indicating it may be well suited as an approximation for face recognition. However, our generated samples were also successful in uncovering misalignments. Optimized samples were very successful at generating potential cases of false positives and similarly positive and interpolation samples led to cases where the decision-making model indicated less similarity compared to human raters. Visual inspection of the samples with the largest disagreement suggests that the tested model struggled with distinguishing children and was affected by changes like age, pose, or accessories that had no effect on the identity ratings by human raters. This has different implications for different groups. As users of the model, one has to be aware, that small everyday changes can largely impact recognition performance. As a developer of such a system, this gives starting points for what to improve. For our method, the strong disagreement for samples with meaningful but minor edits implies that specifically sampling for such differences may be a good way of finding potential false negative samples and should be considered as an addition to our proposed samples.

6.5.2 Considerations

Using a generative model to explore a decision-making model can introduce new biases or hide existing shortcomings of the tested models if biases align. While this cannot completely be avoided, we believe that introducing human raters can uncover some of those effects. At the same time, the inclusion of human raters also limits our approach to tasks that actually can be reliably rated by humans. This includes in particular decisions that people already make in their daily lives, like recognizing others by their face, voice, or the way they walk. However, comparing something like fingerprints or making a diagnosis on medical data is uncommon or requires experts, making it less suited for our method. Finally, we used StyleGAN2 [153] in this work. However, our approach should be compatible with other GAN variants and potentially also other generative methods, as long as their latent space is locally stable and can be navigated.

6.5.3 Further Application Areas

We used our proposed method for face recognition in this chapter. However, we expect it to be applicable to other models as well. The key prerequisites we see are: 1) humans can rate the model's decisions, and 2) generative models exist for the type of input (e.g. voice recognition [215] or gait and gesture recognition [305]).

While our method was designed for exploration, it can also directly support the improvement of models. Our approach inherently generates a dataset of human-labeled synthetic samples. This can be used in training to improve the model performance for those challenging inputs.

6.5.4 Leveraging generated Samples for Biometric Interfaces

Our analysis has shown that our proposed sample sets reliably produce potential mismatches between human perception and model decisions. Thus, our approach could be used to illustrate model performance for individual users. In that case, an embedding of the user's face (or other biometric features for other biometrics) in the latent space [1] can be used to generate relevant samples. No other human ratings will be needed, as the user can make the comparison of their own perception to the given system response by themselves to then decide to e.g. adjust decision thresholds, avoid failure cases, or – more generally – make an informed decision if, and how to use the model. Note, how the different sample types can provide different types of insights. Positive and negative samples allow the user to verify, that their perception of similarity generally aligns with the model, while interpolated and optimized samples can give them an impression of the robustness of the model against changes and its performance when confronted with other similar-looking people. This provides a graspable improvement over the abstract textual warning that is currently used (see Fig. 5.3c in Chapter 5) and gives users insights into the actual performance they can expect.

In addition, the visual analysis we performed could be adapted to be used in a user interface to illustrate areas of disagreement for the model as a whole in a more graspable manner than using e.g. False Positive Rates or Equal Error Rates.

For a more active user involvement, our idea could also be extended to an interactive application, generating more variations on demand. We found that meaningful but minor edits like changing the pose or adding accessories can have meaningful impacts on the recognition performance. A biometric interface could allow users to actively explore such changes, e.g. in the form of a slider interface that offers to edit such dimensions as pose, age, or lighting conditions. Related work has both demonstrated that such edits are possible [123, 137] and what interfaces for them could look like [67].

6.6 Implications

In this chapter, we proposed to leverage generative models to strategically produce alterations (positive, negative, interpolation, and optimized) to the input of a decision-making model and compare its ratings to answers by human raters. We collected a dataset of 11,200 ratings of similarity and identity for pairs of face images and compared them to the output of a face recognition model, providing insights into how the perception of humans and the algorithm differs.

We highlighted, how our method can be leveraged to *gain deeper insights into the performance of a biometric model.* We proposed how it can enable improvements to the model by using the generated labeled dataset and discussed how generated samples can be utilized to support end-users in gauging the performance they can expect from a biometric model.

As such, this chapter provides a valuable *tool* for both model developers to introspect and improve their work and designers of biometric interfaces to support users in making deliberate and informed decisions.

7 Leveraging Context Cues to Inform Authentication Choice

This chapter is based on the following publication:
Sarah Prange, **Lukas Mecke**, Alice Nguyen, Mohamed Khamis, and Florian Alt. 2020. *Don't Use Fingerprint, it's Raining! How People Use and Perceive Context-Aware Selection of Mobile Authentication.* In Proceedings of the International Conference on Advanced Visual Interfaces (AVI '20) [208]

In the previous chapter, we explored how the enrollment process for a biometric system can be improved by giving users more insights into the model's performance. After a user enrolls in a biometric system, the main point where they further come in contact with the biometric model is whenever they want to authenticate. From related work we know, that context factors like light conditions or moisture can have an impact on biometric authentication [29]. In this chapter, we explore if and how informing users of the impact of such context factors can help them in choosing an appropriate authentication mechanism in different situations.

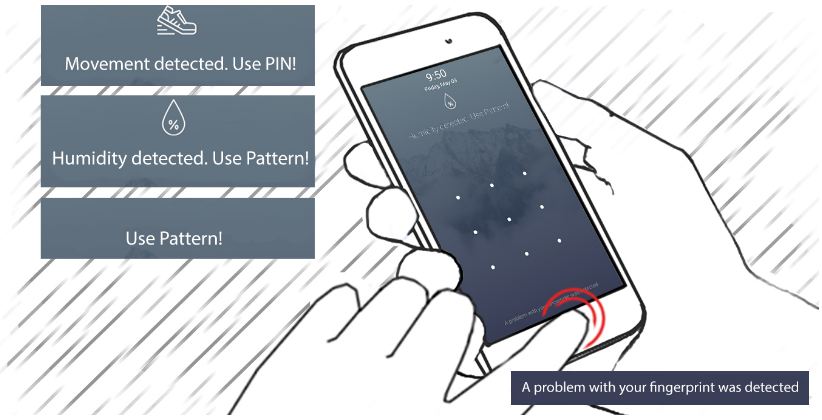

Movement detected. Use PIN!
Humidity detected. Use Pattern!
Use Pattern!
9:50
A problem with your fingerprint was detected

Figure 7.1: We investigate how people use and perceive context-aware suggestions to switch mobile authentication mechanisms. This is useful when the primary mechanism is likely to fail (e.g., wet fingers when using a fingerprint).

While related work suggests that it is technically possible to choose appropriate authentication based on context [290], we look into how context-aware selection of authentication is used and perceived by mobile users in the wild.

To gain a better understanding of the context factors impacting users in their actual interaction with authentication mechanisms we conducted an online survey and focus group collecting problems and coping strategies. This highlighted both a wish for explanations and an easy way to switch to a context-appropriate mechanism. Informed by this investigation we designed a prototype suggesting users of fingerprint authentication to switch to their knowledge-based fallback based on context information.

In a two-week field study (N=29), we tested how users used and perceived this biometric interface and if giving explanations for suggested switches was helpful for them to gain a better understanding of the model.

Our results show that users were willing to switch their authentication scheme when prompted and found our app helpful and beneficial in daily use. Participants preferred receiving an explanation. However, sometimes the app behavior and given explanations did not match participants' perceptions highlighting the need for good explanations and activation strategies. We discuss context factors, authentication switches, and use cases.

> In this chapter we contribute 1) an investigation of context factors impacting user authentication. We derive 2) a prototype for suggesting adequate authentication based on context information. A two-week field study provides 3) empirical insights into its perception and use, in particular with respect to the presence of explanations.

7.1 Background and Research Approach

Here we give an overview of some background on the utility and previous use of context factors for authentication. We extend this knowledge by conducting both an online survey and a focus group to gain a better understanding of the actual factors impacting users when using biometrics to authenticate in Section 7.2. Based on those results we derive the design of a prototype and study to investigate context-based authentication switches in the wild.

Users protect access to a plethora of personal data on their smartphones, using authentication methods such as knowledge-based or biometric schemes. However, authentication on mobile devices is error-prone [120, 171] and perceived as time-consuming – in particular because interactions on smartphones are usually short [120].

Beyond improving the security of existing mechanisms [155, 279], concepts have, hence, been suggested to reduce authentication overhead (e.g., [44, 146]). One such option is to use *context*, which refers to any (explicit or implicit) information that characterizes the user's current situation [235]. Factors include environmental properties (e.g., location), but also human factors [236]. Context can be leveraged to a) skip authentication in certain situations (e.g., *Google Smart Lock*[1]) or b) choose adequate (e.g., biometric) authentication [290].

Adapting mobile authentication based on context is very useful as we authenticate around 40 times a day [120] in varying contexts in daily life [126]. Related work shows, that context can be considered for authentication (e.g., location, proximity, or other sensor values [125, 147, 169, 253, 290]). Wójtowicz and Joachimiak [290] presented a generic model that allows choosing the "optimal biometrics" for mobile authentication based on contextual factors (e.g., no voice biometrics in silent mode). However, no such thing has been proposed or tested in practice.

7.2 Understanding Context-Aware Authentication

In this section, we report on results from an online survey (N=35) and a focus group (N=5) we conducted to inform our design of a prototype to suggest switches to a fallback based on context factors (e.g. moisture, see Figure 7.1).

7.2.1 Online Survey

Related work found that fingerprint authentication is sometimes problematic, e.g., while walking, in dark environments, or after using moisturiser [21, 29]. Only artificial environments were tested. To understand the contexts in which users encounter difficulties when

[1] `https://support.google.com/android/answer/9075927`, last accessed October 16, 2024

authenticating on arbitrary devices, we conducted an online survey (N=35, 20 female, mean age=28). We did not limit this to biometrics to get a broad spectrum of experiences.

Respondents were asked to describe any problems they encountered in as much detail as possible, followed by open-ended questions about the context of the incident and the perceived reason behind the problem. We asked for the time of day, weather, and location as those might have an impact. Participants were recruited via university mailing lists and took part in a raffle for three 20 € vouchers.

We discarded one response since it did not contain a problem situation. In the remaining 34 responses, smartphone-related issues were predominant (22 out of 34). The majority of those were about issues with lockscreens (14) and fingerprint authentication (12). From those, we identified wet or dry fingers as the main source of failed authentication attempts, e.g., P30: "*When [my] hands are sweaty the smartphone can't be unlocked using fingerprint. This mostly happens in crowded [public] transport.*". Reported contextual causes for wet or dry fingers were temperature (rain or muggy weather), location (kitchen, bathroom, or public transport), and activities (walking, applying moisturizer, washing hands, or exercising).

7.2.2 Focus Group

We conducted a follow-up focus group (N=4, 2 female, mean age=26.3) to further investigate problems with fingerprints and coping strategies. Participants were compensated with 10 €. We asked about *issues* encountered when authenticating using fingerprint authentication on mobile devices and their *coping strategies* to overcome said issues. We concluded by collecting feedback on the idea of *leveraging context to suggest switches to fallback*.

Named *problems* were dirty/wet fingers, similar to the online survey. The reasons were cooking, the winter season (dry fingers), and neurodermatitis. The predominant *coping strategy* encompassed repeated scanning of the finger. Other options were registering multiple fingers, addressing the problematic state (e.g., drying or moisturizing fingers), or using different methods (e.g., a fallback mechanism). Participants were very positive about the suggested switches to fallback based on context information. Their design priorities were saving time (e.g., by omitting the manual swipe gesture to get to the fallback), having a visual indication of the currently used method, and receiving (brief) explanations for system decisions.

In addition to confirming the online survey's results, we found that switching to fallback mechanisms is not among the common coping strategies due to the required effort. However, participants thought positively about alleviating the need to actively switch the authentication method. Overall, participants favored concepts that are transparent and save time, which aligns with previous work [120, 197].

7.2.3 Lessons Learnt

From related work, we learn that fingerprint authentication is error-prone (e.g., for wet or moisty fingers [29]). This was also reported by our survey and focus group participants. While related work showed that choosing an appropriate authentication scheme based on context is possible [290], it is not known whether users will follow derived suggestions *in the wild*. Participants of our focus group also wished for an explanation, which was highlighted by prior work to be important for intelligent systems in general (e.g., [209]). Indeed, one of the usability heuristics is to maintain the "visibility of the system status" [198]. This is also in line with our findings in previous chapters. Our online survey and focus group further show that there is a need for transparent and straightforward ways to switch to fallback mechanisms when authentication is not possible due to contextual factors.

Thus, we also investigated in our field study if *explaining* suggested switches to fallback mechanisms impact users' decision to follow this suggestion. We made concrete suggestions for a switch to make this step easy and straight forward for users.

7.3 Leveraging ContextLock to explore authentication switches

As outlined in section 7.1, we see great potential in leveraging context to (proactively) suggest switches of authentication methods. As a prerequisite for this, the aim of our work is to evaluate user perception towards authentication switches in the wild.

7.3.1 Prototype: ContextLock

Informed by our survey and focus group we developed an Android application, *ContextLock*, to provide suggestions to switch to fallback based on context. Due to Android's security limitations and ethical concerns, we did not replace the lock screen but simulated a failed authentication attempt by showing a fallback screen after successful user authentication.

We built an Android fallback authentication screen allowing for PIN[2], pattern, and fingerprint[3] authentication. The presented fallback was determined by a question in the initial questionnaire to match participants' routines. Fingerprint was always provided, as we did not want to force a switch but allow users to retry using fingerprint (which we found to be a common coping strategy) if they wished.

[2] PINLockView: `https://github.com/aritraroy/PinLockView`, last accessed October 16, 2024

[3] Lock-Screen: `https://github.com/amirarcane/lock-screen`, last accessed October 16, 2024

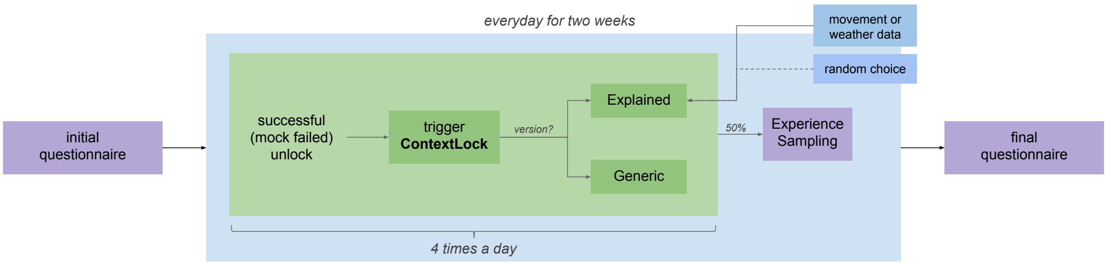


Figure 7.2: Procedure of our two-week field study, including all data sources (marked in violet).

To acquire the user's context, we integrated OpenWeatherMap[4] and Google's activity recognition API[5]. If no fitting context data was available, we randomly chose either *"Humidity detected"* or *"Movement detected"* as mock context reason (see Figure 7.1).

7.3.2 Study Design and Procedure

We designed our study as a two-week within-subject field study with the presence of EXPLANATIONS for a suggested switch to fallback as an independent variable with two levels: *generic* or *explained* (see Figure 7.2). Dependent variables were participants' subjective ratings from questionnaires and experience sampling probes as well as decisions on whether to switch to the fallback.

The study duration was initiated and concluded with a questionnaire, asking for demographics and a final comparison of the study conditions. During the study period, participants were presented with a proposed switch to their fallback four times (average authentication failures as reported in the initial questionnaire) in random intervals between 8 a.m. and 9 p.m. every day. Explanations were given depending on the current study condition (see Figure 7.1), which would automatically switch midway through the study. After a successful unlock (by either using the suggested fallback or fingerprint) a dismissable experience sampling (ES) questionnaire was shown with 50% probability. Conditions were counterbalanced.

7.3.3 Participants

We recruited participants via university mailing lists and social media. From a total of 42 installations, 29 participants (12 female) between 18 to 45 years (Mn=23.6) completed the field study. Participants were located in the UK, Central America, Russia, and Italy when using the app. The majority came from Germany. Participants used PIN fallback (16) and pattern fallback (13).

[4] Open Weather: `https://openweathermap.org/api`, last accessed October 16, 2024

[5] Activity Recognition: `https://developers.google.com/android/reference/com/google/android/gms/location/ActivityRecognitionClient`, last accessed October 16, 2024

7.3.4 Limitations

Due to strict battery handling on Huawei phones, our application was sometimes terminated by the operating system. To counteract this, we showed an icon in the taskbar to indicate that the app was active and kindly asked participants to manually restart it if the symbol disappeared. We analyzed all records with at least four (of seven) days of data for each condition. For the context of our study, we decided to trigger our app in certain intervals rather than triggering based on actual context factors. We made this decision to ensure consistent data collection, though a real application would do it the other way around. Furthermore, our sample was self-selected and biased towards younger European students.

7.4 Results

Here we present our results from the initial and final questionnaire as well as experience sampling probes (compare Figure 7.2). Significance was determined using Wilcoxon and McNemar's tests and is reported at a significance level of p = .05.

7.4.1 Prior Authentication Behaviour

We found the most common coping strategy (stated by 24 of the 29 participants) with fingerprint failure to be switching to the fallback after *multiple* failed attempts. Fewer participants would switch immediately (2), try again later (2), or ignore it (1). Some participants reported lockouts (complete loss of access to their device) at least once a day (5) or more than once a week (4). Nine participants experienced lockouts once a month at most and eleven never encountered this problem. The responses about perceived fingerprint error frequency were also mixed, from once (8) or more than once (7) a day, once (7) or more than once (6) a week, to less than once a month (1).

Overall, this shows that fingerprint errors and, in some cases, resulting lockouts are indeed a problem and there is room for improving the current coping strategies.

7.4.2 Experience Sampling

After data clean-up, we had a total of 253 complete sets of experience sampling data for the *generic* version and 261 for the *explained* version. On a 5-point likert scale (0=strongly disagree; 5=strongly agree) the situational *annoyance* level was rated as neutral for both versions (*generic*: Mn = 2.02, SD = 0.976; *explained*: Mn = 2.03, SD =1.126). We found no significant difference between the versions.

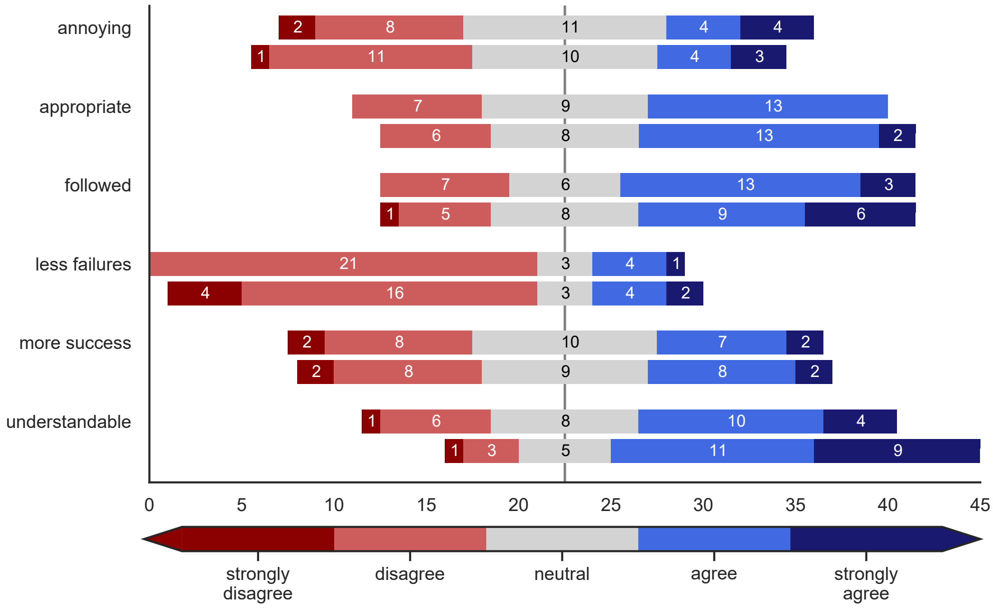


Figure 7.3: Responses for each Likert item in the final questionnaire for the *generic* (upper)/ *explained* (lower) version.

Though both were rated about neutral, a significant difference can be observed for the perceived *appropriateness* (Z = -3.031, p = .002). The *generic* design was perceived significantly more adequate (Mn = 2.13, SD = 1.073; *explained*: Mn = 1.85, SD = 1.143).

Participants were asked for possible reasons for fingerprint failure while using the *generic* version of *ContextLock*. Overall, wet (74) and dirty (62) fingers constituted the majority of perceived reasons. Weather and ambient influences such as rain (5), snow (2), humidity (14), and heat (14) were indicated as influential factors which are in line with prior work [290]. Other reasons were “movement” and “damaged fingers from climbing”. This confirms our survey’s results.

7.4.3 Switching Behaviour

We recorded if participants followed our suggestion to switch to their fallback mechanism. We collected 645 datasets for the *generic* (no explanation) and 611 for the *explained* design. Users showed no significant differences (p > .05) in following our suggestion with 67.13% (SD= 0.470) and 67.76% (SD = 0.468) of the cases, respectively.

7.4.4 Overall Rating

Figure 7.3 shows participants' overall rating of the two conditions with regards to understandability, appropriateness, increased success, fewer failures, annoyance, and if they used the fallback. We found no significant impact of the conditions.

Besides conditional questions, we also asked for overall opinions. In summary, users found *ContextLock* somewhat helpful (Mdn=4), thought that the automatic recommendation was beneficial compared to their current lock screen (Mdn=4), and would use a similar system in the future (Mdn=4). The majority preferred the *explained* version (22) over the *generic* one (6); another six participants remained undecided.

15 participants made comments about situations in which they would have wanted *ContextLock* to activate (but it did not). Reasons were "humid" environments, "wet fingers while walking in the rain" or "misplacement" of the finger on the sensor. Participants also mentioned increased battery usage and a wish to customize *ContextLock* to more closely resemble their real lockscreen. Three participants liked the usability of the app and two commented to have enjoyed the design. We saw no feedback indicating participants noticed some of the given reasons being random.

7.5 Discussion

For current biometric authentication mechanisms, users need to switch to a knowledge-based fallback in case the primary mechanism does not work (as expected). This takes time and is annoying to users, as reported in our focus group. We suggest considering context to switch authentication mechanisms, not only in case fallback authentication is necessary, but on a per use case basis. We now discuss further aspects of our concept as well as opportunities for future work.

7.5.1 Appropriateness of Suggestions

Participants perceived suggested switches as significantly less appropriate when they were given an explanation. We believe the reason for this to be the use of fake context information (when no real data was available), hampering trust in the system [163]. However, we saw this effect only for the experience sampling and not the final rating. While we did not find significant differences, participants rated the explained version more understandable and perceived fewer failures. It was also rated as the preferred choice. This shows that participants generally appreciate explanations, though the use of real context data would be necessary to make the system transparent (as, e.g., suggested by Nielsen [198]). Specific use cases and related context factors are subject to future work.

7.5.2 Extending the Concept

From related work, we learn that environmental as well as technical factors [125, 235, 236, 290] may influence the choice of authentication.

It's Raining vs. I'm Tired

We propose to consider not only technical but also further human factors. This may, on one hand, refer to users' concrete characteristics, as, e.g., hand size has an influence on the accuracy of touch interaction [40, 207]. On the other hand, more abstract factors like users' current cognitive and physiological state may be worth considering when choosing authentication. As an example, using face recognition may be more usable than entering a PIN for switching a song while doing sport. Users who are at home and tired may as well not want to enter a knowledge-based secret but rather rely on their trusted environment.

Socially-Aware Authentication

Frankel and Maheswaran [97] showed that human social interaction is a feasible authentication factor, thus also social context could be leveraged for authentication switches. This may, on one hand, lead to a switch to an easier, potentially less secure mechanism, if only trusted entities are present. On the other hand, users may want to hand over their device to someone else. While it is easy to share a knowledge-based secret, a biometric secret cannot be shared. Context-aware authentication could thus switch to knowledge-based models if the device is in the hands of a trusted, but foreign entity.

7.5.3 System vs User-Initiated Switches

Our prototype *suggested* the switch to knowledge-based fallback, but did not force users to do so. However, our participants did follow the recommendation in the majority of the cases (67%). Other approaches may provide users with the possibility to choose context factors to be considered themselves (compare to, e.g., *Google Smart Lock* or aCapella [78]). At the same time, context-aware authentication could also force the switch of authentication mechanisms based on appropriate factors. This would introduce a trade-off where users loose agency and insight into their used authentication but may gain convenience.

7.6 Implications

In this chapter, we explored context-induced problems users have with authentication and their coping strategies. We derived a concept for an app –*ContextLock* – which helped us to understand users' willingness to follow context-aware suggestions (both mocked and real) for authentication switches in the wild. We tested *ContextLock* in a 14-day field study with 29 participants.

We found no significant differences between giving explanations or not, but explanations were overall preferred. Participants liked the concept and found it useful and worth using in the future. This was also reflected in about 67% of the cases in which participants followed our suggestion and switched to their fallback. While we saw no feedback indicating participants noticed some of the given reasons being random, we believe this had an impact on the perceived appropriateness of suggestions to switch.

This chapter highlights, that users both *wish to understand context factors* for their authentication and have an *easy and straightforward way to cope* with them. We found, that fingerprint errors and lockouts are indeed a problem and there is room for current coping strategies (e.g. just retrying multiple times) to be improved. Our proposed biometric interface was actively used and *preferred to having no explanations*. However, this chapter also highlights, that it is important for explanations to match users' actual experience to be useful in building a mental model and support user literacy.

8
Communicating the State of Continuous Authentication Systems

This chapter is based on the following publication:
Lukas Mecke, Sarah Delgado Rodriguez, Daniel Buschek, Sarah Prange, and Florian Alt. 2019. *Communicating Device Confidence Level and Upcoming Re-Authentications in Continuous Authentication Systems on Mobile Devices.* In Fifteenth Symposium on Usable Privacy and Security (SOUPS '19) [182]

In addition to being used as an explicit authentication mechanism, biometrics (in particular behavioral) can also be used to provide security implicitly [39, 71, 76, 192, 226, 247]. Implicit authentication has two major use cases: a) as an effortless, independent main authentication mechanism [156]; or b) as a second line of defense against unauthorized access [168]. The first use case is particularly suitable for smartphone users who currently do not use any kind of authentication on their devices due to the required effort of explicit mechanisms.

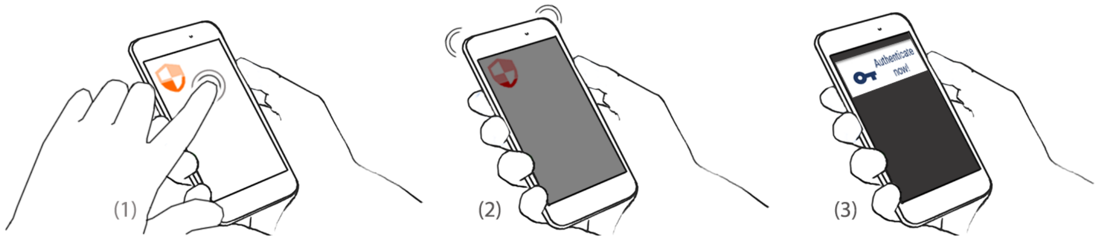


Figure 8.1: We propose to use indicators to communicate both the current device confidence level (*DCL*) and the need for re-authentication for continuous implicit authentication systems on mobile devices: 1) a *long term* indicator illustrates the current *DCL* and its development over time via a taskbar icon, and 2) a *short term* indicator announces an upcoming re-authentication via darkening the screen. Our system also allows for 3) *voluntary* re-authentication to avoid system-side locking of the device.

Those users would need to authenticate less frequently than with traditional explicit authentication approaches [121, 156]. The second use case provides an additional security barrier for devices that were already unlocked using an explicit mechanism [168].

One caveat of such implicit authentication systems is that they can trigger explicit re-authentication; that is: asking users to confirm their identity via a second factor, in case the mechanism is unable to confirm the current user's identity [93, 156, 168]. Such re-authentication events are likely to interrupt other tasks and annoy users [157]. Reasons for this annoyance include the unpredictability of interruptions and the sensation of not being correctly informed about the current state of the implicit authentication system [5, 62, 157]. Moreover, users wish to influence the timing of the interruption in some way [5, 178]. The addition of a biometric interface could help to address these user needs.

We thus propose 1) a long term indicator (*LT*), informing users about the current device confidence level (*DCL*) and thus enabling anticipation of upcoming re-authentications, and 2) a short term indicator (*ST*), enabling users to finish their task. To avoid system-side locking of the device we 3) provide *voluntary* re-authentication (see Figure 8.1).

We investigated these indicators in a field study (N=32) where participants used them in everyday life. We found that participants preferred our indicators to a system that interrupts them in an unpredictable way. Their perception strongly depended on the importance of the interrupted task. Voluntary re-authentication was perceived as less annoying. Our research is complemented by deriving implications for future implicit authentication systems.

> In this chapter we contribute 1) a biometric interface to announce upcoming re-authentications and allow for voluntary re-authentication; 2) findings from a 4-week field study, testing the two indicators and their combinations; and 3) a set of implications for future implicit authentication mechanisms based on our findings.

8.1 Background and Related Work

In this section, we give an overview of related work on the perception of implicit authentication, re-authentications, and interruptions. We conclude with an overview of the implications we derive from this related work and the research questions guiding our work.

8.1.1 Perception of Implicit Authentication

There are several works pointing out the positive effects of implicit authentication. Hayashi et al. [125] found that implicit authentication could reduce explicit authentication by 68%. Riva et al. [218] report a decrease of 42%. Several studies report on implicit authentication being perceived as convenient and easier to use than traditional methods [58, 110, 157]. In a study by Crawford and Renaud [62] 90% of the participants indicated they would consider using implicit authentication and 73% felt it was more secure than authenticating explicitly.

8.1.2 Research on Re-Authentication

While implicit authentication is generally perceived as positive and can indeed reduce authentication overhead, previous work found that the need for re-authentications can strongly disrupt those positive effects. Khan et al. [157] found that re-authentications, due to *false rejects (FR)* (i.e., cases in which the system rejected the legitimate user), were perceived annoying by 35% of their participants. This was due to both the unpredictable nature of the interruption and the need to switch the context for re-authentication. Another finding, also supported by the study of Crawford and Renaud [62], was that security barriers – like re-authentication – helped users to build a mental model of the system's security and thus led to a stronger perception of security.

8.1.3 Research on Interruptions

Work by Bailey et al. [18] found that interrupting users is perceived as rude and decreases task performance. The timing of an interruption was highly important, as interrupted tasks were perceived as more difficult. Thus, they suggest using *attention manager* systems to detect phases of low memory load and schedule interruptions during these.

Adamczyk and Bailey [3] further investigated the impact of triggering interruptions at opportune moments. They were able to show that better timed interruptions are perceived as less annoying, less frustrating and more respectful. They also require less mental effort. Fischer et al. [96] aimed at identifying such opportune moments for interruptions with smartphones with the goal of identifying the best timing for delivering notifications. Although their participants did not clearly prefer the suggested interruptions after finishing a task compared to random interruptions, they found people attending faster to notifications in the task-dependent condition.

McFarlane [178] studied interruptions in general and found that making interruptions more predictable made them less annoying and had a positive effect on user performance in the interrupted task. He also found that letting users determine the moment of interruption made interruptions less annoying. Agarwal et al. [5] found similar results in their study. They tested different mechanisms to delay the re-authentication interrupt, using gradual dimming of the screen and transparent overlays to reduce context switch overhead and unpredictability of the interrupt. They found indications that participants were less annoyed when they could predict the interruption. Participants liked the introduced *grace period* (i.e., the delay of the re-authentication) and performance was increased as users tried to finish their tasks before the device was locked.

8.1.4 Implications of Related Work

Based on the previous work we derived three main user needs, that biometric interfaces for continuous authentication should address.

Show current state: Crawford and Renauds [62] found that users disliked the idea of a totally invisible authentication mechanism. Khan et al. [157] suggested indicating the current system status to address similar concerns voiced by participants of their study. This suggests that users' general desire for system feedback is particularly true for authentication as well.

Announce interrupts: Agarwal et al. [5] and McFarlanes et al. [178] found that predictable interruptions make users feel less annoyed.

Delay interrupts: Instantly locking the device when re-authentication is required can heavily disrupt the interaction flow [18]. Prior work showed that users liked having a *grace period* to finish their tasks in these situations [5].

8.2 Concept Development

In this section we report on the development process for our re-authentication concepts: We introduce design considerations revolving around *presentation strategy* and *integration with the smartphone*. These considerations provide the framing for a subsequent focus group in which participants brainstormed about specific designs. In the next section, we describe our final concept for indicating upcoming re-authentications based on related work, our design considerations, and our findings from the focus group.

8.2.1 Design Considerations

Based on the requirements derived from related work we considered different aspects of our design. In particular, we considered how indicators should be presented and how this presentation could be integrated into a smartphone interface.

Presentation Strategy

Here we derive two approaches for presenting a re-authentication indicator: long-term and short-term. We consider and investigate both.

To show the current state of the system we propose a *Long Term Indicator* as a permanent indicator displaying the device confidence level (*DCL*) to show that the system is active. This also serves as a means to anticipate upcoming re-authentication. To inform users about the imminent need for a re-authentication, we propose a *Short Term Indicator*, granting a grace period similar to the gradual dimming used by Agarwal et al. [5].

Integration with the Smartphone

The biometric interfaces can be integrated with the smartphone in different ways: by means of static elements with the main purpose of permanently showing the current system status; by using dynamic elements, announcing an upcoming re-authentication request; or a combination of both approaches (hybrids).

A well-suited *static element* on mobile devices is the taskbar, as it is (with few exceptions) always shown. Possible elements are icons, percentages, progress bars, or changes to the bar itself (e.g., changing color) to indicate the current *DCL*. Possible On-screen *dynamic elements* include distortions of the screen content (e.g., darkening, desaturation, pixelation, etc. [5, 9]) or a notification. Off-screen elements include vibration, sound, the use of the flashlight, or the notification light. Finally, *hybrid* elements could be used. An element that can be used both statically and dynamically is a floating action button, overlaying screen content. Such buttons can show both *DCL* and upcoming re-authentication requests, either color-coded or in the form of e.g., a counter. In particular, a floating action button could also remain invisible and only (gradually) appear to announce a re-authentication.

Freedom of Authentication

To address annoyance due to having to wait for the grace period to finish [5], we propose allowing explicit re-authentication at any time and in particular during the grace period.

8.2.2 Focus Group

We conducted a focus group to gain further insights towards the design of our biometric interfaces. The focus group served two purposes: 1) To collect novel design ideas for re-authentication concepts, focus group participants engaged in an open brainstorming session. 2) To understand users' preferences regarding the design opportunities, participants discussed several designs, covering different aspects of our considerations. We recruited five HCI students from our university (4 female, 1 male) for their expertise in interface design.

Procedure

We first introduced participants to the concept of continuous implicit authentication and explained the terms 'device confidence level' (*DCL*) and 're-authentication'. Afterward, we asked them to sketch ideas of how the current *DCL* and the need for re-authentication could be communicated to users. We provided print-outs of smartphone home screens. Furthermore, we nudged them to think beyond visual cues. Following the sketching phase we asked them to present their ideas and discuss them. We then presented a set of our own indicator designs and asked participants to discuss those. Finally, we asked participants to rank all designs (their own and our presented ones) and comment on why they chose a ranking.

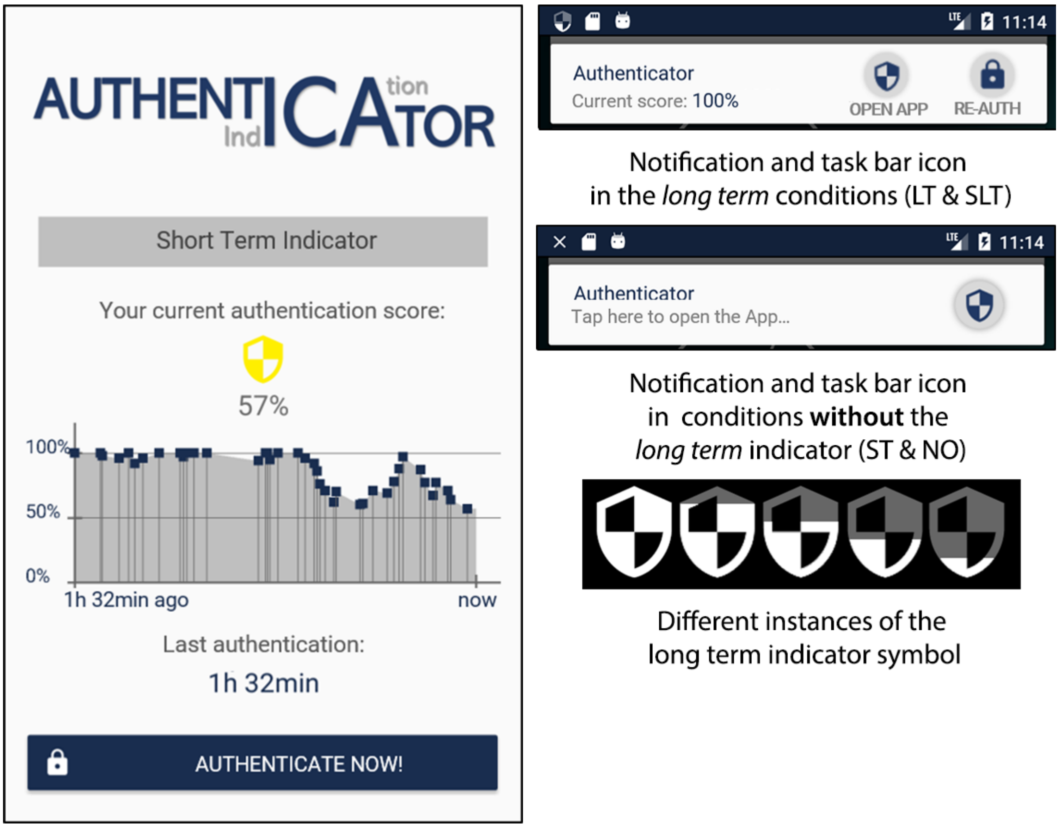
Focus Group Results

The results of our focus group covered integration with the smartphone, visual design, modalities, and re-authentication mechanism.

Participants favored approaches that subtly *integrate the indicator with the smartphone*. In particular, they felt that the indicator would optimally be placed in the taskbar. Floating action buttons were perceived as too intrusive. Notifications received mixed opinions: While some participants argued that they were intrusive, others described them as the natural way the device would communicate announcements. Regarding the *visual design*, participants suggested indicators gradually changing appearance (such as color) to make users aware of diminishing *DCL*. Abrupt color changes were considered too intrusive. A positively perceived idea was dimming the screen (similar to the method used in [5]). Regarding *modality*, participants mentioned notifications and vibration to announce upcoming re-authentication. As *re-authentication mechanism*, most participants mentioned biometric methods (fingerprint or face recognition) to make the process as smooth as possible. This is in line with feedback from participants in the study by Khan et al. [157].

8.3 Authenticator Prototype

Based on the recommendations and suggestions both from related work and the focus group we built an Android app, called *Authenticator*. The app simulates an implicit authentication system. It provides two different types of indicators that can be combined but also work independently.



Notification and task bar icon in the *long term* conditions (LT & SLT)

Notification and task bar icon in conditions **without** the *long term* indicator (ST & NO)

Different instances of the long term indicator symbol

Figure 8.2: Different elements of the Authenticator app. Left: the main application with the device confidence level (*DCL*) visualized as a graph. Right: The notification and icon shown in the *long term* conditions (top), in the conditions without a *long term* indicator (middle), and the instances of the indicator symbol showing the current *DCL* in the task bar.

8.3.1 Indicator Designs

Our prototype supports two indicators, namely a *short term* and a *long term* indicator.

Long Term Indicator (LT)

To realise the long term indicator, our application places a permanent (non dismissable) notification in the task bar (see Figure 8.2 right top). As an icon, we used a shield that gradually darkens in five steps, according to the *DCL* (see Figure 8.2 right bottom). In the notification, we displayed the current *DCL* value together with a button to open the control application and *re-authenticate voluntarily*. While we decided to permanently display the indicator in our study, it could also be implemented as an on-demand information source (comparable to e.g., battery level) to free up space in the taskbar.

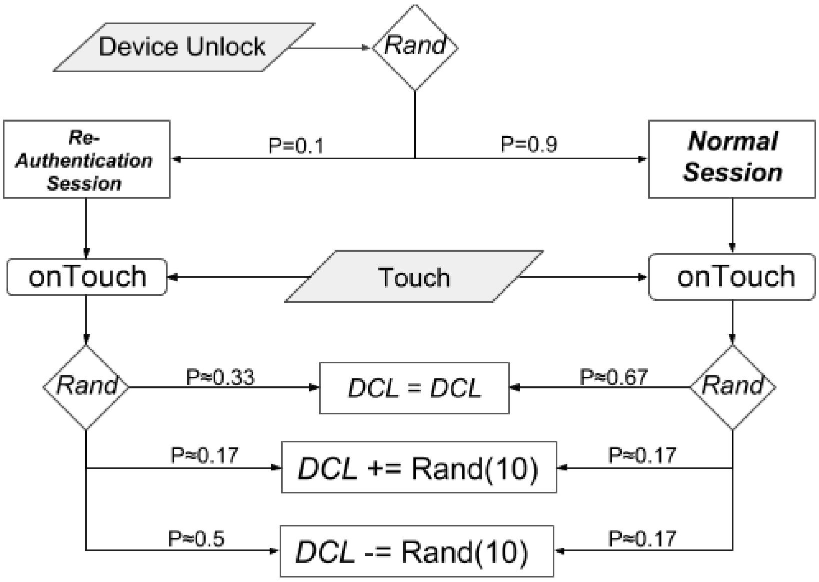


Figure 8.3: Schematic presentation of our simulated implicit authentication mechanism: Upon unlock of the device we determined (based on the desired false acceptance rate of 10%) whether a re-authentication should be triggered in this session (*re-authentication session*). The probabilities of user touches influencing the device confidence level (*DCL*) are altered accordingly; leading to decreases being more likely in *re-authentication sessions*. In *normal sessions* the *DCL* is more likely to remain stable.

Short Term Indicator (ST)

The short term indicator gradually darkens the screen once the *DCL* falls below 20% (Figure 8.1 center). It is therefore only visible when a re-authentication is imminent. To avoid annoyance by waiting for the grace period to end (see [5]), we display a notification as the dimming period begins. It shows a button to allow the user to *voluntarily re-authenticate* at any point within the grace period (Figure 8.2 right top).

In the study by Agarwal et al. [5] a duration of 4 seconds was chosen as shorter amounts did not allow for anticipation of the re-authentication and for longer duration testers had to wait too long for the re-authentication to appear. Due to the introduction of voluntary re-authentications the latter finding does not hold in our setting so we also explored longer grace periods. Through testing with five participants, we determined a grace period duration of 8 seconds to be suitable. To address the remaining uncertainty we included a question about the desired length of the grace period in the final questionnaire.

8.3.2 Simulated Implicit Authentication

We followed related work and used a simulated system: Khan et al. [157] interrupted sessions after a random time period of between 5 and 30 seconds. Using a simulated system provides more control for our evaluation of the indicator concepts and helps to avoid differing false reject rates (e.g., due to hand posture) that might have an influence on the results [44, 62, 157]. We thus favored a simulated system based on the number of touch interactions over a real implicit authentication system to keep conditions comparable. Following the medium-level false reject rate of 10% used in related work [157], our system triggers re-authentication in approximately one out of ten sessions[1]. To achieve this, we simulated DCL fluctuations as follows (see Figure 8.3):

Selection of Re-authentication Sessions

We flagged a session as a *re-authentication session* with a probability of 0.1 (to achieve 10% false rejects) upon unlocking the device. This flag influenced the random DCL fluctuations (see Figure 8.3) such that a re-authentication would likely appear in this session. For cases where sessions were too short for a re-authentication request to appear (i.e., the *DCL* did not fall below the threshold before the session ended), the flag would persist until a re-authentication was triggered. Depending on the flag being set or not, changes to the *DCL* were simulated differently, as explained next.

Alterations to the DCL

Depending on the chosen type of session (*re-authentication* or *normal*) the goal was to either decrease *DCL* or keep it stable while adding some fluctuation to make the results more believable. Each touch by the user had a chance to either trigger a change to the *DCL* (0.67 if it was a *re-authentication session*, 0.33 in a *normal session*) or leave it unchanged (with inverse probability accordingly). For *re-authentication sessions*, a decrease of the *DCL* was more likely (0.5) in comparison to increases (0.17). In *normal sessions* the probability for decreases and increases was equal at 0.17 (compare Figure 8.3 for an overview of the whole process). Both decreases and increases to the *DCL* could trigger a random change between 1% and 10%. Decreases resulting in a *DCL* below 20% were only executed in *re-authentication sessions*.

All probabilities were determined through a pre-study with five testers so as to create fluctuation of the *DCL* that seemed natural. A re-authentication was triggered as the *DCL* fell below 20% and completing a re-authentication reset the *DCL* to 100%. Re-authentication was suspended during calls.

Usage

Using this method we achieved an actual false reject rate of 7.65% in our 4-week field study. The deviation from the goal (10%) is a result of sessions that were too short to trigger a re-authentication. While we forced the next session to be a re-authentication session in those cases as described above, we did not adjust probabilities afterward to mitigate effects on the overall false reject rate.

[1] A session refers to the time between two unlocks.

8.3.3 Re-Authentication

Voluntary re-authentication was possible using the control application (Figure 8.2 left) or one of the notifications tied to the indicators (Figure 8.2 right), i.e., the permanent notification or the notification displayed during the grace period. Information about the current *DCL* was provided by the permanent notification icon (discretized), the permanent notification, and the control application. The latter additionally featured a graph, displaying the history of the *DCL* over time (Figure 8.2 left).

The *re-authentication process* itself was implemented by locking the device and, hence, forcing the user to authenticate by using their default unlock mechanism. Due to technical restrictions, it was not possible to offer biometric methods for re-authentication as Android requires using the backup authentication scheme in cases where the device is locked by an app. Using those methods was still possible for normal locks, i.e., locks that were not triggered by our app.

8.4 Evaluation

Here we describe the research questions guiding our work and the study design we derived to answer them as well as the procedure and participants of our study.

8.4.1 Research Questions

Our evaluation was guided by these research questions:

RQ1 **Can indicators reduce annoyance caused by unpredictable re-authentication requests?** We hypothesize this to hold true due to results from related work [5, 178].

RQ2 **Are there other factors influencing annoyance caused by re-authentication requests?** We propose the location, task, importance, and sensitivity of the interrupted task as possible factors.

RQ3 **Do indicators nudge users to voluntarily re-authenticate?** We expected an increasing number of voluntary re-authentications for short term (due to the option to re-authenticate during the grace period) and long term indications (due to the added feedback from the taskbar symbol and the graph visualization of the *DCL*).

RQ4 **How do users perceive and respond to the introduction of voluntary re-authentication?** We expected users to like this feature, as prior work showed that letting users determine the interruption time reduced annoyance [178].

8.4.2 Study Design

To answer our research questions we conducted a field study (N=32) with a within-subject design. This is (to the best of our knowledge) the largest and, next to the study by Khan et al. [157], the only field study on this topic. As an independent variable, participants tested a set of four INDICATOR CONFIGURATIONS for one week each, resulting in a total study length of four weeks. The order of conditions was counterbalanced. Details are given below.

NO ***N**o Indication*: Our (simulated) implicit authentication scheme runs transparently in the background. Re-authentication is requested without prior indication, which resembles the current practical standard. Voluntary re-authentication is only possible from the control app, but not from notifications.

ST ***S**hort **T**erm*: Only the *short term* indicator is shown. Voluntary re-authentication is possible from the control app and the notification triggered with the grace period.

LT ***L**ong **T**erm*: Only the *long term* indicator is shown. Voluntary re-authentication is possible from the control app and the permanent notification.

SLT ***S**hort & **L**ong **T**erm*: Both indicators are present. All options for voluntary re-authentication are possible.

Note how both *NO* and *ST* can serve as baselines here. The *NO* condition, i.e., locking the device without giving an indication, is the current *practical* state of the art and thus a natural baseline. Furthermore, our *ST* condition is based on the best-performing method from the study by Agarwal et al. [5] (including their recommended change of allowing for re-authentication during the grace period). As such, *ST* serves as a baseline for the best currently known scheme for indicating re-authentications.

8.4.3 Procedure

We recruited participants through a University mailing list and via social media. They were asked to sign a consent form and install our app from the Google Play Store, using an installation guide we provided on a dedicated website. This website also provided additional information about all study conditions and answers to frequently asked questions.

Participants had to *use the application* for four weeks with conditions automatically switching each week. They used their phones as usual with occasional interruptions by our system and a maximum of three (dismissible) *experience sampling questionnaires* per day after successful re-authentication. After each condition switch, we asked participants to fill a *weekly questionnaire* about their experience. In the end, we concluded with a *final questionnaire*.

After four weeks, participants could uninstall the app and we invited them to participate in a *final semi-structured interview* to collect qualitative feedback (in person or via telephone). Participants received 20 €, plus 5 € if they participated in the interview.

8.4.4 Collected Data

We collected *usage data* on participants' devices, including executed apps, and aggregated touch interactions, unlocks, and re-authentications. Collected data was stored on the device and transferred to our server once per day.

The *experience sampling questionnaires* asked for the current location and interrupted task. We also asked if the interrupted task was perceived as sensitive and important and if the interruption was perceived as annoying.

In our *weekly questionnaires*, participants rated on a 5-point Likert scale if they felt rewarded by an increasing *DCL*, if they felt motivated to re-authenticate voluntarily, and if they perceived the system as obstructive, annoying, and easy to use. We also asked for free feedback on what they liked and disliked about the current indicator and the system in general.

In the *final questionnaire* we asked participants to rank the four conditions and explain their decision. In particular, we asked which features of the first and last choices contributed to their decision. For the specific indicators, we asked participants whether they would modify the duration of the grace period, if they were stressed due to the grace period, and if the long term indicator helped predict re-authentications.

Furthermore, participants rated several statements on a 5-point Likert scale: Did they like the system, were they annoyed by the vibration or notification (*ST*), did they feel that the system influenced their behavior, and did any bugs influence the system performance? Similarly, we asked participants if the experience sampling was annoying, and if it influenced their behavior or their perception of the system.

Moreover, we asked if participants had read the introduction on the website and watched the introductory video we provided, if they had previous knowledge about implicit authentication, and if they had looked up app functionality or how implicit authentication worked in general on our website or other sources. Finally, we asked if they always locked their phone after use, if they thought re-authentication interrupts were more annoying than traditional authentication, and if they would consider using implicit authentication.

In the *final interview*, we asked participants to share their experiences with the systems guided by a few questions.

8.4.5 Participants

We recruited 36 participants. Four were excluded since their data was not properly transferred to our server. The remaining 32 people had a mean age of 28 years (18 male and 14 female; Table 8.1). Three participants did not submit a final questionnaire, resulting in a reduced set of 29 answers for these questions. For practical reasons, we conducted the study in two runs (i.e., not all participated in parallel).

Gender	14 (44%)	Female
	18 (56%)	Male
Mean Age	28.3	
Occupation	2 (6%)	Homemaker or retiree
	8 (25%)	Working
	22 (69%)	Student
Primary Unlock Mechanism	1 (3%)	Password
	2 (6%)	PIN
	2 (6%)	Face Recognition
	6 (19%)	Pattern
	21 (66%)	Fingerprint
Secondary Unlock Mechanism	3 (9%)	Password
	8 (25%)	PIN
	10 (31%)	Pattern
	11 (34%)	None
smart phone usage (mean)	52.7	Estimated daily unlocks
	3.6	Estimated daily usage (h)

Table 8.1: Demographics of the participants of our four week field study (N=32).

All but two participants partially agreed (n=7) or agreed (n=23) that the restriction of access to their smartphone (authentication) was important (5-point Likert scale). Participants self-reported their technical knowledge as high (median=4).

8.4.6 Study Limitations

As participants were self-selected, our sample may not represent the general population. Our simulation might differ from the dynamics when using real implicit authentication systems. Moreover, our prototype added re-authentication on top, whereas a real system could in turn remove the initial device unlock authentication. This might have negatively affected participants' perception of our system. However, the goal was not to evaluate the general concept of implicit authentication itself but indicators for re-authentication.

8.5 Results

In the following report, quantitative results were tested for significance using repeated measures ANOVA with Greenhouse-Geisser correction and Bonferoni post-hoc tests. Ordinal results were tested using a Friedman test with Conover's post-hoc tests. We report significance at the level of $p < 0.05$. No effects of ordering were observed.

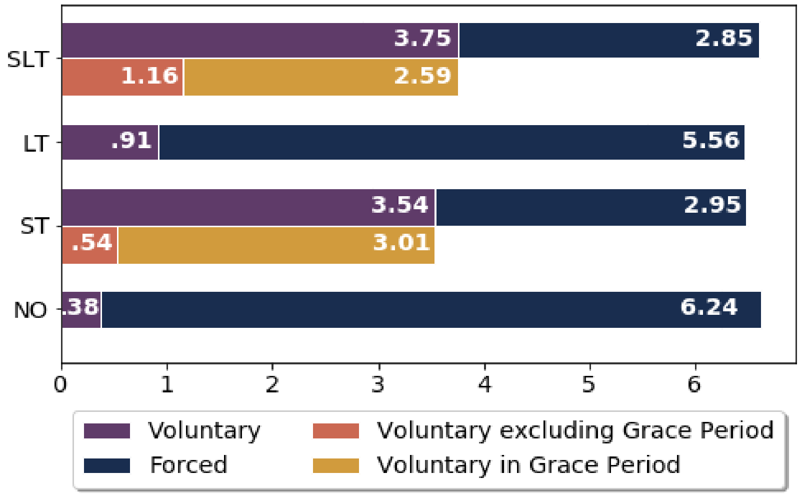


Figure 8.4: Average daily re-authentications by condition. Re-authentications are divided into voluntary and forced re-authentications and voluntary re-authentications are again subdivided into re-authentications during and excluding the grace period (where applicable).

8.5.1 Usage Data

Over the course of the four-week field study, we observed a total of about 3.6 million touches and about 74.200 unlocks (average 84.7 unlocks per day and user) of which 5679 (7.65%) were re-authentications (1910 were voluntary including 646 outside of the grace period).

The *average number of daily re-authentications* per condition is shown in Figure 8.4. We found no effect of the indicators on the average number of daily re-authentications. However, we found a significant difference for the average number of daily *voluntary* re-authentications (F(1.95, 60.44)=14.75, p<.001, η^2 =0.322). Post-hoc tests revealed significantly more voluntary re-authentications for all indicators (p<.04) compared to none (*NO*); and also significantly more for *ST* (p=.001) and *SLT* (p=.003) compared to *LT*.

We also analyzed re-authentications *excluding* those in the grace period, since these are arguably not strictly voluntary: We found a significant difference for relative daily voluntary use, that is, the ratio of voluntary to all re-authentications (F(2.82, 84.53)=59.09, p<.001, η^2 =0.165). Post-hoc tests revealed significantly higher relative voluntary re-authentication for both *LT* (p=.014, Mn=14.56%) and *SLT* (p=.008, Mn=17.63%), compared to *NO* (Mn=5.67%). Relative voluntary re-authentications *during* the grace period were significantly higher (F(1.0, 30.0)=5.01, p=.032, η^2 =0.144) for *ST* (Mn=47.49%) than for *SLT* (Mn=38.93%).

In 49.6% of cases, participants re-authenticated *before* the grace period was over, that is, they did not wait for system-triggered re-authentication (Mn=3.29s, SD=1.46). Outside of the grace period, there was no particular *DCL* at which people preferred to voluntarily re-authenticate (Figure 8.5), but we saw a slight increase below 40%.

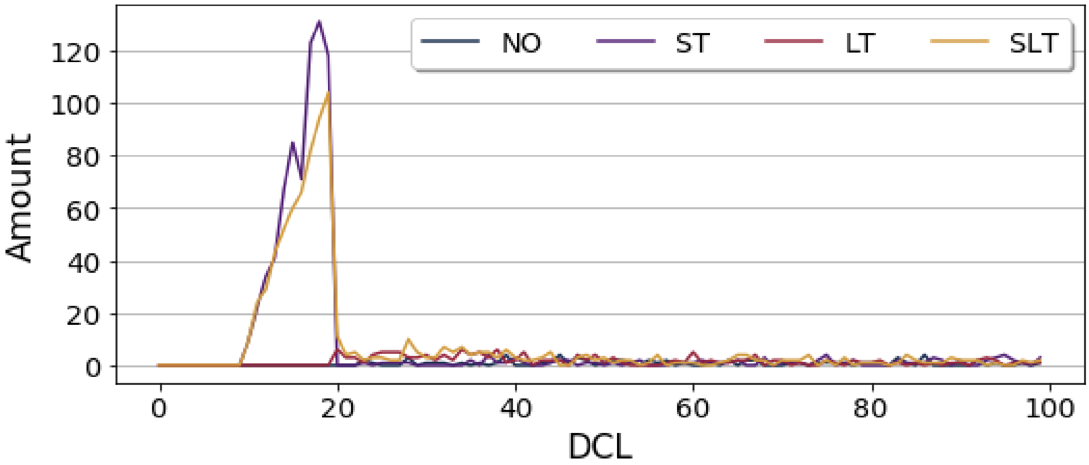


Figure 8.5: Distribution of *DCL* at voluntary re-authentication. There are no re-authentications below 20% for *NO* and *LT* as they had no grace period but instantly locked the device.

In summary, we did not observe an effect of the indicators on the *total* average daily re-authentications. However, *voluntary* re-authentications were more common when using indicators. This can be mainly attributed to re-authentications *outside* the grace period for conditions including the long term indicator and re-authentications *during* the grace period for conditions using the short term indicator.

8.5.2 Experience Sampling

General Results

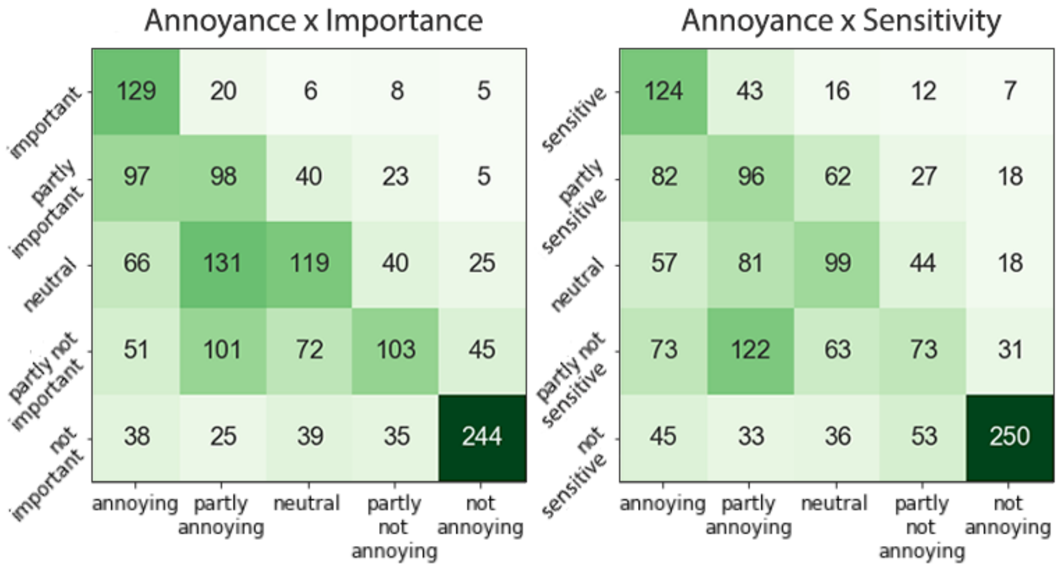
We collected 1557 answers for the experience sampling questionnaires. On a 5-point Likert scale, annoyance was rated neutral *over all conditions* (Mdn=3). The statements that the interrupted task was sensitive and that the interrupted task was important were also rated neutral (both Mdn=3). We could not find a significant impact of indicators on any rating.

Regarding the *authentication context*, participants most frequently reported "at home" for the *place* where they were interrupted, followed by transit and work. The most frequent *tasks* that were interrupted were chatting, reading, searching for information, "nothing"[2] and writing. This aligns with our logged data about the interrupted apps.

Annoyance

We found significant positive (Spearman) correlations between perceived annoyance and importance of the interrupted task (r_s=0.569, p<.001) and between perceived annoyance

[2] This includes both cases where participants actually did nothing in particular or were not interrupted, as the re-authentication was voluntary.



Annoyance x Importance

	annoying	partly annoying	neutral	partly not annoying	not annoying
important	129	20	6	8	5
partly important	97	98	40	23	5
neutral	66	131	119	40	25
partly not important	51	101	72	103	45
not important	38	25	39	35	244

Annoyance x Sensitivity

	annoying	partly annoying	neutral	partly not annoying	not annoying
sensitive	124	43	16	12	7
partly sensitive	82	96	62	27	18
neutral	57	81	99	44	18
partly not sensitive	73	122	63	73	31
not sensitive	45	33	36	53	250

Figure 8.6: Frequencies of reported annoyance by the importance of the interrupted task (left) and by the sensitivity of the interrupted task (right). Color encodes the shown counts.

and sensitivity of the interrupted task (r_s=0.489, p<.001), see Figure 8.6. We could not find effects of the day of the week or the day since the specific condition started.

The annoyance of voluntary re-authentication was perceived as neutral (n=273, Mdn=3), similar to forced re-authentication (n=1277, Mdn=3). The degree to which people were annoyed by voluntary re-authentication did not significantly differ based on whether it happened during (n=76, Mdn=3.5) or outside of the grace period (n=136, Mdn=3). Voluntary re-authentication was labelled as such in the experience sampling in only 18.3% of the cases.

When comparing annoyance for the most frequently reported tasks in the experience sampling, a Friedman test revealed a significant effect of task on annoyance through re-authentication ($\chi^2(5)$=36.16, p<.001, W=0.604). Conover's post-hoc tests found that the interruption of the task "voluntary/nothing" was perceived as less annoying (Mdn=1) when compared to chatting (p<.001, Mdn=4), reading (p=.002, Mdn=3), searching for information (p<.001, Mdn=4), writing (p<.001, Mdn=4) and all other tasks (p<.001, Mdn=4).

In summary, we found that the annoyance caused by an interruption was influenced by a) the sensitivity of the data accessed during the interrupted task, b) the importance of the interrupted task, and c) the task itself, as the reported task "voluntary/nothing" was perceived as less annoying.

8.5.3 Weekly Questionnaires

Voluntary Re-authentications

For the weekly questionnaires we found significant differences for the motivation to voluntarily re-authenticate ($\chi^2(3)$=10.05, p=.018, W=0.498) and the feeling of reward by an increased *DCL* after re-authentication ($\chi^2(3)$=21.74, p<.001, W=0.618) with regards to the different indicators. Post-hoc analysis revealed that for *SLT* (Mdn=3) participants felt significantly more motivated to voluntarily re-authenticate than for *NO* (Mdn=1, p=.009). For all conditions using an indicator participants felt significantly more rewarded (Mdn_{ST} =2, Mdn-LT=2, Mdn_{SLT} =3) than in the *NO* condition (Mdn=1, p<.02). We found no significant differences on *perceived annoyance* of the system.

Thus, while we cannot provide evidence for a general effect of our indicators on the annoyance, we did find a positive influence of the long term indicator on the motivation to voluntarily re-authenticate. The feeling of being rewarded for re-authentication by the increased *DCL* was also significantly higher for the conditions including the long term indicator.

Perception of Indicators

Participants liked about the indicators that interruptions were less sudden compared to no indication (mentioned by 22 people) and that the *DCL* was visible at any time for the conditions with a long term indicator. In the *NO* condition, participants liked that re-authentication was fast (9 mentions). The gradual darkening was positively mentioned by ten participants for *ST* and by eight for *SLT*.

Interrupts were perceived as sudden by fifteen participants in the *NO* condition and by ten, four, and three participants in the *LT*, *ST*, and *SLT* conditions, respectively. Seven participants reported they overlooked the *DCL* visualization in the *LT* condition. Interrupts were in general perceived as annoying in all conditions (mentioned by 10, 9, 7, and 8 participants for the *NO, ST, LT* and *SLT* conditions, respectively).

8.5.4 Final Questionnaire

Ranking

In the final questionnaire, participants were asked to rate their experience with the system in general. The *overall ranking* of the different conditions (Figure 8.7) reveals that the combination of both *long term* and *short term* was preferred. No indication (*NO*) was ranked last. Long term (*LT*) and short term (*ST*) ranked second and third. Based on the open questions, the following reasons contributed to their choice: Sixteen participants stated to not like the sudden interruptions without indication. The combination of both short and long term (*SLT*) was particularly liked for the best overall overview and control and the continuous visualization of the *DCL* (10 and 9 mentions).

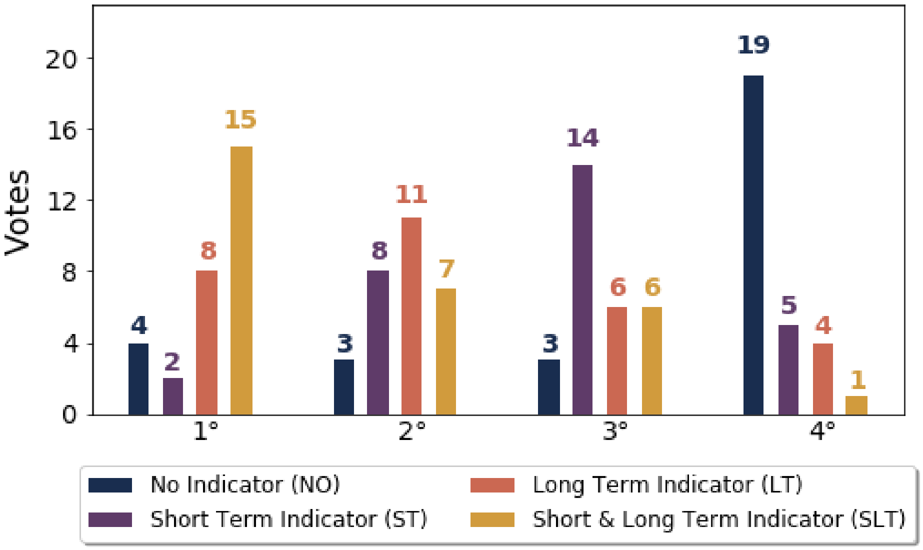


Figure 8.7: Participants' ranking of the different indicators. The combination of long- and short term indicators was the most preferred method while no indication was least preferred.

General Perception

As a response to our Likert scale questions, participants did not find vibration and notifications particularly annoying (Mdn=2). They felt neutral towards being stressed by the dimming during the grace period (Mdn=3). The long term taskbar symbol was considered to be helpful (Mdn=4) to predict re-authentications.

Participants remained neutral (Mdn=3) towards a possible influence of the system on their behavior. They partly liked the design (Mdn=4) and partly disagreed with being negatively influenced by bugs (Mdn=2). They felt neutral (Mdn=3) about the experience sampling being annoying or influencing their behavior or perception.

No one had profound knowledge about implicit authentication before the study nor did they review implicit authentication from other sources than the material provided by us (Mdn=1). There was general agreement on having read the introduction on the website and having watched the whole introductory video (Mdn=5).

In general, participants agreed to always locking their device (Mdn=5) and to authentication interrupts being more annoying than traditional authentication up front (Mdn=5). Regarding whether they would use the concept of implicit authentication in general, participants remained neutral (Mdn=3; 10 agreed or partly agreed, 5 neutral, and 14 disagreed or partly disagreed).

Finally, people would have liked a slightly longer grace period. On average they suggested 10.14 s (range 2 s–60 s).

8.6 Discussion

Here we give an overview over our findings and discuss their implications.

8.6.1 Importance & Sensitivity

While we did not find a significant effect of indicators on perceived annoyance via experience sampling, we gained related evidence and insights: We found a significant impact of *sensitivity* and *importance* of an interrupted task on the perceived annoyance. This was also pointed out in the final interviews where five of the eight participants found the system interrupting an important or stressful task to be a particularly negative event: "*I remember when I had to make a really important call and my screen was locked before I could do it. I had to answer the feedback, too, before I could finally call. Then, it was really annoying, but usually the interrupts were no problem.*"

As a key insight, the situations in which participants perceived interruptions as annoying were also those that they rated as sensitive, hence, those that would require increased protection when relying on a real implicit authentication system. It might be possible that users were biased as they knew their phone was protected by their primary locking mechanism anyway in this study. Nevertheless, we believe that this topic should be investigated further.

8.6.2 Voluntary Re-Authentication

In contrast to related work on general interruptions [178], we could not find a positive effect of deciding when to re-authenticate on reducing annoyance. For the grace period, one explanation is that participants might not have perceived the option to re-authenticate as voluntary (as re-authentication was inevitable). More generally, our results on importance, sensitivity, and interrupted tasks all indicate that for our participants annoyance was mostly determined by the interrupted activity and not by whether re-authentication was voluntary or not.

Nevertheless, voluntary re-authentications were mentioned as positive in open comments and the interviews, and indeed accounted for a considerable proportion of 33.6% of re-authentications (11.4% excluding grace period). Moreover, users felt significantly more motivated to re-authenticate for the combined short and long term indicator. All indicators also resulted in significantly more common use of voluntary re-authentications.

Hence, a promising approach to reduce user annoyance might be to investigate concepts that provide options for users to voluntarily re-authenticate with awareness of current activities. For instance, one person suggested allowing for voluntary re-authentication when opening an app, which often coincides with the beginning of a new activity.

8.6.3 Grace Period

We received mixed feedback on the grace period. Many participants liked it, in particular the more predictable nature of the interruption. For example, one participant said: "*The more sudden the interruption happened, the more annoyed I felt about it. Surprisingly, it did not depend so much on the frequency of the interrupts. It only depended on the announcement.*"

However, some participants complained that they could not use the grace period to its full extent due to light conditions and wished for a longer duration. Others used our introduced option to voluntarily re-authenticate before the device was locked. In general, the desired length was very different amongst the participants which implies that an option to customize this (as also suggested by Agarwal et al. [5]) might indeed be promising for future work. We also believe that there is an impact of the personal *usability-security trade-off*, as having a (longer) grace period also implies a security risk in cases where an attacker would get hold of the device. Steps to address this might be, e.g., adapting the length of the grace period to the derivative of the *DCL* (i.e., the strength of change in system confidence) or the importance of the interrupted app.

In general, we see the approach of gradually dimming the screen only as a first step. Moreover, as proposed by participants of our focus group, future systems could, for example, use biometrics for re-authentication. In this case, dimming the screen could be an indicator for the user to present their face to the camera or quickly put their finger on a fingerprint scanner and thus avoid a full context switch.

8.6.4 Interruptions

Based on the previously discussed results, we present three recommended aspects to consider with regard to scheduling re-authentication interrupts.

1. **Sensitivity of the task:** If the user is accessing non-sensitive data (e.g., while reading a book), an upcoming re-authentication could be delayed or triggered when the task is finished, as suggested by related work [3, 18] and done in practice[3]. However, while accessing sensitive data (e.g., banking app), re-authentication should be triggered instantly to restrict further access.

2. **Importance of the task:** As users found interruptions of important tasks particularly annoying, selectively delaying such interruptions could improve users' experience with the system. This assumption is further supported by Adamczyk and Bailey [3, 18].

[3] e.g., Smart Lock: `https://support.google.com/android/answer/9075927?hl=en`, last accessed October 16, 2024

3. **Recent changes in confidence:** Changes in device confidence level (*DCL*) over time may be used as an indication of the necessity of an immediate interruption. While a sudden decrease in confidence most likely corresponds to an intruder taking hold of the device, a slow decrease is more likely to be caused by natural variations in the legitimate user's behavior. However, those assumptions are, as of now, speculative, and further research with a functioning implicit authentication system is necessary to verify this hypothesis.

The focus of our work was on interruptions caused by a continuous authentication system. Some lessons learned may generalize to other interruptions, such as notifications. A further factor to consider in that case is the importance of the interruption itself – which we assumed to be high for implicit authentication due to the security risk.

8.6.5 System Design

For our study, we introduced a novel method to more realistically simulate an implicit authentication system. Our approach extended previous approaches (e.g., Khan et al. [157]) and made some of our evaluations, like the *long term* indicator, possible in the first place. We believe this to be a valuable step to enable future evaluations but also acknowledge that using our system has limitations. In particular, as the system was touch-based we introduced a bias towards interrupting tasks that used many touches, such as writing, whereas very short interactions were interrupted less. One way to address this would be to track the current app and schedule interrupts to distribute re-authentication requests equally over the different tasks. Due to our use of a simulated system, we were also not able to remove the primary unlocking mechanism, as this would have left participants unprotected.

However, our results from the final questionnaire suggest that neither the system itself nor the introduced experience sampling had a major effect on participants' perceptions or behavior. Furthermore, vibration feedback and notifications were not perceived as annoying, and the overall design was rated as very positive.

8.6.6 Adoption of Implicit Authentication

Our participants remained neutral towards using implicit authentication and only 10 of 29 agreed or partially agreed to wanting to use it. This contrasts with results of previous studies: Crawford and Renaud [62] report 90% of their participants to be interested in adopting implicit authentication. Participants also generally agreed that re-authentication was more annoying than unlocking up front.

Possible reasons could be that users underestimate the actual number of authentications they perform (on average by 38% in our study) and the accompanying benefit of implicit authentication. Other explanations include authentication overhead of a simulated system or habituation to users' traditional unlocking methods. On the other hand, studies from related work were a lot shorter (several lab studies [5, 62, 218] and shorter field studies[157]) and thus user perception in our study developed over a longer period of time (e.g., we potentially observed a lower novelty effect). Moreover, effortless fingerprint authentication in particular has become an established method in the years between some of the earlier related work and our study, potentially shifting users' views.

As a next step, we suggest evaluations with a functional implicit authentication system for a more realistic scenario. In cases where such a system cannot robustly provide sufficient security, conducting the study with users who do not lock their phones anyway might be an option. Targeting this user group has also been suggested as a major application area for implicit authentication in related work [156, 292].

8.6.7 Research Questions

Regarding our initial research questions, we found all our indicators to be preferred to no authentication. We found no effect of indicators on annoyance. Annoyance was rather determined by the interrupted activity (RQ1). We found sensibility, importance, and the specific interrupted task to be further factors influencing the perceived annoyance of interrupts (RQ2). We also found all indicators to have a positive effect on the use of voluntary re-authentications (RQ3). Finally, we found that users felt particularly motivated to voluntarily re-authenticate by combined short and long term indicators. They overall perceived voluntary re-authentication as positive and used it to a considerable extent (RQ4).

8.7 Implications

In this chapter, we introduced a biometric interface by providing a Long Term Indicator of the current system state, a Short Term Indicator preparing users for an upcoming re-authentication, and the option to voluntarily re-authenticate to avoid interruption.

From the results of our four-week field study (N=32), we found that both indicators were preferred to having no indication. The importance and sensitivity of the interrupted task had a strong influence on annoyance while we did not find significant effects of the indicators. Voluntary re-authentications were perceived as less annoying and were more prominent in all indicator conditions. While they were mostly used to bypass the grace period we also found increased voluntary re-authentications when only showing users the system state. Our work thus both highlights the value of providing insights into the state of continuous models and the importance of considering user needs when designing interventions. Future work may consider delaying re-authentications depending on the task or explore other nudges for users to confirm their identity already at higher confidence levels. Participants felt rewarded for re-authenticating when they could see the score increase, implying that gamification could be a viable approach for this.

IV

Biometric Interfaces to Support User Agency

PART IV: BIOMETRIC INTERFACES TO SUPPORT USER AGENCY

Literacy alone is only of limited value to users of biometric systems if they cannot use it to take agency over their authentication and data. However, this can be a challenging task, in particular for biometrics, where no natural (user) interfaces exist so far. In this part, we take a depth-first approach to creating biometric interfaces for such a case, namely biometric authentication using keystroke dynamics. In this part, we show, that users are able to actively control their typing behavior and present a game to support them in doing so. We conclude with a prototype that allows for externally influencing behavior and thus reducing user effort. This opens up possibilities for users taking agency over if and when to be recognized.

- ❖ **Chapter 9** proposes a visualization of typing behavior that we leveraged in a lab study to show that participants could alter their typing behavior towards a given target.
- ❖ **Chapter 10** extends this lab study by developing a game around the task of modifying typing features and showing differences between a lab and a remote setting when playing our game.
- ❖ **Chapter 11** explores the use of electromagnets to influence user typing and to free users from having to actively control typing features themselves.

9

Exploring Intentional Keystroke Control

This chapter is based on the following publication:
Lukas Mecke, Daniel Buschek, Mathias Kiermeier, Sarah Prange, and Florian Alt. 2019. *Exploring intentional behaviour modifications for password typing on mobile touchscreen devices.* In Fifteenth Symposium on Usable Privacy and Security (SOUPS '19) [181]

In the previous part of this thesis, we explored how biometric interfaces can be leveraged to communicate information about the model, its state, and influencing factors to a user to support their literacy in using their biometric authentication. In this part, our focus is on additionally giving users agency over their authentication.

As a concrete example, we here explore how users can gain control over authentication using keystroke dynamics, i.e. user identification based on features of their typing like rhythm [262], finger placement [42], and other related features [38, 263, 292]. Even if attackers gain knowledge of a password, they also have to enter it with the same behavior as the legitimate user to gain access. The underlying assumption of such behavioral biometric authentication systems is that humans differ *implicitly* in how they type.

In this chapter we present the first systematic exploration of a fundamentally different view: We study users' ability to *explicitly* modify commonly utilized biometric features of their typing behavior. Our goal in this chapter is not to design a new authentication system but to better understand users' fundamental ability to control their typing behavior. As keystroke dynamic systems are naturally designed to run transparently in the background, this ability lays the foundation for users to be able to take agency over this type of recognition.

Better understanding the ability to intentionally modify interaction behavior is vital considering the growing number of biometric systems, as illustrated by the following use cases:

1. *Extending the password space:* Instead of only using different characters to compose a password, each character could be entered in a different manner. For instance, although both use the same eight characters, "password" is different from "pass[*hold long*]word", where the user keeps the second "s" pressed for longer than usual.
2. *Recovering from a leak of behavioral data:* A leak of behavioral information usually implies that this biometric can no longer be used if we assume that behavior is unchangeable. However, this is worth challenging. As an analog example, some people decide to intentionally change the way they write their signature. Similarly, it might be possible to intentionally change, for example, password typing behavior features to recover from a leak to be able to continue using this biometric.
3. *Hiding identity and private information:* Keystroke dynamics can not only be used to identify people but also give insights into other features like their gender [91, 109] or emotional state [90] they may want to keep private. Modifying typing in untrusted environments like on the web or when using an unknown device can thus help users protect their privacy and identity.

In all these examples, users have reasons to intentionally modify aspects of their behavior that they do not need to control for the underlying input method (e.g., typing rhythm does not matter for entering an email). Prior work on intentional changes of typing behavior has exclusively studied this ability for attackers with technical support [23, 158, 159] or for limited features in desktop settings without changes and learning over time [54, 133, 134]. Thus, it still remains unclear to what extent users can control and modify fundamental biometric features of their mobile touch typing behavior.

Following the approach of this thesis, the first step in that direction is supporting users in gaining literacy about their own typing. To facilitate this we designed different visual text annotations to communicate typing behavior and conducted a prestudy (N=144) to find the one that was most clear. To explore users ability to intentionally modify their typing we then conducted a lab study (N=24) where participants entered given passwords with such modification instructions on a smartphone in two sessions a week apart.

Our results show that users can successfully control and modify typing features (flight time, hold time, touch area, touch-to-key offset) given our visualization. Modifying multiple features was significantly more difficult, in particular, if they were distributed over the input instead of being co-located on a single key and if temporal features were involved. We discuss influencing factors on users' success in changing their behavior, implications for the usability and security of mobile passwords, such as informing behavioral biometrics for password entry, and extending the password space through explicit modifications.

> In this chapter we contribute 1) visual text annotations to communicate typing behavior modifications, developed in a prestudy (N=114). 2) A lab study (N=24) using this scheme to investigate intentional modifications for different features and their combinations, for password typing on smartphones in two sessions a week apart. 3) A discussion of implications for mimicry attacks, research on behavioral biometrics, and usable passwords with intentional modifications.

9.1 Related Work

In this section, we relate our work to research on keystroke biometrics and mimicry attacks. These areas motivate our investigation of intentional modification of typing features and our choice of the specific features we studied.

9.1.1 Keystroke Biometrics

Our work is related to keystroke biometrics (or "keystroke dynamics"), which describe users' individual behavioral characteristics when entering text on a keyboard. This information can be used by the system to identify users, for example, to protect accounts, devices, and data. A rich body of related work examined this idea first for typing on physical desktop keyboards (for example, [195, 196]; survey [262]), then on early mobile phones with physical keys (for example, [35, 49, 57, 133, 149, 173, 303]). More recent work investigated keystroke biometrics for on-screen typing on smartphones (for example, [41, 42, 81, 292]; recent survey [263]), including keyboards operated via gestures instead of tapping [38].

For entering passwords in particular, recognizing users based on *how* they enter the secret word provides an extra (implicit) layer of security [42], for example, to protect against cases in which the attacker got to know the password via shoulder surfing [234], smudge [17, 281] or thermal attacks[2].

Due to the origin of keystroke biometrics on physical desktop keyboards, the most commonly used typing behavior features are temporal [262]: Users' typing is characterized by their typical *hold times* (i.e., time between key down and up event), and *flight times* (i.e., time between key up and down on the next key). Mobile touch devices offer further spatial features, such as touch area and offsets between touch locations and key centers. Offsets, in particular, showed higher biometric value, that is, they facilitated more accurate distinction of users [41, 42]. Related work motivates our choice of features: hold time, flight time, offsets, and touch area.

In summary, related work on typing behavioral biometrics used features as they occur "naturally" as an *implicit* part of typing. Our work is fundamentally different: We examine these typing features as *explicit* and *actively controlled* by users, for example, to increase the password space. In particular, we study how well users can indeed control these features when entering passwords on a smartphone.

9.1.2 Mimicry Attacks

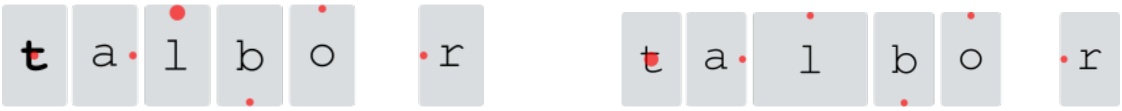
Attacks on keystroke biometric systems can be performed either automated or manually. Automated attacks use generative models to synthesize forgeries from observed data and were shown to be effective against handwritten signatures [23] and keystroke dynamics on a PC [193, 211, 244]. Some work also tested such attacks when proposing a new keystroke biometric system. For example, Stefan et al. found their system resistant to inputs generated from a first-level Markov model [256].

The most commonly considered attack on behavioral biometric systems is the so-called *mimicry* attack: Here, an impostor tries to manually reproduce (mimic) the (known) behavior of a legitimate user to gain access.

As a simple case, a *zero-effort attacker* model evaluates a biometric system against natural behavior collected by other users who did not intend to actually bypass the system. While this model has been commonly used to evaluate the vulnerability of behavioral biometric systems, related work found that it underestimates attack success [23, 211]. This calls for evaluations with means for more skilled and targeted attacks.

To support attackers in launching successful mimicry attacks they need to know the behavior to imitate. In the case of handwritten text, for example, this could be a sample signature. Researchers mounted successful mimicry attacks against touch input behavior [158], keystroke dynamics on a PC [265], and keystroke dynamics on mobile phones [159]. Key to those attacks were systems that both visualize the target behavior and provide the attacker with feedback on their attempts. For example, Khan et al. [159] used augmented reality using a phone's camera to show visual cues on top of its view on another phone's keyboard. This guided correct timing and touch behavior. In another approach, they used audio stimuli to guide the timings.

In summary, prior work used representations and active modifications of typing behavior to support mimicry attacks. In contrast, we aim to better understand the human ability to control mobile typing behavior per se.



(a) **‘Bold Letter’** using a bold font to indicate large touch area and circle size for hold time. Circle location shows offset, gaps indicate flight time.

(b) **‘Long Key’** using circle size for touch area and key width for hold time. Circle location shows offset, key gaps indicate flight time.

Figure 9.1: Main design candidates for visualizing target feature values for studying intentional behavior modifications. Both were evaluated in our prestudy. Based on the results we decided to use the ‘*Long Key*’ concept for our main study.

9.2 Visualizing Typing Behavior Modifications

For users to be able to produce targeted changes in their typing we need a common language to communicate such behavior modifications. This also serves as a tool for users to explore and understand how changes in their typing translate to feature changes in the input to a biometric model. Here we describe our approach to finding such a visual representation.

9.2.1 Selection of Features

There are a multitude of possible features that can be used for biometric authentication in the context of mobile touch interaction. An extensive list was compiled by related work [42] and covers 24 spatial, temporal, and contact features. Khan et al. [159] found this extensive feature set hard to simultaneously control for their mimicry attack. They thus removed highly correlated features, resulting in a set of six: key hold time, flight time, down pressure, down area, down x, and down y.

We combine x and y together as touch offset. Furthermore, pressure and area were highly correlated on our test devices, since most Android phones[1] estimate pressure from the area. We thus decided to omit pressure and use the area directly.

To sum up, we decided to study a set of four features, namely *touch area*, *flight time*, *hold time*, and *touch-to-key-offset* with the latter being two-dimensional (x, y).

9.2.2 Visualization Design

We developed several designs that communicate modifications of the four features to instruct participants, for example, to perform a long key press for the second character in a password. We first tried simple markup (e.g., p̣– . **ȧ**s . . sẉ–ȯ**r** . d—) but found this representation to become cluttered quickly and to offer very limited expressiveness.

[1] We used LG G6 phones in our study.

We thus chose a pictorial approach: We showed letters with a key metaphor to visualize behavioral changes (Figure 9.1). We explored a range of possible visual features, including offsetting the key or its label, writing bold or italics, and using underscores and colored dots.

We narrowed the options down to two final designs (see Figure 9.1). Both used whitespace gaps between keys to indicate flight time and a red dot to indicate touch offset. One variant ('Bold Letter') visualized a larger touch area by rendering the key in bold and using the size of the offset dot to represent hold time. The other ('Long Key') used the size of the dot to visualize the touch area, and key width to show hold time. While 'Bold Letter' resulted in a more compact format, 'Long Key' unified both temporal features on a shared axis (time flows from left to right).

9.2.3 Online Survey

We conducted an online survey to determine our final design.

Survey Design and Procedure

To assess the intuitiveness and readability of our designs, we created an online survey that showed example passwords with visualized modifications. Participants had to indicate which parts of the visualization were used to encode which behavioral cues, without prior explanations. People did this for both designs in counterbalanced order. Afterward, they were asked to rate on a 5-point Likert scale how intuitive and readable they found the two visualizations.

The survey was distributed over a university mailing list. It took 5 minutes to complete. Participants had a chance to win a 10 € gift voucher.

Results

A total of 114 participants answered our survey (56 % female; mean age 27 years, range 18 to 63 years). Both *offset* and *flight time* were correctly interpreted by 90 % of the participants for both designs. *Area* and *hold time* were correctly interpreted by 81 % and 82 % in the 'Long Key' condition, respectively. However, these two features were only correctly interpreted by 50 % and 51 % in the 'Bold Letter' condition. 'Long Key' was rated as more intuitive (Mdn = agree, Mdn_{bold} = neutral) but 'Bold Letter' was rated to be more readable (Mdn = strongly agree, Mdn_{long} = agree). When asked for their preferred method, 59 % of the participants reported the 'Long Key' notation while 39 % voted for the 'Bold Letter' visualization. The rest had no preference.

9.2.4 Final Visual Representation

We decided to use the 'Long Key' visualization: It has the advantage of encoding temporal features on a shared axis and all features allow for continuous representation of values (in contrast to the binary bold letter).

In conclusion, we used the following visual encoding shown in Figure 9.1b: *Touch-to-key-offset* is marked by a red dot at the position where the key should be touched. *Flight time* is represented by a gap between two key shapes that scales with duration. Analogously, *hold time* is represented by scaling the width of the key rectangle with duration. Finally *touch area* is visualized by the size of the red dot used for offset (larger size indicates larger area).

9.3 Study

9.3.1 Study Design

As our study design is quite complex, the following subsections each explain one main component. The most complex one is *task*, which is given both as an overview and in detail.

Passwords

In general, participants had to repeatedly enter given *passwords* ("football", "princess", "password"). While these three are obviously not great passwords in terms of security, we selected them since they have comparable properties and are common passwords[2]. Moreover, they do not require switching keyboard mode (e.g., between characters and symbols), which we wanted to avoid as a simplification for this first investigation. Similarly, we favored simple passwords to ensure that task difficulty was mainly determined by behavior variations and not affected by memorability or search time for rare symbols.

Features

We studied the intentional modification of four features: *touch-to-key-offset* (on five levels: center/left/right/top/bottom), *flight time* and *hold time* (both on two levels: default/long), as well as *touch area* (on two levels: default/large).

Tasks

Participants solved 37 *tasks*, each using one of the three passwords. The tasks differed in various aspects described below. While the design is complex, the overall goal was to cover six aspects, namely (1) different *passwords* with (2) different *feature modifications* at (3) different *locations* within each word. We also include (4) different *combinations* of features that are modified in the same password, either (5) at the same character/keypress (we call this *co-located*) or (6) *distributed* across several characters/keypresses within the word.

We iterated the task design several times by means of prestudy runs with two to three people in each version. We gradually narrowed the tasks down to an acceptable study duration of

[2] `https://teampassword.com/blog/worst-passwords-2024-password-security-tips`, last accessed October 16, 2024

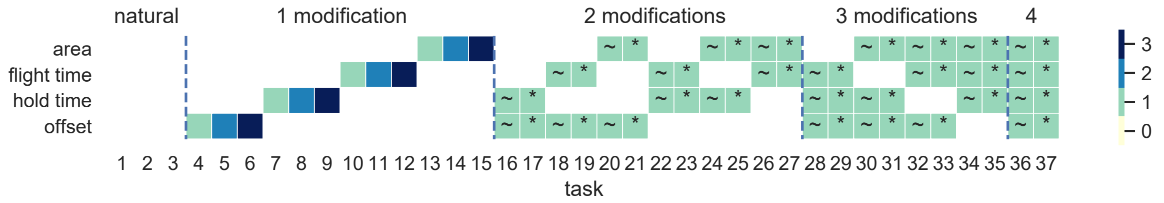


Figure 9.2: Overview of the tasks in each session. In the beginning (tasks 1–3), participants were asked to enter the passwords naturally, afterwards (tasks 4–15) a single feature had to be modified with an increasing number of occurrences (color of the cell). Thereafter, two (tasks 16–27), three (tasks 28–35), or four (tasks 36, 37) features had to be modified at once. All possible feature combinations were tested and features were either *distributed* (~) over the password or *co-located* (*) on a single key.

one hour. In full detail, the tasks used in the main study were structured and designed as follows (Figure 9.2):

Natural tasks (1–3): The first three tasks simply asked people to enter each password six times without presenting any intentional behavior modifications.

Modifying a single feature (tasks 4–15): In each of these tasks participants had to modify one feature (e.g., hold time). There were three such tasks per feature, namely one per password (i.e., 4 features × 3 passwords = 12 tasks). Across the three tasks per feature, all feature levels occurred at least once, while covering different locations: The first task per feature modified the 2nd character of the password, the second task modified the 2nd and 7th characters, and the last task modified 2nd, 4th, and 7th characters. The assignment of passwords across these tasks was counter-balanced, such that modifications overall occurred in all passwords at all locations.

Modifying two features (tasks 16–27): In each of these 12 tasks people modified two features (for example, hold time *and* flight time). There were two tasks per combination of two features: The first had one modification on the 2nd character and the other on the 3rd (i.e., *distributed*). The second task had both modifications on the 7th character (i.e., *co-located*).

Modifying three features (tasks 28–35): In these eight tasks, participants had to modify three features, with two tasks per combination of three features: The first had modifications on the 2nd, 4th, and 7th character (*distributed*). The second one had all three modifications on the 5th character (*co-located*).

Modifying four features (tasks 36 and 37): Finally, participants had to modify four features: The first one had modifications on the 2nd, 4th, 6th, and 8th character (*distributed*), the last had all modifications on the 5th character (*co-located*).

The task order was not randomized, in favor of gradually increasing the number of modified features per password, which we suspected to have an influence on task difficulty.

Sessions

The whole procedure was repeated two times, in two sessions about a week apart. In this way, we observed the typing behavior of each participant at two points in time.

Summary

For the following report of our data analyses and results, it is useful to think of our study design as follows:

Tasks 1–3 are used to analyze natural (i.e., unmodified) behavior, while the other tasks are used to analyze user behavior when modifying the four behavior features.

Note that from task 16 onward (i.e., all tasks with feature combinations), our study is a typical repeated measures design with *number* of modifications (2, 3, 4) × *distributed* multiple modifications (distributed, co-located) × *session* (1st, 2nd). We use this for typical ANOVAs to study in particular the impact of modification of multiple features.

9.3.2 Apparatus

We developed an Android app that controlled the study process (e.g., counterbalancing, task progression, explanations).

The values used for scaling our visualizations (e.g., default flight time for default key gap) were informed by prestudy experiments and related work [41] (flight time 260 ms normal, 1000 ms long; hold time 80 ms normal, 300 ms long; area 0.2 normal, 0.4 large, unitless as reported by the Android API; offset x ±40 px, offset y ±70 px). To avoid visual clutter, we limited the scaling to minimum and maximum threshold values, beyond which the visualization did not change.

We integrated a modified version of the Android open source project LatinIME[3] keyboard. This enabled us to log all typing events and touch features. To reduce distraction, we disabled the context menu for special characters shown on the long press. In addition, our study app logged the expected key and behavior modifications, as well as the current user and task for each keystroke.

9.3.3 Procedure

Upon arrival, participants were introduced to the goal of the study and asked to sign a consent form. After an initial demographics questionnaire, they performed the tasks (see Figure 9.2) as described in section 9.3.1 on our test device. We asked participants to enter passwords with their right thumb to keep results comparable.

[3] LatinIME: `https://android.googlesource.com/platform/packages/inputmethods/LatinIME/`, last accessed October 16, 2024

When first confronted with a new type of modification, participants got a short explanation of what to do and prior to every task, they had the option to train entering the password. Except for the tasks without modifications (natural tasks), they were provided with real-time feedback, using our visualization, to show their behavior next to the expected one. Every task had to be completed successfully six times and without feedback. The number of attempts was not limited.

Each task was followed by a short Likert questionnaire containing the statements: 1) *"I was able to adjust to the specified behavior."*, 2) *"I was successful in completing the task."*, and 3) *"The task was difficult for me."*.

After completing all tasks, participants were asked to come up with a modified password on their own and could take notes to remember it. The same process was repeated in the second session, excluding the initial demographics questionnaire. Creating a custom password was replaced with recalling and performing the password from the previous session. After the second session, we conducted a short interview. Sessions were scheduled one week apart.

9.3.4 Participants

Study invitations were distributed over a mailing list of our local university. Requirements were right-handedness and familiarity with typing on mobile phones. We recruited a total of 24 participants (14 female; mean age 27 years, range 14 to 54 years). Half of the participants were in their twenties. 58 % were students, 30 % were employed, and the remaining ones were in school. Participants were compensated with €20 for completing the whole study.

9.4 Results

Significance tests were conducted using ANOVA with Greenhouse-Geisser correction and Bonferoni-corrected post-hoc tests (significance at alpha level $p < 0.05$). If not reported otherwise, data for analyses is aggregated for both sessions.

As a first overview, we report key descriptive measures: The grand mean task completion time across all tasks (i.e., completing all six successful password entries of a task) and participants was 38.3 seconds. For typing speed, the grand mean was 28.7 words per minute (WPM [288]). The grand mean of the number of incorrect entries per task was 1.74.

We report on participants' natural typing behavior (Section 9.4.1), their ability to modify it (Section 9.4.2), and their accuracy in doing so (Section 9.4.3). We analyze the effect of multiple simultaneous modifications (Section 9.4.4) and the impact of modifications on the individuality of behavior (Section 9.4.5). We conclude with details on technically detecting modifications (9.4.6) and participant feedback (Section 9.4.7).

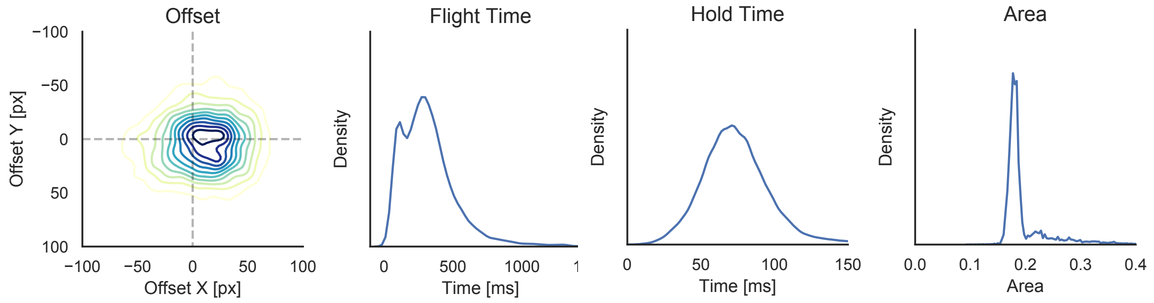


Figure 9.3: Overview of participants' natural typing behavior (i.e., typing without being presented with any modifications), as measured in the first three tasks of each session.

9.4.1 Natural Behavior

We first report on "natural" behavior – typing *without* any modification instructions (tasks 1–3). Figure 9.3 presents the results. They match our expectations based on related work:

Touch offsets are slightly shifted to the lower right, as typical for input with the right thumb [40]. Moreover, median flight time (290 ms) and hold time (72 ms) are in line with related work [41] and close to the ones we chose as defaults for scaling key width and gaps in our visualization (flight time 260 ms, hold time 80 ms). Thus, our chosen values indeed matched people's natural behavior.

Touch area significantly correlated with the x location of the target key (r=-0.252, p<.001): Due to thumb stretching, typing keys on the left of the keyboard resulted in a flatter thumb posture and thus a larger touch area. Flight time showed a main and secondary peak (Figure 9.3). The latter was caused by zero finger travel distance for "double letters" (e.g., pa*ss*word).

9.4.2 Ability to Modify Behavior

Figure 9.4 visualize the distribution of the behavioral features for different *target values*, i.e., expected feature values shown by our visualization. Next, we report on statistical tests comparing these distributions per feature (see Table 9.1). Here we report on the post-hoc tests and further details:

For all features, post-hoc tests showed that directions of differences were as expected (e.g., offset significantly further to the left for *left*, flight time significantly longer for *long*).

For vertical offset and flight time, the interactions of session and target were significant (see Table 9.1), yet the small effect sizes and visual inspection of descriptive plots indicated that this was too tiny to warrant meaningful interpretation.

In summary, the significant results of these statistical tests confirm the "big picture" visible in Figure 9.4: For all features, people significantly modified their behavior in the direction indicated by our visualization.

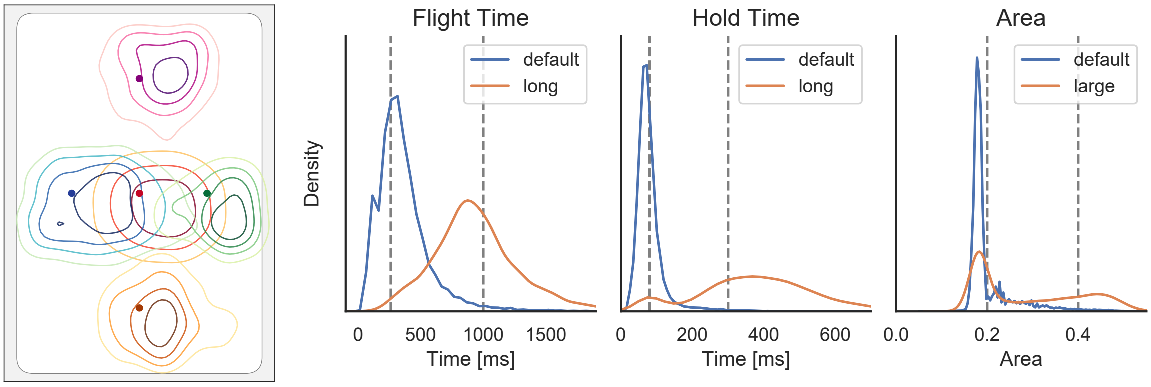


Figure 9.4: Overview of participants' modified typing behavior across both sessions. Overall, this figure shows that presenting modifications via our visualization provoked clear differences in the typing features. For offset, the rectangle indicates key borders. Vertical lines/dots indicate the target values.

Feature	Measure	target	session	target * session
Offset	absolute x	.777^a		
	absolute y	.890^a		.015^c
	relative (error)	.082^b		
Flight time	absolute	.785^a		.010^c
	relative (error)	.332^a	.038^b	
Hold time	absolute	.848^a		
	relative (error)	.624^a		
Touch area	absolute	.737^a		
	relative (error)	.930^a		

a: p < .001, *b*: p < .005, *c*: p < .05, empty cells not significant

Table 9.1: ANOVA results for ability (1) to modify behavior (absolute, Section 9.4.2) and (2) to replicate target feature values (relative i.e., error, Section 9.4.3). The last three columns show the effect sizes (ω^2) for *target* value (i.e., the feature value communicated via our text annotation), *session*, and their interaction. See the text post-hoc test results.

9.4.3 Ability to Replicate Target Feature Values

The previous section investigated differences in absolute feature values. It is also interesting to analyze how *accurately* people were able to replicate modifications. To this end, Figure 9.5 visualizes the distribution of participants' errors when reproducing the target values indicated by our visualization for each feature. Table 9.1 summarizes the ANOVA results.

For *offset*, post-hoc tests revealed errors to be significantly smaller for the target *right* compared to *left* (p=.010, d=-.773), *top* (p=.008, d=-.783), *bottom* (p=.011, d=-.765) and *default* offset (p=.027, d=-.685).

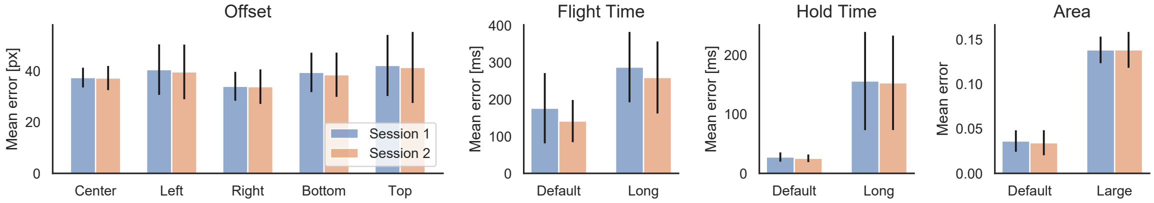


Figure 9.5: Observed derivation of participants' behavior from the target values of the given modifications for both sessions. Participants were generally better at reaching the target value for the default level. For offsets, the lowest error occurred for touches to the right, since this coincides with natural thumb offset [40]. In contrast to the other features, flight time accuracy increased from the first to the second session, indicating a learning effect.

Measure	number of mod.	**session**	**distributed**	number * distributed	session * distributed
Offset error					
Flight time error	.93^a	.017^b	.109^a	.023^c	
Hold time error	.166^a		.178^a		
Touch area error	.039^a			.018^a	
Task compl. time	.032^c	.015^c	.224^a	.039^c	
Typing speed	.172^a		.232^a	.079^a	.002^c
Incorrect entries			.114^b		

a: $p < .001$, *b*: $p < .005$, *c*: $p < .05$, Empty cells not significant.

Table 9.2: Overview of ANOVA results for the impact of modifying multiple features on performance measures (Section 9.4.4) and ability to replicate target feature values (i.e., error, Section 9.4.4). Columns show effect sizes (ω^2) for *number* of modifications, *session*, and *distributed* multiple feature modifications, plus interactions. See text for details.

For *flight time*, we found errors to be significantly smaller for the *default* time than the *long* one (p<.001, d=-1.488), as well as for observations from the *second* session compared to the *first* (p=.004, d=-.645). The latter matches the observation that people typed slightly faster in the second session.

Regarding *hold time*, post-hoc tests showed errors to be significantly smaller for the *default* time compared to the *long* one (p<.001, d=-1.844). For *touch area*, we found errors to be significantly smaller for the *default* area size compared to the *large* one (p<.001, d=-4.470).

In summary, these results confirm that participants significantly modified their behavior, namely toward the values indicated by our visualization. In addition, people are more accurate in producing the default feature values compared to the more extreme ones, likely because the latter are further away from "natural" typing behavior.

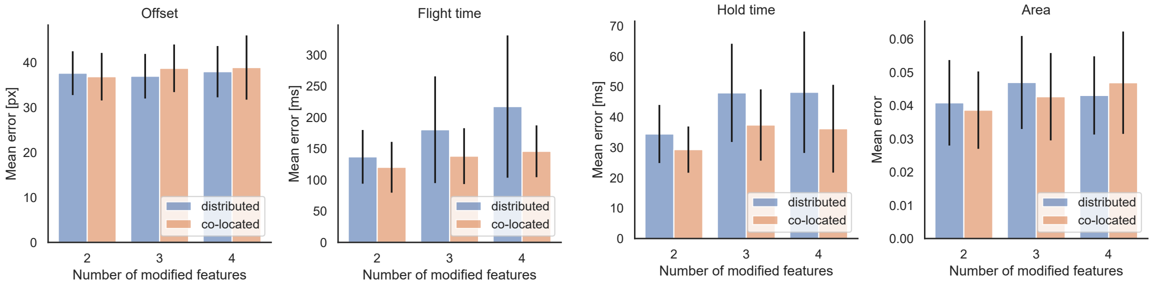


Figure 9.6: Participants' ability to replicate given behavior depending on the number of features that had to be modified in one password and whether those features were co-located on a single key or distributed over the password.

9.4.4 Impact of Modifying Multiple Features

Here we report on users' ability to modify multiple features in one password. Table 9.2 summarises the ANOVA results. Post-hoc tests and further details follow below.

Impact on Time, Speed, and Incorrect Entries

For *task completion time*, post-hoc tests revealed that three modifications resulted in significantly longer times compared to two (mean 40.70 s vs 36.36 s; p<.005, d=0.543); descriptively, this was also true for four modifications compared to two, yet not significantly so (p=.064). Moreover, distributed multiple modifications took significantly longer than co-located ones (mean 42.33 s vs 34.33 s; p<.01, d=1.397). People were also significantly slower in the first session than in the second one (mean 39.76 s vs 36.90 s; p<.05, d=0.444).

For *typing speed*, all pairwise comparisons of the number of modifications were significant (all p<.001), with slower typing for higher numbers (mean 2: 30.18 WPM, 3: 27.33 WPM, 4: 25.15 WPM). Moreover, distributed multiple modifications were typed significantly slower compared to co-located ones (mean 26.91 WPM vs 30.46 WPM; p<.001, d=-2.445).

Finally, significantly more *incorrect password entries* occurred for distributed compared to co-located multiple feature modifications (mean 2.44 vs 1.45; p<.005, d=0.677).

These results show that users take significantly longer to enter passwords as the number of modified features increases, in particular, if the behavior is modified for multiple features across different characters (i.e., *distributed*). In that case, people also produce significantly more incorrect password entries.

Impact on Replicating Target Feature Values

Figure 9.6 shows participants' behavior deviation from the given target behavior (i.e., error), based on the *number* of features that had to be controlled within a single password and whether those features were *co-located* or *distributed*.

For *offset*, we found no significant effects (see stable distribution of errors in Figure 9.6).

For *flight time*, errors were significantly lower for *co-located* modifications compared to *distributed* ones (p<.001, d=-.965), and for the second session compared to the first one (p=0.02, d=-.699). Regarding the number of modified features, we observed significantly lower errors for two compared to three (p<.001, d=-1.149) and four (p<.001, d=-1.522), as well as for three compared to four modifications (p<.001, d=-.867).

Post-hoc tests for *hold time* revealed significantly lower errors for *co-located* features (p<.001, d=-1.004) and for two modified features compared to both three (p<.001, -1.479) and four (p<.001, d=-1.073) modifications.

Finally, for *touch area*, post-hoc tests showed significantly lower errors for two modified features compared to both three (p<.001, -1.565) and four (p<.001, d=-0.868) modifications.

The results are in line with the findings from the previous section. Participants generally performed better when features were *co-located* (i.e., not distributed over the password, Figure 9.6) and performance decreased for increasing *number* of modifications. Offset error was stable regarding all factors.

Impact on Subjective Rating

Participants answered three Likert items after each task: 1) *"I was able to adjust to the specified behavior."*, 2) *"I was successful in completing the task."*, and 3) *"The task was difficult for me."* We compared users' ratings on these questions between tasks with co-located and distributed modifications: Wilcoxon signed-rank tests revealed significant differences for all three questions (Q1: Z=3.828, Q2: Z=4.074, Q3: Z= -3.765, all p<.001). Thus, participants subjectively perceived tasks with multiple feature modifications at the same character as significantly easier (i.e., better able to adjust behavior, higher success, less difficult), compared to tasks with feature modifications distributed over several characters.

9.4.5 Impact of Modifications on Individuality

The previous analyses have shown behavior differences *within users*, caused by modification instructions. Complementary, we now investigate how natural behavior differences *between users* are influenced by modifications. This is interesting, for example, to inform behavioral biometric security layers. We will return to this in our discussion.

We thus compared the individuality (or "biometric value" [41]) of typing behavior between natural and modified behavior. To do so, we employed a user identification model [41, 43]. Note, that we do *not* intend to present this model as a practical biometric identification system. We rather use it as an *analysis tool* to quantify the impact of explicit behavior modifications on individuality. Thus, we are not interested in optimizing identification accuracy, but in measuring the differences obtained between natural and modified behavior.

Evaluation Scheme

We used the established Gaussian model for mobile touch typing, with a Gaussian distribution per feature per key [22, 115, 119, 300]. For touch location, for example, it defines the user's spread of touch points when aiming for that key. Thus, each user u is represented by a set of Gaussians (the model m_u), fitted to the touches from the training set for that user. We used the data from the first session to fit these models.

For each user $u \in U$, we then fed the data from u's second session to this user's model m_u, which yields likelihoods for u (for an ideal model, these should be high). In particular, we computed the joint likelihood for all touches for each task t, that is, the likelihood that u is the one who typed the password in task t. Note that the features are per touch, not per password. Complementary, we fed the data from all other users $v \in U \setminus \{u\}$ to the model m_u as well (for an ideal model, these likelihoods should be lower). We repeated this for all pairs of users $u, v \in U$, such that we obtain 24 (user models) × 24 (user data) likelihoods per task. We repeated the whole analysis twice, once for natural and modified typing data.

On these likelihoods, we computed the standard measures for typing biometrics (e.g., see [41, 262]): receiver-operating-characteric (ROC) curve, area-under-curve (AUC), and equal error rate (EER).

ROC Analysis Results

Figure 9.7 shows ROC, AUC, and EER. Compared to random guessing (dotted line, 0.5 AUC), both natural and modified typing clearly yield biometric information. The values are in line with related work using this model for password typing on smartphones with the right thumb in the lab [42]. The results also show that people retain aspects of their individual behavior when asked to perform the same modifications.

The key observation is the *gap* between the curves in Figure 9.7. It quantifies the loss in individuality: To summarize, when measured using an established typing model, the individuality of participants' typing behavior was *reduced* by intentional behavior modifications such that AUC dropped by .07 (relative -8.9 %) and EER increased by .06 (relative +20.7 %).

9.4.6 Detecting Modifications

Finally, we analyzed how well behavior modifications can be technically detected. This is important, for example, to build an authentication system that allows these modifications to be used as part of a password. For instance, to check a password like "pass[*hold long*]word", the system needs to be able to distinguish between normal and long hold times.

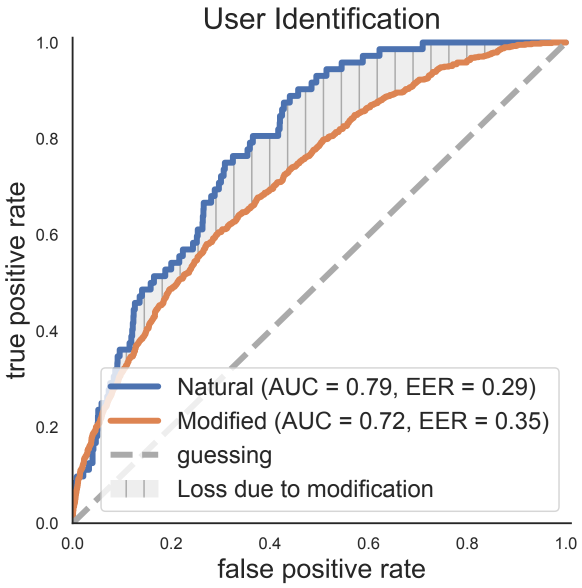


Figure 9.7: Impact of behavior modification on the individuality of typing behavior, quantified by measuring the difference (shaded area) between ROC curves for user identification on natural (blue) vs. modified (orange) typing. Typing behavior becomes less individual through performing modifications. Clear individual characteristics remain, as evident from the modified (orange) line well above chance (dashed line)

.

We employed Random Forest classifiers with 100 trees and default parameters[4]. We used all typing features as input (hold time, flight time, area, offset x, y) and trained one model per modification (e.g., to classify normal vs. long hold times).

We used leave-one-user-out evaluation across sessions: For each user u, we trained the classifier on the first session's data of all users except for u. We tested this model on u's data from session two. Thus, the model could be shipped pre-trained and would not require data collection during enrollment.

We report mean (std) classification accuracy over all users: hold time 97.9 % (1.36 %), flight time 96.14 % (1.84 %), area 94.71 % (1.16 %), and offset 94.29 % (0.96 %). Note, that the remaining error includes user errors (e.g., the user accidentally performed normal instead of long hold time). For these user errors, the model has to give an incorrect classification.

These results demonstrate that modifications can be reliably detected. It is thus technically feasible to implement an authentication system that allows users to use these modifications as part of their password. We provide the model code and trained model[5] to facilitate implementations and further research on such password systems.

[4] Random Forest: https://scikit-learn.org/stable/modules/ensemble.html#forest, last accessed October 16, 2024

[5] Dataset: https://www.unibw.de/usable-security-and-privacy/research/datasets/intentional-behaviour-modifications, last accessed October 16, 2024

9.4.7 User Feedback

After the study we conducted short interviews: Half of the participants (12) stated to be interested in using passwords with behavioral modifications and four were strictly against it. The other eight had concerns (e.g., security, being able to reproduce their behavior under different circumstances, or the technical feasibility of such a system), but stated they would be interested in using a system utilizing intentional modifications if those concerns could be addressed.

Many participants said they struggled with offset modifications as they would often hit the wrong key. Some also had difficulties distinguishing large areas and long hold times.

When creating passwords, users often first observe their natural behavior to then emphasize it. For example, P20 stated: "*When I created the password I first typed it and observed what I automatically did. For example, I typed a 'g' rather to the left, entered a 'b' rather [long]; That's what I adjusted [the password] to.*". Another common strategy was putting modifications at salient positions, such as at the beginning of words or syllables.

9.5 Discussion

9.5.1 Controlling Password Typing Behavior

As a key insight, we revealed that people are able to significantly modify temporal and spatial features of their mobile typing behavior in given directions. It is also possible to train a model that distinguishes between these feature levels (e.g., default vs. long press) with high accuracy (Section 9.4.6).

People were more accurate (i.e., deviated less from target feature values) in reproducing default values rather than extreme ones. We thus conclude that people are better at replicating behavior that is close to their natural behavior.

For flight time, accuracy was higher in the second week. We attribute this to people getting accustomed to our devices, modifications, and tasks, indicating a learning effect.

In some cases, participants performed default behavior when expected to show a modification (see secondary peaks in distributions in Figure 9.4), likely due to the cognitive load of actively controlling their actions, especially when modifying multiple features. Controlling touch area is partly affected by the usage of the right thumb, which naturally leads to larger areas towards the left of the screen, due to stretching.

9.5.2 Modifying Multiple Behavior Features

Overall, modifying an increasing number of behavior features in a password becomes significantly more difficult to control. A possible explanation is the likely higher cognitive demand for intentionally modifying several aspects of typing behavior, as supported by participants' comments and a higher number of incorrect inputs.

Specifically, modifying multiple features at different characters within one password ("distributed modification") is significantly more difficult than modifying multiple features at the same character ("co-located modification"). This conclusion is supported by all quantitative measures (task completion time, typing speed, incorrect entries, error measures), as well as participants' subjective Likert ratings and comments.

Control of temporal features particularly suffers when other modifications are present, likely since focusing on those others distracts users from keeping the timing for the temporal modifications. Controlling spatial features is more robust.

In summary, our findings show that multiple features are harder to control when spread over multiple different characters; in particular, if temporal modifications are involved.

9.5.3 Methodology

We developed a visual text annotation scheme (Figure 9.1) to communicate target behavior modifications. We chose this approach to be able to use text entry research's most common and established transcription task (i.e., enter given text) with our new concept of intentional behavior modifications.

An alternative would have been to visualize desired feature values directly on the keyboard (e.g., show a cross-hair on the key for offset modifications). However, this would have turned the task into a *reaction* exercise (i.e., hitting such cross-hairs), which likely leads to different behavior. This approach also borrows heavily from the technical support work on mimicry attacks. Yet we were interested in users' ability to modify behavior without such scaffolding. With our task, we thus gave clear instructions while participants were left to implement those modifications as they saw fit.

Future work could compare the two approaches. For example, work on systems for mimicry attacks could use our results here as a baseline for unsupported modification ability.

9.5.4 Deployment

As shown in Section 9.4.6, it is possible to reliably detect behavior modifications, which enables building authentication systems that utilize them as part of a password. With backends that store passwords as hashes of strings, this could be easily integrated by inserting a special symbol depending on the preceding character's modification (e.g., "pass$holdlong$word" where $ stands for any character not allowed to be used directly for passwords in the system). Therefore, this technique can potentially be used in any context in which passwords are currently used – given that client software and hardware are capable of detecting modifications. For non-touch keyboards, only temporal features would be available.

Moreover, our visualization (Section 9.2) could give users feedback on their typing, analogous to revealing entered characters in a password field on demand.

Finally, it is not clear how different devices and keyboard layouts influence behavior and control, which could be investigated in future work.

9.5.5 Implications for Usable Passwords with Intentional Behavior Modifications

Intentional behavior modifications increase the space of possible passwords. We focused on the fundamental ability of users to control behavior features. Our results offer plenty of opportunities for future work, e.g., investigating observability and memorability. We summarize practical recommendations for usable passwords with behavior modifications:

Flight time, hold time, and touch-to-key offset present suitable behavior features for intentional modification for password typing on smartphones. Modifications of the touch area for thumb input should be avoided. The area is harder to control since it is partly determined by stretching the thumb.

Flight time and hold time can be controlled on two levels (normal vs. long). Offsets can be controlled on five levels though they were the most difficult modification for participants. We see several options to improve this for future work. This includes tolerance for miss-typing (i.e., accepting input that hits a neighboring key in the direction of the executed modification) and using offset modifications only with larger keys (e.g., on tablets or for PINs). Modifying offsets may also be easier when typing with a different finger which allows for more precision (e.g., index). Modifying behavior for one character in multiple ways should be favored over distributing feature modifications across several characters. Combinations of feature modifications across multiple characters in particular for temporal modifications should be avoided.

Based on user feedback after creating own passwords, a promising creation strategy is to observe one's own natural behavior and add emphasizing modifications.

9.5.6 Implications for Mimicry Attacks

Related work [41, 42] found that spatial features (particularly offsets) have higher biometric value, that is, they lead to more accurate user identification, compared to temporal features. Our results show that it is difficult to intentionally modify multiple temporal features, or temporal features combined with others. In contrast, for modifying offsets, users are not inherently under time pressure when controlling them.

We thus revealed a novel trade-off: Spatial features have higher biometric value than temporal ones in the literature, yet they might be easier for informed attackers to modify. Future work can investigate such mimicry attacks: In particular, our results suggest 1) to compare mimicry attacks on biometric systems that use either spatial or temporal features; and 2) to compare such attacks for "victims" that do or do not intentionally control these features as part of their passwords.

In contrast to most previous work on mimicry attacks, these new study ideas do not focus on technical support for attackers or specific protection methods, but rather on better understanding the fundamental human capabilities for copying and controlling otherwise uncontrolled input behavior details.

9.5.7 Implications for Biometrics Research

We showed for the first time that when multiple people follow the *same* modification instructions, their mobile typing behavior becomes less distinguishable (here relative +20.7 % equal error rate for user identification across sessions).

Earlier work on typing on desktop keyboards [54, 134] and phones with physical key pads [133] discussed "artificial rhythms" (e.g., inserting a pause), which *increased* biometric value, contradicting our results. This difference may be due to typing on touchscreens in our work and the fact that related work studied behavior in one session only, ignoring changes over time. Moreover, users received "open" instructions to modify the rhythm as they liked and thus likely responded in more individual ways [134]. Typing biometrics for desktops can only utilize temporal features. In contrast, mobile touchscreens enable rich spatial features and it can be difficult to coordinate modifications of multiple features in one password entry. This might have caused less consistent behavior across sessions, reducing the accuracy of user identification.

On one hand, this suggests that authentication systems need to be careful with applying *both* behavioral biometrics (e.g., as an extra security layer) and intentional modifications (e.g., for extended password space). On the other hand, suggesting *different* modifications to different users could improve biometric value, as we find users able to follow modifications of the most important features in typing biometrics.

Other work examined related ideas that might be investigated in our context as well: (1) nudging users towards creating more diverse lock patterns via subtle visual cues [280]; and (2) facilitating user exploration of "original" behavior [287].

Our results guide future work on the idea of provoking more diverse behavior: For example, a future study could ask users to set up a password not only with composition instructions (e.g., minimum length) but also suggest (random) behavior modifications for how to enter it. Based on our results, we expect to achieve higher biometric value in this way, compared to 1) suggesting no behavior modifications, or 2) suggesting the same modification to all users.

password length	8	7	6	5
no modifications	**49.36**	43.19	37.02	30.85
1 modification	**55.14**	48.77	42.38	35.94
2 modifications	59.84	**53.27**	46.63	39.90
3 modifications	63.90	57.10	**50.20**	43.16

Table 9.3: Entropy (*bits*) of random passwords with and without (random) modifications on an alphabet of 72 characters (upper and lower case letters, numbers and 10 special characters).

9.5.8 Security Considerations

Using intentional behavior modifications impacts password capture and guessing attacks [31]. *Capture attacks* like smudge attacks[17] may be deflected, as temporal features leave no marks. Video-based attacks like shoulder surfing[234] or thermal attacks [2] may still be possible, though potentially harder, as extracting exact timings may prove difficult and fingers occlude the concrete touch points as long as no feedback is given (compare 9.5.4). Phishing may only be successful if the interface can capture and transmit modifications.

Assuming random passwords and modifications, adding modifications makes both online and offline *guessing attacks* harder (Table 9.3). Including one modification adds up to about 5 bits of entropy (calculations in Appendix B). Thus, modifications may enable shorter passwords with similar entropy. For instance, under the given assumptions, an eight-character password can be reduced to six characters when using exactly 3 modifications. This is promising as passwords on mobile devices tend to be weaker and harder to enter [187].

Notice that these are upper bounds; there may be common patterns of choosing modifications, which reduce theoretical entropy in practice (e.g., participants reported choosing beginnings of words or syllables for modifications, see Section 9.4.7). Moreover, focusing modifications on a single key instead of spreading them out makes guessing easier. However, our calculations assume that the attacker knows the exact number of modifications, thus (slightly) underestimating entropy. While suggesting concrete modifications might solve some of those drawbacks it may introduce usability issues. We suggest practical security as an area for future work.

9.5.9 Limitations

We examined a limited set of typing features with a commonly used keyboard app (modified Google open-source keyboard). We did not measure pressure or shape features from the full capacitive image (see [176]). Nevertheless, we covered the most commonly used temporal and spatial typing biometrics features (see [262, 263]), found to be the most important ones among a larger set for mobile password typing [42].

To avoid an impact of password complexity we chose a limited set of easy passwords for our study. Our findings may not generalize to more complex passwords.

To keep an acceptable study duration, we only observed one-handed use with the right thumb. This is one of the most considered postures in research [28, 111, 112, 300] and one of the most frequently used ones in daily life [41]. All participants were right-handed and used to this posture. Future studies could compare our results to using the index finger.

During the analysis of the results, we noticed that the target behavior in task 34 contained an additional hold time modification instead of the intended flight time modification. Thus the combination of area, hold, and flight time was not tested.

Our sample is biased towards younger people and might not represent the overall population. Finger precision and timing might change with age (see [274]). Future work could compare our results to samples with children and older adults.

9.6 Implcations

Typing behavior can be analyzed to identify users based on features such as typing rhythm [262] and finger placement [42]. So far, research has studied these features as they occur "naturally" as an implicit, uncontrolled part of typing, or in the context of supporting mimicry attacks with technical means.

This chapter addressed the gap in the literature with the first study on users' ability to intentionally modify their behavior when typing passwords on smartphones: We developed a novel visual text annotation in a prestudy (N=114), before using it to study intentional modifications in the lab (N=24). Overall, our results reveal that users can successfully modify the features most commonly used in typing biometrics systems for smartphones. This fundamental insight has several implications for users, threat models, and biometrics research. We conclude by outlining some of them here:

It is worth further investigating the idea of using intentional modifications as a part of passwords. This could extend the password space (e.g., "password" vs "pass[*hold long*]word") and possibly also reduce observability, as attackers would have to guess the modification, not just the entered word.

Our results also motivate novel research directions for touch and typing biometrics systems: These might suffer from "standardizing" typing behavior across users with given modifications, as revealed in our study. However, nudging different users to use different modifications in turn promises to increase user identification accuracy (see [280]). Related, threat models for evaluating such biometric systems need to take into account that some target behaviors are inherently more difficult to attack: In particular, our results strongly motivate comparing attacks that require modifying temporal vs. spatial features to mimic behavior.

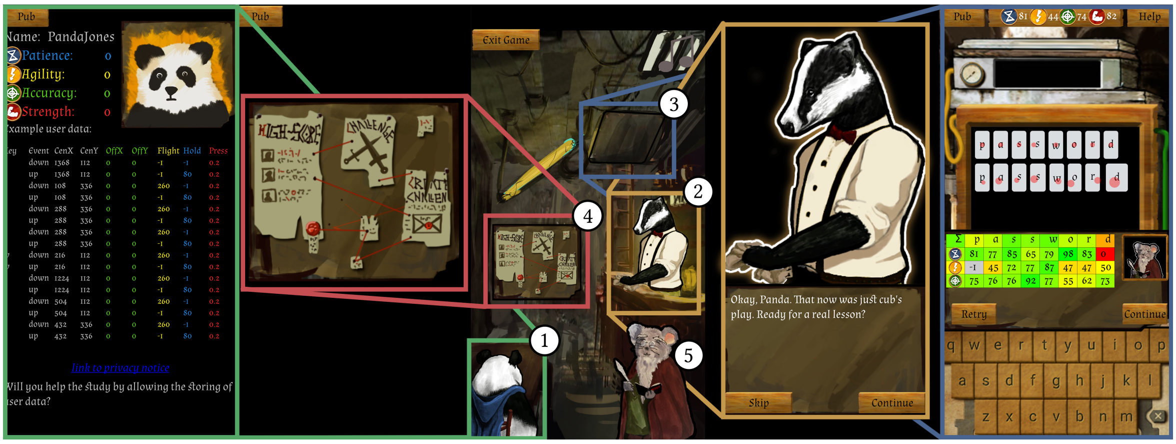
Overall, we show the *rich capabilities* of users to intentionally control typical input behavior features previously considered as an implicit "information byproduct" of interaction. As such, this chapter lays the foundation for our further explorations of ways to support users in modifying their typing to gain agency over their authentication in the remainder of this part.

10
Extending Intentional Keystroke Control to the Wild with *Imitation Game*

This chapter is based on *yet unpublished* work:
Lukas Mecke, Rupert Oxenius, Sarah Delgado Rodriguez, Daniel Busckek, and Florian Alt. 2024. *Imitation Game: A Mobile Game to Support Players in Learning to Control Features of their Typing Behavior*. In preparation for publication.

In the previous chapter, we presented our investigation of users' ability to intentionally modify their typing. We developed a visualization for typing-related features and could show in a lab study, that participants could successfully modify features of their typing towards a given target behavior.

The ability to control typing features can open up several options for new applications. Intentional modifications can be included as additional features of a password and thus either shorten the password while keeping entropy high or increase entropy at no additional length. They can be utilized to make text messages more expressive by applying markup based on typing features [45].



Name: PandaJones
Patience: 0
Agility: 0
Accuracy: 0
Strength: 0
Example user data:

Key	Event	CenX	CenY	OffX	OffY	Flight	Hold	Press
	down	1368	112	0	0	-1	-1	0.2
	up	1368	112	0	0	-1	80	0.2
	down	108	336	0	0	260	-1	0.2
	up	108	336	0	0	-1	80	0.2
	down	288	336	0	0	260	-1	0.2
	up	288	336	0	0	-1	80	0.2
	down	288	336	0	0	260	-1	0.2
	up	288	336	0	0	-1	80	0.2
	down	216	112	0	0	260	-1	0.2
	up	216	112	0	0	-1	80	0.2
	down	1224	112	0	0	260	-1	0.2
	up	1224	112	0	0	-1	80	0.2
	down	504	112	0	0	260	-1	0.2
	up	504	112	0	0	-1	80	0.2
	down	432	336	0	0	260	-1	0.2
	up	432	336	0	0	-1	80	0.2

link to privacy notice

Will you help the study by allowing the storing of user data?

Okay, Panda. That now was just cub's play. Ready for a real lesson?

p a s s w o r d

p a s s w o r d

Σ	p	a	s	s	w	o	r	d
⧗	81	77	85	65	79	98	83	0
⚡	-1	45	72	77	87	47	47	50
⊕	75	76	76	92	77	55	62	73

Figure 10.1: We present *Imitation Game*, a mobile game where players embody a panda (1), taking missions from a badger (2) to break into vaults protected by typing patterns. Players can train for their missions (3), create and take challenges from other players (4), and contribute to our research by filling out questionnaires (5).

Most relevant in the context of this thesis, intentional control can help end-users to (re)gain agency over their authentication: typing can constantly be tracked in the background and there is no clear point for a user to express their intent to be authenticated. Compare this to a user entering a password if, and only if, they want to access an account. Thus, the ability to control typing features could be largely beneficial for end-users to help them keep their identity and other personal aspects private when interacting with non-trusted devices or websites and decide if and when to be recognized.

Those examples show that it can be largely beneficial for users to be able to control their typing behavior. While our previous study (Chapter 9) showed, that this is fundamentally possible, it was limited to a constrained lab setting and did not explore how this skill can be supported in the real world. Related work mainly focused on attack scenarios (i.e. someone trying to mimic a victim's typing patterns to gain access to their data protected by a typing biometric system) [159, 265] which use humans rather as a proxy to enter the text. However, lab studies are not necessarily representative of real-world settings and our goal is to create additional value for the user rather than facilitating attacks. In this chapter, we thus extend our previous work by exploring if those effects can also be achieved outside the lab and under less constrained conditions where they could be of actual value to the users themselves.

To this end, we developed *Imitation Game* as a mobile game built on the typing behavior visualization we proposed in Chapter 9. We used the player types proposed by Bartle [25] as a guideline in designing our game to support players with different preferences in learning how to modify their typing behavior on their own and in a playful manner.

We evaluate players' ability to modify their typing behavior both in a lab study (to validate if the results of our lab study translate to a game approach) as well as in the wild with 24 participants each. We further evaluate the success of our introduced game elements and the effect of player types on their perception. We find that our design was effective in addressing different player types. Participants were able to control the temporal features of their typing but struggled with controlling touch area and horizontal offset. Based on the results, we reflect on our approach of transforming a security lab study into a game to motivate users to learn a difficult skill. We discuss the differences between playing our game in the lab and in the wild and outline future extensions to our work.

> In this chapter, we contribute 1) a mobile game (*Imitation Game*) to support and motivate users in the process of learning to control their typing and 2) two studies (N=24 each) using the game in the lab and in the wild to investigate participants' ability to control their typing.

10.1 Background and Related Work

Here, we provide some background on games and game design with player types in mind as well as previous uses for games in typing research. We omit an introduction to the use of typing behavior and previous work on controlling it in the context of mimicry attacks and instead refer to Chapter 9 for details on those.

10.1.1 Gamification and Serious Games

As described by Deterding et al. "*Gamification* is commonly known as the use of game design elements in non-game contexts" [77]. Other definitions see gamification as "the process of game-thinking and game mechanics to engage users and solve problems" [309] and "A simple concept of making non-gaming systems more engaging through applying gaming principles to them" [32]. In extension, applications are *serious games* if they go beyond the goal of entertainment and offer features with learning and educational purposes [100].

More generally, gamification and serious games can be used to improve activities otherwise unrelated to games like behavior training [77] or learning and acquiring skills [11, 100]. According to Stapleton [255], good gamification offers the same or equal challenges as before, but gives users a different framing and motivation and can thus be more engaging and motivating than standard approaches to training and education. Freitas and Liarokapis therefore see immersion as a central design goal [100]. According to McGonical [94, 179], games offer happiness often not found in reality. Problem recognition and problem-solving are only a few of their various benefits. In a learning environment studies showed more voluntary participation in activities [11].

10.1.2 Game Design and Player Types

Game elements can bring motivational benefits to the gamified application [77, 100, 255]. Basic elements used in the context of gamification are points, badges and levels [167]. Based on Charles et al. [52], the static game design might not account for all players' needs. They proposed an adaptive approach to accommodate individual players' preferences, e.g. by adjusting the difficulty level or feature prominence to a player's style and talent.

Bartle [25] introduced the concept of player types, describing groups of players that are motivated by similar elements of a game. Arkün Kocadere and Çağlar Özhan [11] discuss the need to take those player types into consideration when choosing game elements. They conducted a survey and collected what elements and features different player types enjoyed. Here we give an overview of those player types together with the proposed game elements. The *Killer* is a type of player strongly engaged with other players. As such they seek status and competition which can be achieved through levels, a leader-board, and a competitive point system. *Achievers* are after rewards and achievements and want to beat the mechanics of the game. Potential motivators include achievable elements and progression in the form of points or levels. *Explorers* are mostly about experiencing the game world. They also look for progression but enjoy rewards and a narrative. Elements for them are levels, badges, achievable elements, and a story. Finally, *Socialisers* enjoy interacting with other players. Transaction, cooperation, and narrative are the mechanics addressing them. This can be implemented through a story, badges, and achievable elements. Cooperation can mainly be achieved by a team element. While the player base can be sorted by identifying major character traits, many players usually do not completely fall into one player type [25]. Other approaches (e.g. [95, 102]) exist but we focus on Bartles classification in this work.

10.1.3 Using Games for Typing Research

Previous research has used games both to teach typing skills and assess typing. Costagliola et al. [60] proposed a game to help players learn the "KeyScratch" text entry method. Other games were designed to assess typing errors and utilized the game context to be able to generate a database of expected and actual player spelling [220, 260]. Chen et al. [53] used a typing game to understand user typing timings and Henze et al. [128] observed user typing behavior to implement a touch correction and improve error rates for mobile text entry.

10.2 Research Approach

Our goal for this chapter is to develop a mobile game to support players in learning to control their typing behavior based on our findings from Chapter 9. As such, the game should motivate players to keep playing, be challenging, and last but not least: be successful in supporting players to modify their typing.

From related work we learn, that gamification is an effective tool to increase player motivation and support learning new skills [11, 60, 100]. In Chapter 9 we developed both a visualization of (expected) typing behavior and a set of tasks containing increasing amounts of modifications. However, we used this approach only in a lab setting and with a focus on analysis rather than as a tool to help users actually control their typing. Here we extend this scope by converting our work into a game while still using it as a baseline to compare participants' success in learning to modify their behavior in our game approach.

The work by Bartle [25] and Arkün Kocadere and Çağlar Özhan [11] offered great insights into designing game elements around the types of players. We use this as a guide for our work but mainly focus on elements for Explorers and Achievers, as the other types require a strong social component. However, we design our game with all types in mind and will also evaluate the success of this approach.

Based on those considerations, our research is guided by the following questions:

RQ1 **Game Design**: How to design a game based on a complex study setup to yield results that can be compared to previous work and that is motivating and engaging for players?

RQ2 **Comparison between Settings and Sessions**: How does players' abilities to modify their behavior differ between a (constrained) lab setting and playing in the wild?

RQ3 **Impact of Player Types**: Is our design successful in addressing different player types and does their ability to modify their own typing behavior differ?

10.3 Imitation Game

We conceptualize *Imitation Game* as a story-driven skill game. We describe the constraints arising from our goal to compare results to our previous lab study and the setting chosen for the game. We illustrate the core game loop, how it relates to the replicated study setting, and explain which game elements we chose to account for player types [25]. Figure 10.1 gives an overview of the game and Figure 10.4 an overview of the game elements used in *Imitation Game*.

10.3.1 Adaptations from Our Previous Work

To be able to compare our results to the work introduced in Chapter 9, we had to put some constraints on our game design and study execution. Here we give a short overview of the tasks and changes, for details we refer to Chapter 9.

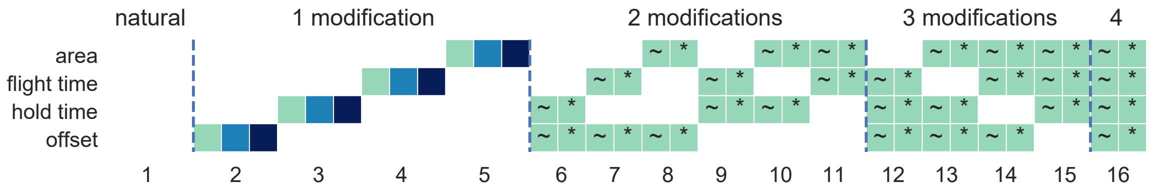


Figure 10.2: Illustration of the feature modifications across tasks used in our previous study (Chapter 9). A darker color indicates more modifications. Features are either distributed (∼) over keys or co-located (*). Below we indicate the mapping of tasks to missions for our game.

Tasks

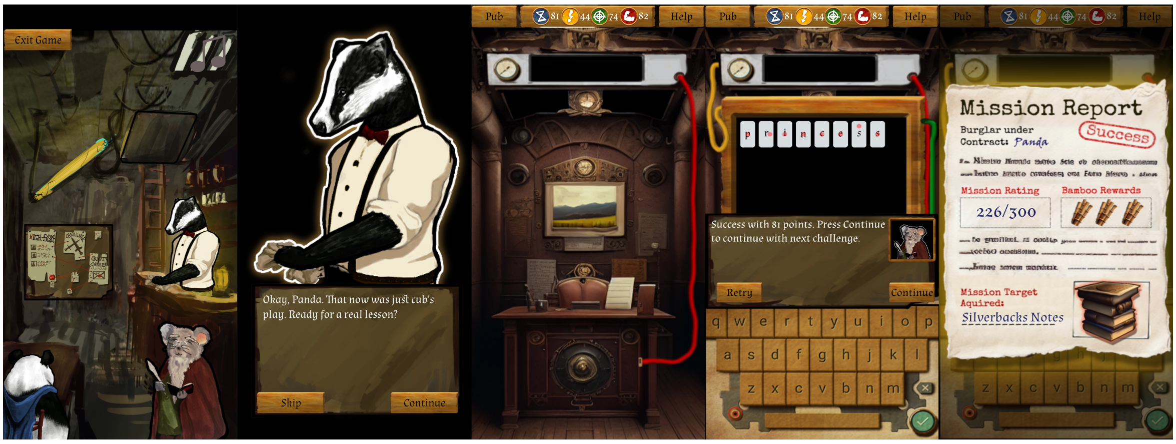
In our previous study, we used 37 tasks. Each task consisted of an eight-character-long password and a pattern of expected typing features. We used the features *flight time* (time between releasing a key and pressing the next), *hold time* (time between pressing a key and releasing it), *touch offset* from the center of the key, and *touch area*. The tasks had an increasing number of features to modify in each password (starting from natural typing in the first tasks and ending in having to modify all four features in a single password in the end). For each combination, we explored different locations within the password to place the modification.

For our game, we preserve both the tasks and their order but group them into 16 missions with each mission having a fixed combination of features to control (e.g. hold time and touch offset). See Figure 10.2 for an overview of the tasks and how they were mapped to missions. We checked for area/pressure[1] sensitivity in the beginning and removed area modifications when they were not supported by the participant's phone. This means, that those participants experienced the same amount of missions but never had to modify the area and did only up to three features.

Keyboard

In our lab study, we used a custom keyboard app to track typing features. This was a viable approach for a lab study but installing a separate keyboard to play a game can not be expected by players in the wild. We thus re-implemented the keyboard as an in-game element and removed all unused keys (e.g., control keys and numbers).

[1] On the device used in our lab study (LG G6) pressure was calculated from area, making the two features interchangeable



Okay, Panda. That now was just cub's play. Ready for a real lesson?

Success with 81 points. Press Continue to continue with next challenge.

Mission Report

Burglar under Contract: Panda

Success

Mission Rating: 226/300

Bamboo Rewards

Mission Target Aquired: Silverbacks Notes

Figure 10.3: Example of how a mission in *Imitation Game* looked like. Players talked to the badger in the pub and received an objective. They would then be taken to a mission location with a vault and had to enter 2 to 3 passwords with behavior modifications. After each attempt, they got a score and the option to make another attempt. After the mission, they got a report including their total score, the bamboo they earned, and the artifact they had to bring back as their mission objective.

10.3.2 Game World and Characters

We chose to situate our game in a steampunk-inspired world populated by talking human-like animals. We made this choice both to create a captivating setting for the story and to have more freedom on how the world works: things that would not be possible in our world do not have to be a problem in this world. The main character embodied by the player is a panda burglar who takes contracts from a badger in a shady pub to break into vaults protected with typing patterns to earn their living. Together with a mouse, the badger serves as the main story driver by sending the panda on missions and explaining the world. The mouse is a researcher and serves as a proxy for our study setting: it explains the game mechanics and typing features and wants to better understand the panda's ability to modify their typing behavior. It thus has some further questions from time to time, which take the players to the three questionnaires that were part of our study. Whenever the mouse had no questions, players could tap it to give feedback about the game. Overall, the story was designed to appeal to Explorers.

10.3.3 Missions and Scores

The main mechanic of the game is to mimic given typing patterns that the panda gets as missions from the badger to access safes and steal artifacts. We avoid the question as to how the badger gets into possession of those typing patterns. Figure 10.3 shows an example of how such a mission would look like. We reused the target typing patterns from our previous study described in Chapter 9 and regrouped them as missions (see Section 10.3.1).

Modifications were applied to three easy passwords (password, football, princess) to not distract from the task. In the game, we explain this by the adoption of typing patterns as a de-facto security mechanism, making the passwords themselves irrelevant.

When on a mission, the player only sees the target typing pattern but gets no feedback on their own input. After each attempt they get a score between 0 and 100 for their success as well as a golden bamboo stick in case they reached more than 50 points (see Figure 10.4e). If they achieved less, we nudged them with a text message to try again to be able to gain the bamboo. We included this element to appeal to Achievers. As an alteration to our lab study, players did not have to enter each password six times. In a game context, this would be too repetitive and we instead utilized the score and bamboo rating to encourage players to attempt to improve their own performance. However, players were free to continue once they had correctly (i.e. without spelling mistakes) entered the password if they so chose.

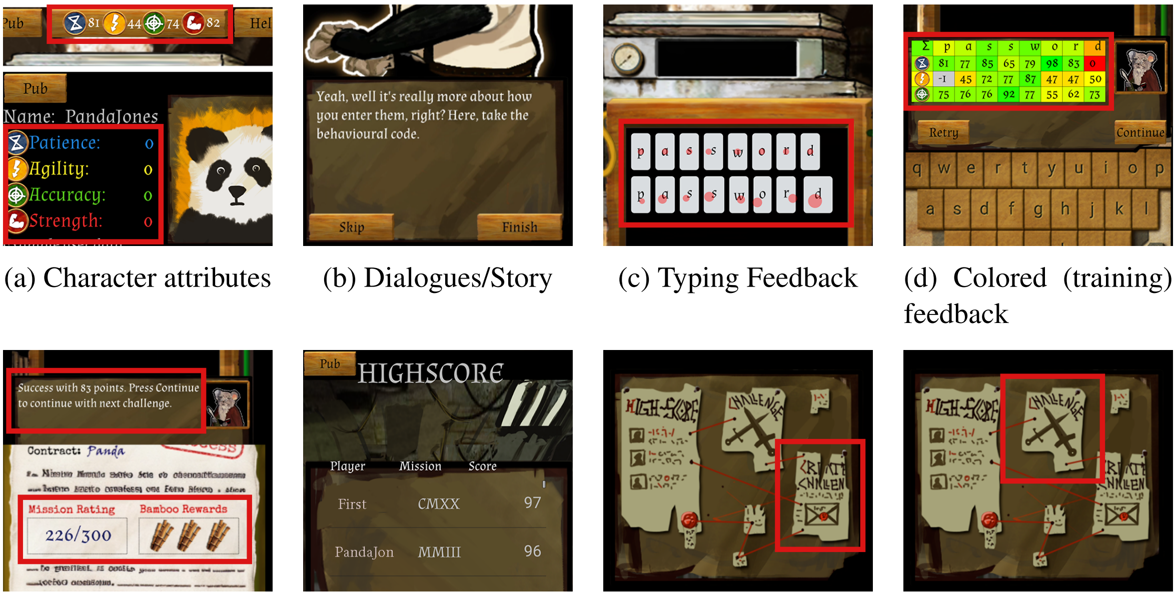
We calculate a score for each feature on each keystroke and then combine them to the score the participants get to see in the game. We implemented a linear mapping of the distance from the target (e.g., the time difference between an expected long press and the player's actual hold time) to the score. We started with no difference mapping to 100 points and a distance of twice the mean error we found in our lab study mapping to 0 points. We refined those values through several rounds of testing. Our aim here was to roughly reflect the participants' performance while allowing them to achieve scores that would be motivating. To give a stronger weight to modified features (i.e. features that did not have their default value), we separated all scores within a password accordingly and calculated the final score as equal weights of the mean of both groups.

10.3.4 Further Activities

We used the steampunk pub as the hub area of the game. This is where players start each mission and where they return once they finished. From here, players also had access to the other features of the game (see Figure 10.1).

Training Room

In our lab study (see Chapter 9) we had participants train their input before each task. We decided to detach training from the missions for our game approach, but in turn, give more feedback when players chose to use it. In addition to seeing their own typing visualized in the same format as the typing pattern they were to emulate (see Figure 10.4c), we also showed them a breakdown of how close they were to the target pattern with each feature of their typing (see Figure 10.4d). We included this element to appeal to Achievers but also to encourage players, in general, to explore further how their behavior would affect the different features. The training room progressed together with the rest of the story so that players could always train for their current mission.



Yeah, well it's really more about how you enter them, right? Here, take the behavioural code.

Σ	p	a	s	s	w	o	r	d
⧖	81	77	85	65	79	98	83	0
ϟ	-1	45	72	77	87	47	47	50
⊕	75	76	76	92	77	55	62	73

Success with 83 points. Press Continue to continue with next challenge.

Player	Mission	Score
First	CMXX	97
PandaJon	MMIII	96

(a) Character attributes (b) Dialogues/Story (c) Typing Feedback (d) Colored (training) feedback

(e) Mission scores (f) Highscore (g) Creating challenges (h) Taking challenges

Figure 10.4: Overview of the game elements we used in *Imitation Game* and that we asked participants about in the questionnaires.

Highscore and Challenges

We included a highscore board (see Figure 10.4f) tracking the best scores achieved together with the player who managed that. In addition, players could also create their own challenges that other players would then be able to compete in (see Figures 10.4g,10.4h). Challenges work similarly to missions, only that the players are free to choose a password and related typing pattern they want to define as a challenge. To make sure players would not just randomly type a sequence nobody else would be able to mimic, we had them first complete their challenge themselves. This is the only social feature in our game and is aimed at appealing to Killers.

Character and Attributes

When tapping on the panda, players could access a character sheet (see Figure 10.1). Here (and in the top bar during other activities), players could see their character's attributes (see Figure 10.4a). They describe how good the players are at modifying certain features and thus are a representation of the player's skill in the game. This is again aimed at Achievers. In addition, the players had access to the data privacy policy and could accept or decline the upload of their typing data from this screen. We showed an excerpt of the data we captured from their last input to make it more graspable what kind of data they would agree to share.

10.3.5 Implementation

We implemented *Imitation Game* as an Android application and created an in-game keyboard so that no further downloads were necessary. One researcher created the story, lore, and dialogue for the game. Based on their descriptions another researcher created all visual elements. Most visual content was hand-drawn, but we used images created with Stable Diffusion [222] as a base for the mission locations and artifact icons. This allowed us to include a greater variety of visual content. We distributed the game through the Play Store to make it easy to access for remote participants.

10.4 Evaluation

We now illustrate our study design to support players in learning to modify their typing behavior in the wild. Each study session encompassed the participants playing through the game once and filling out three questionnaires.

10.4.1 Study Design

Our study design follows a mixed methods approach with both within- and between subject components and two independent variables: We conducted the study both in a *lab* and a *remote* SETTING. In the lab condition, participants were limited to using their right thumb for typing and were provided with a device. Those constraints reflect the lab setup of our previous study (including the exact device used). In the remote setting, participants would play the game on their own devices without constraints. For the lab condition, we also added a second SESSION at least one week after the first to see potential learning effects. We omitted this in the remote condition to more closely mimic a real playing scenario. Replaying a story-driven game just one week later seemed like an unrealistic scenario in this case.

10.4.2 Measurements

Throughout the study, we captured both participants' typing behavior when completing the tasks and their answers to a total of three questionnaires. Here we give an overview of the data collected.

Keystrokes

For each password entered in our study, we capture the *touch offset* from the key center and key *hold time* as well as *flight time* between key presses. If supported by the device we also capture *touch area*. In the lab condition, this was always the case as we provided a device capable of recording touch areas.

Participant Descriptors

We collect participant demographics (age, gender, and country of origin), the hand posture they use for playing the game and their player type [25] based on the questions of an online questionnaire[2].

Interaction with Game Elements

We collected participants' ratings on various game elements (see Figure 10.4). We asked if they liked and were motivated by them with the goal of correlating those ratings with the previously collected player types to asses if our game design (see Section 10.3) was effective in appealing to the respective groups.

Perception of Security and Behavior Modifications

We asked participants how difficult they perceived it to modify the four features. In addition, we presented them with a row of Likert statements assessing their perceived success and improvement in modifying their typing behavior. We asked if they would use our game outside of a study context and how they perceived passwords and behavior modifications for security purposes.

User Experience and general Feedback

We asked participants to fill out the short version of the user experience questionnaire UEQ-S [238] to assess their experience with the game interface. As a last block in the questionnaire, we asked open-ended questions with regard to what participants liked and disliked about the game and gave them space to describe potential problems they had encountered and further remarks they had.

10.4.3 Procedure

We sent remote participants an email with instructions on how to install the game on their devices. From there, their only task was to play through the game once. Lab participants received a prepared device from us and were instructed to use only their right thumb. Otherwise, their task was identical. After one week we invited lab participants to the second session, reset the game, and repeated the process. Consent to the data collection was given inside the game and the mouse prompted participants whenever they were supposed to take a questionnaire.

[2] Bartle Test: `https://matthewbarr.co.uk/bartle/`, last accessed October 16, 2024

10.4.4 Recruitment and Participants

We recruited a total of 48 participants (24 for each setting) through advertisements on university mailing lists and personal contact. Lab participants (12 male, 11 female, 1 preferred not to say) were between 22 and 59 years old (Mn=29.0, SD=7.2) and came from Germany. Remote participants (9 male, 15 female) were between 18 and 55 years old (Mn=27.3, SD=8.0) and came mostly from Germany with one participant from India and one from Austria.

Participants in the lab predominantly followed our constraint of typing with the right thumb (21) but three participants used both thumbs. Remote participants mainly typed with both thumbs (15) followed by the right index finger (5).

Each session took about one hour (depending on whether participants read the whole story, made multiple attempts to increase their score, or took challenges) and participants were compensated with 10 € for each session. The study was approved by our institute's ethics commission under Nr. EK-MIS-2023-202.

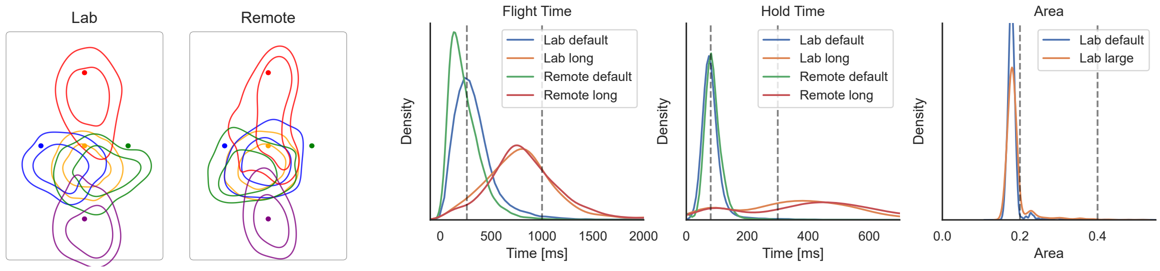
10.4.5 Limitations

Our sample is biased towards younger people from Germany and might not represent the overall population. While the remote setting would have allowed for sampling worldwide, we decided against it to keep the samples for the remote and lab study comparable. Future work could investigate cultural differences.

We used simple passwords in our study, that do not reflect the complexity of real-world typing. However, we allowed players to create their own challenges by choosing both a password and modification, allowing them to extend our concept to more complex tasks.

Participants in our remote condition played the game on their own devices which were mostly not capable of sensing touch area. While this prohibited us from comparing this feature for lab and remote participants, it reflects the actual devices those participants used.

Finally, we designed *Imitation Game* as a game to be played by participants on their own accord. However, in our study participants were externally incentivized to play the game so we have no insights for self-motivated play. We published the game on the Play Store but no other players downloaded it (yet).



(a) Modifications with respect to expected Offset (Colored points: target)

(b) Modifications with respect to expected flight time, hold time, and area. Vertical lines indicate the target values for modifications.

Figure 10.5: Overview of participants' modified typing behavior across settings. Offset (10.5a) is illustrated in relation to a key shape and divided into two plots for better visibility. Participants were overall successful in following the modifications, except for offset to the left and right for remote participants and modifications towards a larger touch area.

10.5 Results

We collected a total of 52,753 keystrokes within missions and 72 answers to each of our questionnaires. Here we report on the results of our study. Our analysis is guided by our research questions (see Section 10.2) and in particular the comparison between settings and the impact of player types. We also compare our results to previous work and explore how participants used and perceived our game.

To account for different screen resolutions of remote participants we corrected absolute offset values to be relative to the key width. Touch area detection was unavailable on almost all devices in the remote setting, so we omitted it in the analysis for remote participants.

To find significant differences, we distinguished between typing data and Likert scales. For Likert scales, we used non-parametric tests (Wilcoxon for paired samples and Mann-Whitney for independent samples). We tested typing data for normality with a Shapiro-Wilk test and used paired/independent t-tests when they were normally distributed and Wilcoxon or Mann-Whitney tests otherwise. Correlations were tested using Pearson correlation. For the sake of brevity, we focus on reporting statistically significant (alpha level p<.05) results. We include further results in Appendix C.

10.5.1 Modifications to Typing Behavior

Here, we explore if participants were successful in modifying their behavior and how their success was impacted by playing the game a second time or in the wild.

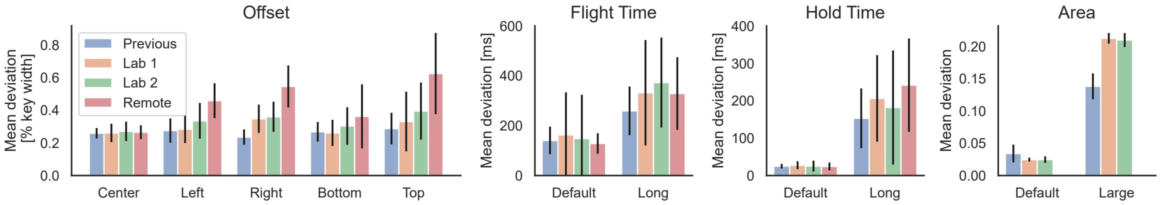


Figure 10.6: Observed deviation of participants' behavior from the target values of the given modifications (i.e. error) for both sessions and settings. Results of our previous study (Chapter 9) are added for reference.

Ability to Modify Behavior

We explored if participants were successful in modifying features of their typing by testing if their typing significantly adjusted when presented with an expected modification. Figure 10.5 gives a visual overview.

We found significant differences in the expected direction (e.g. significantly longer flight time when it was expected) for all feature modifications except for touch area (p=.178) in the lab setting. In the remote setting, participants were able to significantly modify all features in the expected direction except for offsets in the horizontal direction (left: p=.900, right: p=.466). We did not test touch area, as only a few remote participants had this feature.

Impact of Setting and Session

We had participants play the game with fixed hand positions in the lab and without constraints on their own phones remotely. Here, we explore the difference between the two settings. Figure 10.6 illustrates the deviation of participants' behavior from the expected value for all four features and is divided by setting and session.

We found all deviations in offset (except for the center) to be significantly lower in the lab setting (.002<p<.001). Remote participants showed a lower flight time for default modifications (U=117, p<.001) as well as a lower flight time error (U=200, p=.036) for this condition. A visual comparison to the results of our lab study described in Chapter 9 (see Figure 10.6) shows a similar distribution with a tendency towards slightly higher deviations in our study, in particular for the remote condition.

Lab participants were asked to replay the game one week later so that we could see potential learning effects. We found a significant decrease in the flight time deviation (and flight time (Z=37.0, p<.001)) for short key presses (Z=58.0, p=.007) and an increase in the deviation for both top (Z=80.0, p=.007) and bottom (Z=-3.0, p=.045) offset.

10.5.2 Participant Ratings of Difficulty, Success and Future Use

In our questionnaire, we added several Likert statements referring to the perceived difficulty of the different modifications, success in completing the tasks, and attitude towards using the game or authentication involving typing modifications in the future. All statements were rated on a 7-point Likert scale (1=strongly disagree, 7=strongly agree).

Perceived Difficulty of Modifications

We asked participants about their perceived difficulty with different feature modifications. They rated controlling flight time as somewhat easy (Mdn=3) and controlling hold time offset and touch area as somewhat hard (Mdn=5). Participants in the remote setting rated offset to be significantly harder to control (U=172.5, p<.001, Mn_{remote}=5.33, Mn_{lab}=4.21) than participants in the lab. This aligns with our collected keystroke data showing the same effect. Participants in the lab setting rated the touch area significantly harder to control (Z=24.5, p=.035, Mn_{S1}=5.09, Mn_{S2}=4.42) in the second session.

Perceived Success and Future Use of the Game

We asked participants to rate statements with regard to their perceived success in the game. Participants slightly agreed (Mdn=5) to be able – and to have improved their ability – to adjust to the specified behavior as well as be able to influence their typing. They found the tasks slightly difficult (Mdn=5) but felt they were successful in completing them (Mdn=6). Participants slightly disagreed with wanting to play the game outside a study context (Mdn=3). Our tests showed, that participants rated their success in completing the tasks (Z=7, p=.018, Mn_{S1}=5.17, Mn_{S2}=5.80) and their ability to adjust their behavior (Z=16.5, p=.033, Mn_{S1}=4.71, Mn_{S2}=5.25) significantly higher in the second session. We did not find a difference in perception between settings.

Attitude towards Using Typing Behavior Modifications for Authentication

We asked participants to rate statements with regard to their attitude towards using typing behavior for authentication. Participants found both using passwords and using passwords with behavior modifications to be secure (Mdn=6). They were neutral towards using (only) behavior modifications for authentication being secure and wanting to use such a system (Mdn=4). We observed an increase in participants' perceptions of passwords being secure in the second session (Z=28, p=.049, Mn_{S1}=5.21, Mn_{S2}=5.67).

10.5.3 Game Design

We asked participants to rate their agreement on how much they liked and were motivated by a range of game elements on a 7-point Likert scale (see Figure 10.4). As some participants may not have experienced all game elements (e.g. taking challenges) answering these ratings was optional. Results are shown in Figure 10.7. Participants slightly liked (Mdn=5) the character attributes, story, typing visualization, training feedback, and mission scores. They were between slightly agreeing and neutral (Mdn=4.5) for creating challenges and the dialogues and rated the highscore and taking challenges neutral (Mdn=4). With regards to

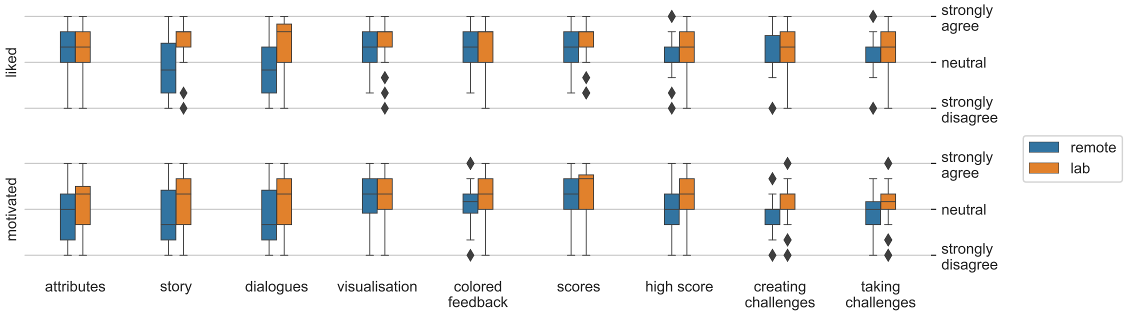


Figure 10.7: A boxplot of participants' responses to whether they *liked* or felt *motivated* by different game elements.

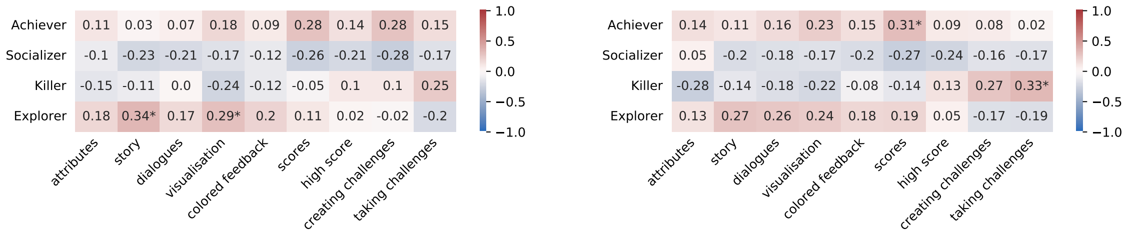
motivation, participants were neutral (Mdn=4) about almost all game elements. They were slightly motivated by the typing visualization and the training feedback (Mdn=5) and rated the mission scores as still slightly more motivating (Mdn=5.5). Participants indicated that they skipped game dialogues (Mdn=4.5). We used the UEQ-S scale to investigate participants' user experience. Overall, participants rated the user experience of *Imitation Game below average* [129] (hedonic quality=0.60, pragmatic quality=1.04, overall score=0.82).

Impact of Setting and Session

Lab participants indicated liking the story (U=146.5, p=.002, Mn_{lab}=5.33, Mn_{remote}=3.71) and dialogues (U=160.5, p=.004, Mn_{lab}=5.08, Mn_{remote}=3.71) significantly more than remote participants. We observed the same effect on their motivation. Lab participants indicated feeling significantly more motivated by the story (U=196.5, p=.029, Mn_{lab}=4.58, Mn_{remote}=3.50) as well as by creating (U=166, p=.034, Mn_{lab}=4.50, Mn_{remote}=3.82) and taking (U=174.5, p=.035, Mn_{lab}=4.54, Mn_{remote}=3.70) challenges. Similarly, lab participants also rated the hedonic and overall user experience of *Imitation Game* significantly higher compared to remote participants (hedonic quality: U=148, p=.002, Mn_{lab}=1.13, Mn_{remote}=0.07; overall score: U=149.2, p=.002, Mn_{lab}=1.177, Mn_{remote}=0.458). Participants liked the typing visualization slightly more (Z=30, p=.044, Mn_{S1}=5.00, Mn_{S2}=5.63) and skipped dialogues more frequently in the second session (Z=11.5, p=.002, Mn_{S1}=3.83, Mn_{S2}=5.75).

Impact of Player Types

Participants mostly aligned with the Explorer player type (lab: 0.69, remote: 0.70) followed by Achiever (lab: 0.50, remote: 0.56), Socializer (lab: 0.48, remote: 0.47) and Killer (lab: 0.34, remote: 0.30). Here, we correlate player types with the game elements they liked and were motivated by. Results are shown in Figure 10.8. We found a significant positive correlation between Explorer types and liking the story (r=0.34, p=.017) and the typing visualization (r=0.29, p=.046). Affiliation with the Achiever type correlated positively with being motivated by the mission scores (r=0.31, p=.032) and Killers were more motivated



(a) Player type vs. Liked Game Elements.

(b) Player type vs. Motivation through Game Elements.

Figure 10.8: Heatmaps illustrating results of the correlational analysis (Pearson's r). (*) denotes significant correlations ($p < 0.05$). We used only data gathered during the first session for the correlational analysis.

by taking challenges (r=0.33, p=.027). Even though no other correlations were significant we observed a tendency of positive correlations with our game elements for Achievers and Explorers and negative correlations for Socializers or Killers.

To see if the player type impacted the actual performance in the game we also conducted a correlation of different player types with participants' achieved score (i.e. the measure combining their deviations from expected behavior that was shown in the game), but we found no significant correlations.

10.5.4 Open Feedback

At the end of the study, we asked participants what aspects they liked and disliked about *Imitation Game*. Participants mostly mentioned the story and dialogues, the characters, and the visual style. Participants liked "*[t]he funny dialog[ue]s [and] emotional investment in the story*" as well as "*the inside jokes about pop culture, literature [and] security*". They appreciated the "*wacky characters*" and found the panda to be cute. Finally, several participants mentioned to like the idea of a game about showing and modifying typing features: *"I like how it can actually sense and track my speed and other attributes of typing. It's actually something unique ".*

At the same time, the dialogues were also the most disliked feature followed by the tasks being repetitive and limited to only a few passwords. One participant summarized it like this: "*The dialogues were sometimes too long and it got a bit boring to do the same kind of task over and over*". Some participants wished for voice acting in the game and more guidance (e.g. a metronome) for modifying their typing. They mentioned the score calculation to be unclear or good scores too hard to achieve ("*Pressure is not recognized as much as I want*"). Some participants encountered crashes with our app and found the navigation unclear: "*I reali[z]ed too late that I had an option for training, it was not clear for me in the pub scenario were to click for tha[t] purpose*".

10.6 Discussion

We designed *Imitation Game* to explore how approaches from gamification research can be used to transform a complex security lab study into a game. We had the goal of supporting players in learning to modify their typing behavior on their own and under less constrained conditions. Here, we reflect on the success of this approach, the limitations of our work, and lessons learned, both for future work and for other researchers attempting such an approach.

10.6.1 Learning to Control Typing Behavior with *Imitation Game*

Both participants in the lab and the remote part of our study were able to significantly control the *temporal* features of their typing. This in itself is already useful, as those features are the most widely used and available on all devices. However, we did not see significant adaptations of touch area in the lab (the feature was largely unavailable on participants' phones in the remote condition) and we found no evidence of remote participants being able to control their horizontal offset. This is surprising to us, given that we used the same tasks, visualization, and – at least in the lab setting – device and hand posture as in our previous lab study (Chapter 9) where we found participants able to modify those features. There are some small differences, that could have had an impact. We called touch area pressure in the game to give participants a clearer impression of how to modify it (pressure is calculated from the touch area on our test device, so both terms are technically interchangeable). This may have led participants to a wrong mental model on how to modify this feature. We decided against having participants train before each input but made this an optional game element. As a consequence, participants did not get direct visual feedback before each task and may not have noticed if they did not correctly modify a feature as long as the other features compensated for the loss in score. Finally, the scoring function could have been too generous, allowing for high scores with little adaptation of behavior. Factors that changed for the remote setting and could have influenced participants' ability to modify horizontal offset include the use of different hand positions, the uncontrolled in-the-wild setting, or an overall smaller screen resolution on participants' own devices. We conclude, that rather small changes (like wording, adding a score, or smaller screen resolution) may have a large impact on participants' success in modifying typing behavior. Future work will be needed to better understand such effects.

10.6.2 Designing for Player Types

We designed *Imitation Game* as a story- and skill-driven game with Achievers and Explorers [25] as potential players in mind. We observed a (non-significant) tendency in the ratings for our game elements to be positively correlated with affiliation to those player types. In addition, we also found significant effects that supported our design decisions: Alignment with the Explorer type correlated with enjoying the story, Killers were motivated by taking challenges, and affiliation with the Achiever type correlated with being motivated by the mission scores. This shows that designing game elements targeting specific player types is indeed effective in appealing to them.

We also observed that most of our participants aligned with the Explorer- and Achiever types. This raises the question if our game was well received because participants naturally aligned with the types we designed for or if our study explicitly attracted such participants. While we did not find differences in the performance of different player types, this can also be relevant for other types of studies that focus on aspects that are valued by certain player types (e.g. a task involving competition might attract more Killers and Achievers).

10.6.3 Reflections on building a Game as a Research Tool

Using a game as a research tool comes with both advantages and drawbacks. First and foremost, development takes a lot of time. A game needs a good game mechanic, should have aesthetic visuals and (if applicable) a story or other motivating elements. The story in our game was both one of the most liked and disliked game elements, showing that it is difficult to cater a game to a general audience and that further refinement would be needed. All this takes time and effort, even though advances in generative models may make this process more approachable (e.g., by generating visuals). In addition, the implementation needs to account for more variables and gets more complex and error-prone: when deploying in the wild, a lot of unforeseen things can go wrong.

In turn, participants may be more motivated to be part of a study including games and put more effort into completing them. Games can also allow for a more realistic view of the actual usage of a system. In our study, we asked participants to use our game both in the lab and remotely on their own devices. We observed that deviations from the expected behavior were larger in the remote condition and lab participants reported to be significantly more motivated by our game elements. This may be an effect of social desirability bias or them putting in just more time to explore the game in the lab, but it may also well be that our lab settings drew an overly optimistic picture. We believe, that lab studies have their place in fundamentally showing that effects (can) exist, but taking the step into the wild, e.g. through implementing a game, results in a more realistic view of actual use. In this study, we found designing a game guided by player types to be an effective way of taking this step.

10.6.4 Extending Imitation Game

We implemented *Imitation Game* as a research tool for our study but also as an extendable game that is – in the long run – intended to be actually played in the wild. From participants' feedback, we gathered some directions for improving the game, including technical improvements, voice acting for the dialogues, or improvements to explanations and navigation. We observed participants typing faster, deviating stronger from the expected offset targets, and skipping more dialogues in the second session. This hints at them putting in less effort, most likely as a consequence of the game being repetitive. A future iteration could improve on this and design for more long-term playability or incentivize replay through e.g. a diverging storyline. Beyond improvements to the game, we also see the potential to adapt *Imitation Game* to explore different aspects of typing. Potential extensions could be the inclusion of more complex passwords or writing full text. The scoring system could be adapted to different functions to understand its effect on player performance. Some players suggested the inclusion of a metronome or other aids to support them in modifying their typing. While we avoided this to both remain consistent with our previous lab study and assess actual learning instead of reacting to presented indicators, this could be a worthwhile comparison for future work to make.

10.7 Implications

In this chapter, we introduced *Imitation Game*, a mobile game to support players in learning to control features of their typing behavior. We designed our game as an extension of previous work on controlling typing behavior and guided by Bartle's player-type model [25].

Our results show that our design was effective in addressing different player types and participants were able to control temporal features of their typing. However, they struggled with controlling touch area and horizontal offset.

This chapter highlights, that using a *game can be an effective approach* for taking usable security research to the wild. We showed, that modifications to typing behavior can be learned and applied by users on their own using our game, opening up the space for future applications using this ability. At the same time, our work showed, that the lab results we found in Chapter 9 may have been in parts overly optimistic and do not fully translate to a more realistic learning scenario in the wild. This highlights the *need and value of investigating security mechanisms in the lab and in the wild* to gain insights into both their potential and their actual use under more realistic conditions.

11
Supporting Key Targeting using Electromagnets

This chapter is based on the following publication:
Lukas Mecke, Ismael Prieto Romero, Sarah Delgado Rodriguez, and Florian Alt. 2023. *Exploring the Use of Electromagnets to Influence Key Targeting on Physical Keyboards*. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA ’23) [185]

In the previous chapters of this part, we showed, that users are able to learn to intentionally change their typing behavior and can also do so in the wild, enabling new security applications and user agency over their authentication. However, they found this task challenging. In this chapter, we explore an option, for actively supporting users in *executing* such modifications and alleviating the effort required for active typing behavior control.

Previous work already investigated, how interaction with keyboards can be extended and enhanced. Examples include modifications to the resistance [26, 131, 233] and sensation

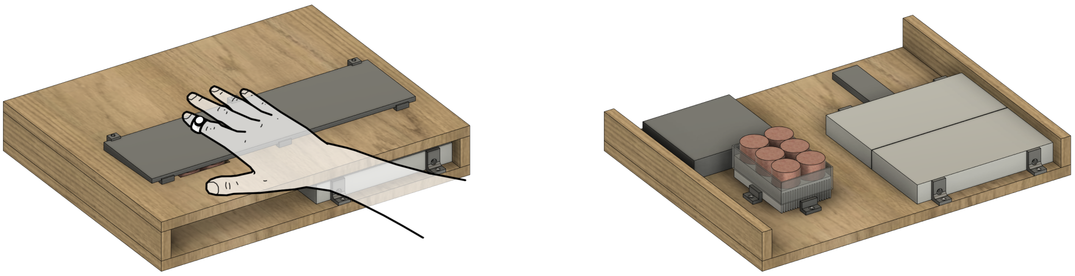


Figure 11.1: We present a prototype to explore if and how users' key targeting on keyboards can be influenced. This is achieved using a magnetic strip on the user's finger (left) that is actuated with electromagnets below the keyboard (right).

when touching a key [51, 191, 194], or lights and vibration to provide feedback [70]. With our work, we aim to extend interaction by not only augmenting touch or providing passive feedback but actively exerting force before, while, and after a key is touched. To the best of our knowledge, there are no other systems in the related literature that are capable of doing this. Our system could be used to provide feedback, feed-forward (e.g., warnings), or subtle guidance during keyboard interactions. Note, how this approach keeps agency in the users hands, as force is applied but it is up to the user to follow it. Compare this to approaches like programmatic changing key timings that are both transparent (and their function thus hard to understand and verify for a user) and take away user agency.

To achieve this, we propose an array of electromagnets below a keyboard to exert forces on a permanent magnet placed on the user's finger and consequently on the finger itself. In this chapter, we prototypically implement this approach and provide an initial technical evaluation and preliminary study with 4 users to understand how the exerted force is perceived and if it can be used to modify key targeting.

We show that we can exert noticeable forces of 3.56 N at a distance of 10 mm. We observe an impact on key press times and errors made as well as a trade-off with the pinkie being easiest to actuate but also liked least. Actuating the index finger allowed for modifying key press times while also being perceived as comfortable. Our work is complemented by a discussion of application opportunities and implications of the introduced approach.

> In this chapter we contribute 1) a prototypical implementation of a keyboard to influence key targeting using electromagnets and 2) an initial technical evaluation and preliminary study to understand the potential and user perception of our approach.

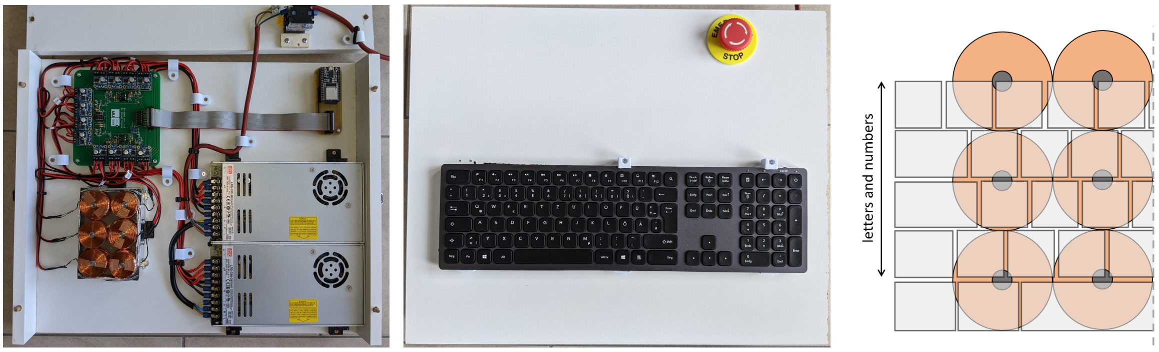
11.1 Related Work

There are different approaches to augment or influence typing. One option is the use of visual and auditory cues [70] or haptic feedback like vibration [170]. Tactile cues [117, 124, 194] can be used to alter touch sensation, e.g., through ultrasonic waves. Other approaches had great success by changing the structure of a touched surface, e.g., through modifying the stiffness of a hydrogel [191] or using stretched fabric [124]. Another approach is to change the resistance when pressing a key through servos [26] or solenoids [131]. Savioz et al. [232, 233] used electromagnets and permanent magnets under the keys (instead of being attached to the users' finger as in our approach) to control key press resistance. While influencing users was not always the goal in the named approaches some demonstrated such abilities. Hoffmann et al. [131] could reduce typing errors with their approach and participants in the experiment by Bell et al. [26] took more breaks. That said, most approaches are limited in that they require touching a surface (e.g., to feel the resistance or vibration) or are passive (e.g., lights [70]). Our aim is to also be able to actively influence users' movements before and after touch. The best option we see for this are magnetic fields which have also been shown effective in the context of physical keyboards [131, 232, 233].

The use of magnetism to influence users has been researched in the past with both electromagnets (EMs) and permanent magnets. Yamaoka and Kakehi [295] moved a permanent magnet under a table through motorized actuators to control the motion path of a pen and guide users (e.g., to replicate or scale drawings). Zarate et al. [304] developed a sphere with three orthogonally oriented EMs, which exerted forces on a ring-shaped neodymium magnet attached to a pen. Mignonneau and Sommerer [189] created an artifact that simulates atomic forces. It contains arrays of large electromagnets that actuated permanent magnets attached to the user's hand at distances of up to 15cm. Similarly, Weiss et al. [286] used an array of EMs and a permanent magnet attached to a finger to guide users' fingers on tabletops. They created an attraction force right below two touch buttons visualized on the screen and repulsing forces around them. This resulted in reduced cumulative drifting in comparison with their baseline without a force field. In our work, we follow a similar technical approach in a different context (keyboards) and with a stronger focus on the impact of design choices (e.g., the placement of the magnet).

11.2 Prototype to Influence Key Targeting

Related work has shown that electromagnets can be used to exert noticeable forces and induce changes in user behavior. With our work, we extend this research to the context of keyboards. The particular challenge is to exert *sufficiently strong* forces also in *mid-air* while using a *minimally invasive* setup (magnetic strip). At the same time, the electromagnets (EMs) need to be as small as possible to fit below the keyboard and have to be placed densely to allow for precise exertion of force. In this section, we describe the design and implementation of a first prototype for achieving those goals.



(a) The matrix of electromagnets is on the left and the power supply units are on the right side. The microcontroller is situated behind the power supply units.

(b) Top view of our finished prototype. The keyboard is fixed with 3D-printed clamps and an emergency button is placed in the right corner.

(c) Dimensions and placement of the electromagnets under the keyboard.

Figure 11.2: Overview of our final prototype consisting of a wooden box a) housing the EMs and the electronics, b) the keyboard mounted on top. Figure c) shows how the magnets are placed under the keyboard

11.2.1 Electromagnets

While smaller EMs provide a higher density of points that can create attraction or repulsion they also produce a weaker magnetic field.

As a trade-off, we chose a diameter of 40 mm and a height of 25 mm. For our prototype, we created a matrix of six such EMs to cover the left half (as we only actuate one hand) of the keyboard (see Figure 11.2c). All EMs were built with a self-made winding machine[1]. We used self-bonding magnet wire (diameter: 0.58mm, resistance: $0.0871\,\Omega \cdot m^{-1}$) to create stable coils. Such coils also have better cooling capabilities as no additional casing is needed to prevent unwinding. We inserted an iron core into the coils to finish the EMs.

Due to minor imperfections in the process, diameters ranged from 35.8 mm to 40 mm and resistance from 7.3Ω to 7.7Ω (measured at 23.7°C). When applying 40 V, we measured currents between 4.2 A and 5.3 A. To achieve a similar force for all EMs we set the maximum current to 4.2 A. We used pulse width modulation (PWM) to dynamically control the force created by the EMs. We used a frequency of 17.5 kHz for the PWM as it is slightly below the maximal audible frequency (20 kHz) and thus hardly noticeable.

[1] Coil Winder: `https://github.com/bonafid3/CoilWinder`, last accessed October 16, 2024

11.2.2 Electronics

Each EM consumes a maximum of 168 W ($4.2A \cdot 40V$). We design for a maximum of three simultaneously powered EMs (504 W) and thus use two 360 W *power supply units (PSU)* that we adjusted to provide 40 V each (9 A); leaving a 216 W margin for current peaks. Each EM has one *current sensor* with a measurement resolution of $0.2V/A$. This allowed us to monitor the consumed current, which is crucial as it is directly related to the created force and can be affected by variations in the coil resistance due to temperature changes. We further included an *envelope detector* to smooth the current sensor's output signal and filter possible peaks. Next, we sample the data using an analog to digital converter (ADC)[2]. We specifically chose an ADC with a high sampling rate to measure the current signal, since it is influenced by the 17.5 kHz PWM signals. To control our setup we used an ESP32 *micro controller (μC)*, which generates 12 independent PWM signals (i.e., two 17.5 kHz signals per driver with a 10-bit resolution). We isolated the μC to protect it from high currents and to allow for a modular circuit design. We used two *drivers*[3] suited for a current of 3.6 A in parallel for each EM. Both drivers in parallel can handle a current of a maximum 7.2 A – enough for the required 4.2 A and potential current peaks. Please refer to Appendix D for detailed schematics of the circuit.

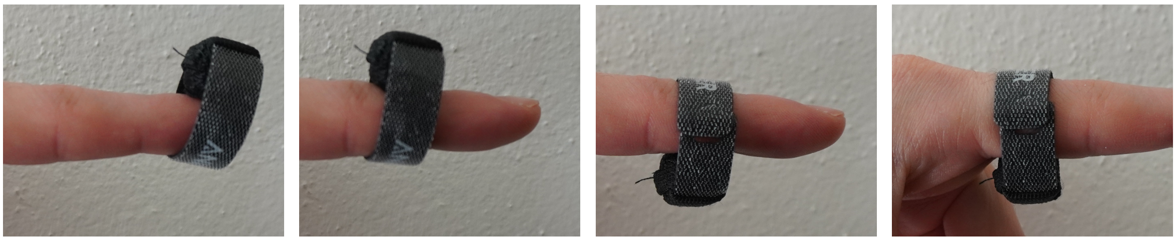
11.2.3 Assembly

We built a portable wood box that contained the EMs and attached the keyboard on top (see Figures 11.2a and 11.2b). A cut-out on the top panel of the box exposes the EMs. We used a generic wireless keyboard[4] with a QWERTZ layout. The keyboard contained a steel sheet, that we cut to accommodate the EM matrix, remove all magnetic elements between keys and EMs, as well as to minimize the distance to the top to 5.84 mm, leading to more exerted force. We used an aluminum cooling block and thermal pads to cool the EMs. To exert forces on the users' fingers we built a magnetic strip that is attached to a finger and actuated by the magnetic field. We sewed a cylindrical N52 neodymium magnet (10×8.5 mm) to a velcro strap (see Figure 11.3). It is reusable, adjustable to different finger sizes, and sufficiently rigid to avoid the rotation of the magnet.

[2] MCP3008, ADC with 8-Channel, SPI capable with a max. sample rate of 200 kilo samples per second and 10-bit resolution

[3] DRV8871 motor driver breakout boards: `https://www.adafruit.com/product/3190`, last accessed October 16, 2024

[4] Keyboard: `https://www.amazon.de/gp/product/B089FF153B/`, last accessed October 16, 2024



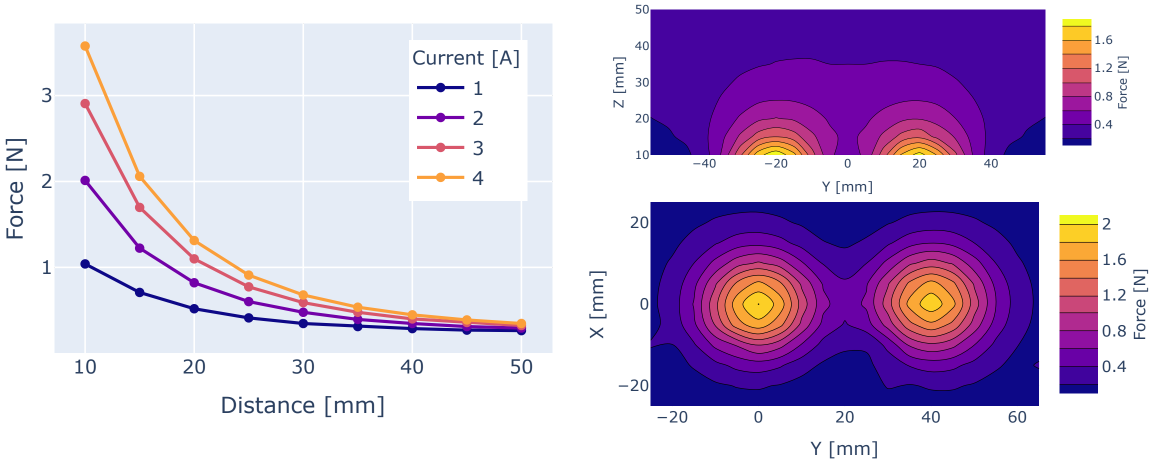
(a) Position 1: distal (top)

(b) Position 2: intermediate (top)

(c) Position 3: intermediate (below)

(d) Position 4: proximal (below)

Figure 11.3: Positions of the magnetic strip in our study.



(a) Force in relation to distance (5mm steps) and currents (1 – 4A).

(b) Side and top view of the force field of two EMs at a current of 2A.

Figure 11.4: Results of force measurements in a regular 5mm grid: a) We measure a maximum force of 3.56N for 4A and a distance of 10 mm that decreases exponentially with distance. When combining two EMs their magnetic fields merge (b, c).

11.3 Evaluation

In this section, we evaluate if our prototype can generate sufficient forces to influence a user's finger movement and could thus in the next step be used to influence typing. To this end, we 1) measure the forces exerted on the magnetic strip and 2) conduct a preliminary user test to determine the best electromagnet configurations (strength and direction of the force) and positioning of the magnetic strip to induce noticeable changes.

11.3.1 Force Measurements

We measured the force exerted to a cylindrical N52 10×8.5 mm neodymium magnet (as the one we used in the magnetic strip) with a force gauge (Sauter FK10) while applying constant currents to the electromagnets.

Maximum Force

To determine the maximum force we measured repulsion exerted by the EM to a permanent magnet centered above the core while changing currents and distances. Figure 11.4a illustrates the exponential decrease of force with respect to distance (e.g., 0.90 N for 4 A at 25 mm vs 3.56 N at 10mm). We measured a maximum force of 3.56 N for a distance of 10 mm at 4A. For 3 A, 2 A and 1 A we measured 2.91 N, 2.01 N and 1.04 N respectively.

Force Distribution

To understand the interaction between EMs we also performed measurements at 2 A in 5 mm steps in the orthogonal and parallel planes with two EMs side by side. Figure 11.4b shows, that the measured force in points between the two EMs is greater than the force in points that are situated on the outer sides. For example, we measured a force of 0.48 N (between) in comparison to 0.3 N (outside). This implies that we can exert more consistent forces within the EM matrix, while the field rapidly decays when reaching the outer border.

11.3.2 User Study

To find the best configuration for influencing finger movement we conducted a pilot study, exploring both the choice of finger and positioning of the strip thereon under varying forces.

Conditions & Measurements

We used a within-subject design with 3 independent variables: We vary magnetic CONFIGURATIONS between *50%* and *75%* of the maximum current (2.1 A and 3.15 A respectively) for both, *attraction* and *repulsion* as well as an additional *off*-condition. We decided against stronger currents to avoid potential overheating. We explored placing the magnetic strip on all FINGERS but the thumb as it is not commonly used for typing letters. We further placed the strip at the top and bottom POSITION of each phalanx of the fingers[5] (see Figure 11.3). We excluded the bottom of the distal phalanx because of the resulting inability to press a key and the top of the proximal phalanx due to the large distance to the EM. To assess the best position we captured typing-related measures like key press duration, flight time, and error rate as well as subjective feedback on the comfort and noticeability of the force.

[5] *Phalanges* are the bones forming the fingers. From fingertip to palm they are called *distal phalanx*, *intermediate phalanx*, and *proximal phalanx*.

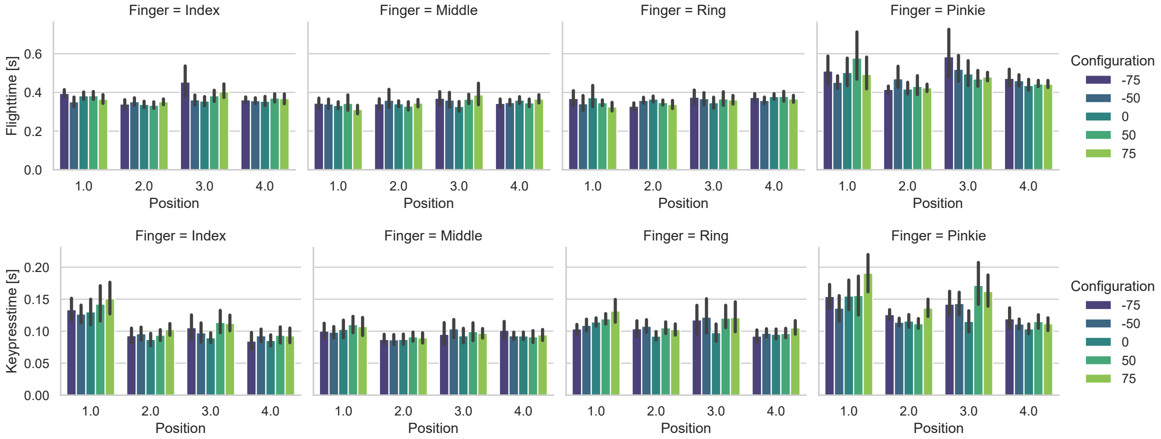


Figure 11.5: Flight time and key press time depending on *finger* used, magnetic *configuration* applied (negative values denote repulsion) and the *position* of the magnetic strip. Errors mainly occurred for the pinkie in the third position under attraction.

Procedure

Participants answered a demographic questionnaire before the magnetic strip was placed in the first position. Subsequently, they entered a specific two-key-sequence five times. Each repetition included pressing Ctrl and then one of the keys *y*, *s*, or *w*. The target key varied depending on the *position* of the magnetic strip and was chosen so that the strip was always situated over the bottom left EM (see Figure 11.2c) when typing the key. Note, that for this preliminary test, only this EM was active and we did not evaluate the interaction effects of multiple EMs. All participants first repeated the task for all EM configurations at the first position (top of distal phalanx) of the first finger (index finger). This procedure was repeated for all positions on the first finger before changing to the next finger. The force configurations followed the order: 1) both attraction configurations, 2) off, and 3) both repulsion configurations. After each configuration participants filled in a questionnaire. Placing the off-level between the attraction and repulsion allowed the EMs to cool down before being turned on again. The order of force magnitudes (50% and 75%) was balanced between tasks. At the end of the study, participants filled out a final questionnaire with questions on the study experience. Overall participants completed 80 tasks (5 configurations, 4 fingers, 4 positions) and filled out 50 questionnaires (48 task questionnaires, demographics, and final questionnaire). The study took about 105 minutes per participant.

Participants

We recruited 4 participants (ages 26–64, two male, two female). They all type for more than two hours per day. *Note*: this evaluation is intended as an initial test of possible prototype configurations so we chose a small sample. We intend to investigate actual typing in a larger study with more participants in the future.

	Finger				Position				Configuration		
Statement	Index	Middle	Ring	Pinkie	1	2	3	4	Rep.	Off	Att.
Force was noticeable	2	2	2	2	2	2	2	2	2	1	3
Typing was comfortable	4	4	3	2	3.5	3.5	3	3	3	3	3
Typing was influenced	1	2	2	3	2.5	2	2.5	2	2	1.5	2.5

Table 11.1: Median participant response to the Likert items presented after the tasks (1: totally disagree, 5: totally agree, Rep.: Repulsion, Att.: Attraction).

11.3.3 Influence on Key Targeting

To account for learning effects we discarded the first repetition of each task. Given the small sample size, we do not conduct statistical tests but report general tendencies. We found that the overall mean key press duration was 0.111s ($\sigma = 0.037s$) and the mean flight time was 0.398s ($\sigma = 0.107s$). Participants made a total of 34 errors (i.e. hit a wrong key) but 306 of the 320 conducted tasks (80 tasks per participant) did not include errors. Results are shown in Figure 11.5.

We observed mostly comparable results across *fingers*. However, both flight time and key press duration were longer when using the pinkie. Similarly, 30 of the 34 errors were made when using this finger. Regarding the *position* of the magnetic strip we saw no clear impact on the flight time but longer hold times in the first and third positions. This effect particularly shows for the pinkie and index finger (only for the first position). Most errors (30) occurred in the third position. The *configuration* had no clear impact on flight time but impacted key press times with time increasing for attraction compared to repulsion. This is most prominent in the first position. Most errors (30) were made in the attraction conditions. In *summary*, errors mainly occurred for the combination of pinkie, third position, and attraction. Effects were generally more pronounced for the pinkie. While flight time was mostly unaffected we observed longer key press times in the first and partially also third positions that increased with stronger attraction.

11.3.4 User Perception

Participants were asked to rate Likert statements from 1 (totally disagree) to 5 (totally agree) after each task block. The results are shown in Table 11.1. Participants generally rather disagreed with *noticing* the force except for the attraction configuration which was rated neutral. *Comfort* was rated best for the index and middle finger (Mdn=4) and decreased towards the pinkie (Mdn=2). Both position and configuration were rated as neutral (3<Mdn<3.5). Conversely to the comfort, the participants' feeling of being *influenced* increased from the index (Mdn=1) to the pinkie (Mdn=3). Participants felt slightly more influenced in the first and third positions (Mdn=2.5) as well as under the attraction configuration (Mdn=2.5) compared to repulsion and the other positions (Mdn=2).

In the final questionnaire, participants rated the force for positions three (under the middle phalanx) and four (under the proximal phalanx) as the most noticeable. They felt the strongest force for the ring finger and pinkie. Participants liked the off-condition best, followed by attraction at 50% and repulsion at 50%. The least preferred options were both 75% configurations. For a direct comparison, 3 of our 4 participants generally perceived the attraction as stronger than the repulsion. Two participants rated attraction to be more comfortable, while the others found both conditions to be equal.

11.3.5 Limitations

Reflecting the preliminary nature of our study our sample was quite small so results may not generalize to the general public. We also simplified the interaction for the study and used only a single EM at a time. Hence, we have no insights into the interaction effects of multiple EMs (as can be seen in Figure 11.4). Furthermore, we limited the supply current to avoid potential overheating. A more effective cooling mechanism could reduce this for future applications.

11.4 Discussion

Here we discuss the results of our evaluation as well as applications for our approach and next steps to improve it.

11.4.1 How to Influence Users' Key Targeting?

We found, that exerting forces on the pinkie was most effective in influencing users' key targeting. It led to more errors, longer key presses, and flight times. It was also perceived by participants as the most influential. Furthermore, our results show that the exertion of forces on the various fingers affects the key-targeting less the closer the finger is to the thumb. Hence, participants rated the index and middle finger as comfortable but participants felt less influenced. We assume that this is connected to the strength of the fingers and their frequency of use in daily life. However, it also implies, that there is a trade-off: Placing the strip on one of the weaker fingers opens more opportunities for manipulation but was rated less desirable. With regard to the positioning, we observed that placing the strip at the fingertip (first position) led to a longer key press duration that increased when moving from repulsive to attracting configurations. This makes sense, as placing the strip at the fingertip means the force is applied right at the touch-point (and with a longer lever). Overall, placing the strip on top of the fingertip of the index finger may be the best option, as it combines the perceived comfort of the index finger and the observed (but not perceived) possibility for targeted key press time manipulations through different electromagnet configurations.

11.4.2 Applications

To use our prototype in a running system, further tests, adaptations, and extensions will be needed. Nonetheless, we would like to outline some application examples and describe how our approach could either enable or improve them.

The addition of feedback is beneficial in most areas of Human-Computer Interaction (HCI). However, feedback (e.g., vibration to confirm a button press) can only be given *after* an action. By inducing repulsion or attraction we can instead provide *mid-air feed-forward information*. A user can thus *anticipate* the consequences of an action before it is executed (e.g., induced resistance on the enter key could indicate missing information in a form).

Mag(net)ic Teacher

Our approach also has potential for *learning applications*. Pangaro et al. [201] showed, that with additional tracking an array of electromagnets could be used to precisely guide a permanent magnet in a 2D plane. This could be transferred to guiding a user's fingers, e.g., for learning to type with 10 fingers. For learning timing tasks (e.g., playing music or gaming) no tracking is required: users could be guided by attraction and repulsion alone.

Preemptive Auto-Correction

The use of auto-correction is common for most typing applications. However, in those applications, the typed text is changed post-hoc and potentially unbeknownst to the user. Our prototype could improve this by making the correction step explicit. Evaluating the already typed letters in a word it is possible to determine all continuations that will result in a dictionary word and add attraction to valid continuations or repulsion to keys that would generate an invalid word. Also notice how this gives the user agency over following the suggestions on the fly, whereas classical auto-correction will need post-hoc revision if undesired.

Behavior Veiling

Keystroke dynamics allow for seamless and continuous authentication but can also happen unwanted or unnoticed (e.g., a website recognizing users without cookies). In Chapters 9 and 10 we showed that this can be mitigated through intentional behavior change. Our approach enables a low-effort alternative. Through random attraction and repulsion a user's unique *typing patterns could be veiled*, making identification harder or potentially even impossible.

11.4.3 Next Steps

In our work, we only influenced a single finger. While this may be enough for many applications (e.g., mid-air feedback or teaching one finger at a time), other approaches may require being able to influence multiple fingers. One way could be the use of multiple magnetic strips per hand (e.g., to influence the pinkie and thumb which are commonly responsible for using the space bar and enter key), though the magnets could interact and lead to unwanted effects. The question of how to address this remains up for exploration, but one solution may be using gloves with small, embedded electromagnets that can be activated on demand. An additional requirement for many applications is tracking. Park et al. [202] made use of a magnetic ring and the smartwatch magnetometer to identify the finger used for interaction. Our prototype could achieve this by measuring the induced current of the magnetic strip on the electromagnet matrix to determine the finger used. Alternatively, Dai et al. [66] have shown, that using magnetic sensors (below the keyboard in our case) it is possible to track the position and orientation of a permanent magnet.

For our prototype, we made specific decisions with regard to the size and placement of the electromagnets as well as the magnetic strip to generate sufficient force to exert noticeable effects on a user's finger. As a next step, we plan to build smaller magnets to be able to more precisely target keys. This may be achieved by choosing thinner wire to enable more windings or experimenting with stronger permanent magnets. Note, that requirements also strongly depend on the application (see Section 11.4.2) (e.g., guiding a user's finger may require a better resolution but could be subtle and thus use less force). Moreover, while we used a regular keyboard, alternatives such as ergonomic keyboards could also be interesting for future research.

11.5 Implications

In this chapter, we presented the design and implementation of a prototype to exert forces on a user's finger with the goal of influencing key targeting. To achieve this, we generate a magnetic field using a matrix of electromagnets under the keyboard. A permanent magnet on the user's finger serves to transmit the force.

We found our prototype to be able to exert noticeable forces and revealed a *trade-off between comfort and the noticeability of force* when placing the permanent magnet.

This chapter highlights the potential of our approach, not only to be used as a biometric interface to support users in taking agency over their recognition with less active effort but also as a *teaching and feedback tool* with potential applications in many areas of HCI.

V

Discussion & Conclusion

PART V: DISCUSSION & CONCLUSION

In this part, we take a step back from the research we did and discuss implications and extensions of our work.

❖ **Chapter 12** discusses our insights towards designing user-centered biometric interfaces, and use cases beyond design and biometrics. We also take the opportunity to reflect on the methods used in this thesis.

❖ **Chapter 13** concludes this thesis by reflecting on the contributions and giving directions for future work.

12
Discussion & Reflection

In this thesis, we took a user-centered approach to both create and enhance biometric interfaces that can support user literacy and agency. To this end, we took steps to assess user needs and perceptions with regard to their current use of biometrics and suggested improvements to the interfaces they use to interact with biometric systems.

Literacy and agency are two interwoven goals and achieving them also depends on user interest and motivation to take action. Given, that biometric models are often not publicly available and behavioral biometric approaches are not yet widely used, assessing users' perceptions and testing our solutions was challenging. In this chapter, we discuss what we learned about designing biometric interfaces and reflect on the methods we used within this thesis. Finally, we outline how our findings can benefit different target groups and be used beyond the design of biometric interfaces.

12.1 Design Considerations for Biometric Interfaces

Here we give a summary and discussion of our findings regarding the design of biometric interfaces, highlighting when and how to communicate information and support user agency. We conclude this section with a short summary of how to practically approach designing interfaces for biometrics.

12.1.1 User-centered Approach

We strongly focused on user perception and interaction with biometric systems rather than technical aspects like their training or performance of underlying models to improve biometrics in this thesis. This naturally raises the question, of whether user interaction with biometric systems is generally needed or if we should rather try to automate processes where possible and focus on improving the models instead. Based on Cranor [61], security systems should keep humans out of the loop where possible. However, it is important to design for the user whenever this is not the case. For biometrics, we are faced with such a case, as they are inherently about the human (i.e. measuring features of human physiology and behavior) and thus humans also are a huge influencing factor in their performance.

In our work, we tried to follow this approach and make processes more transparent to the user, even though options for automation would have existed. Instead of illustrating potential error cases of a face recognition model to the user (see Chapter 6), we could have tried to automatically find the globally best settings for this user instead. Instead of nudging users to type differently by inducing force on their fingers (see Chapter 11), we could have tried to solve the problem programmatically [252] and thus kept the step invisible to the user. However, this also means, that users would have to rely on a solution not graspable to them and without an option to control it. More generally, keeping the user in the loop gives them both the option to make active choices (e.g. choosing their settings or following or ignoring a nudge) and build a mental model of the underlying model that can help them in their future interactions. That said, such active involvement should be optional and design should strive for a default behavior that is as secure as possible to account for users' willingness to engage with biometric interfaces (compare Section 12.1.5).

> Biometrics inherently require a human in the loop and biometric interfaces should be designed accordingly to allow users to take control and build mental models.

12.1.2 Terminology

When investigating the use and perception of biometric methods we also asked participants to explain, what biometrics are (see Chapter 3). Many participants struggled with this question and made connections to other concepts like biology. However, after seeing the definition, most participants named many correct examples of biometric methods. We believe that terminology may be a problem here, as the term biometrics is often only used in connection with passports while biometric mechanisms in practice are referred to by the specific feature they use (e.g. fingerprint). We thus believe, that communicating on the level of concrete examples may be most effective.

Jain et al. [144] describe features suited as biometrics, including the requirement for them to be (sufficiently) distinctive and permanent. However, in our work, we showed that participants could actively control their typing behavior (Chapters 9 and 10), contradicting those requirements. Related work found similar effects for attack scenarios [159, 265], implying, that the term biometrics may be misleading in this case.

Terminology is a complex construct that grows and evolves over time. However, when communicating functionality to users those terms should be used cautiously to avoid associations that are not related to security or imply more security than the system can offer.

> When communicating the functionality of biometric systems to users, concrete examples should be used. The term biometrics should be used cautiously as it can lead to misconceptions about the function and security of a system.

12.1.3 Leveraging Information to Support User Literacy

In this thesis, we argue for user literacy as a basic building block for an informed and secure use of biometrics. This includes basic knowledge about the functioning of biometric systems and their weaknesses [48, 82, 298]. However, generic information is often not useful, when performance for different users and user groups can strongly differ [294] and also depend on external factors [29]. As such it is important for users to be able to form a correct mental model of how their biometric systems work to securely use them [284]. Based on our work we suggest user-centered sources of information to support this.

Personalized Performance Metrics

In Chapter 6 we introduced a method to generate challenging samples for a decision-making model. We used this approach to explore the weaknesses of a face recognition model, but it can also be applied to single users by embedding their faces in the latent space. While global performance measures like model accuracy may be hard to grasp (what does it practically mean that a model is wrong in 1% of the cases?), our approach allows users to draw their own conclusions by comparing the model ratings on the generated samples to their own expectations. As such the user can actively gain insights into the model and may even be intrigued to explore further.

Predictive Information

In addition to understanding how a biometric system generally performs for a single user, it is also important for them to be able to predict its reaction to changes. As an example, our generative method (Chapter 6) could be extended to produce meaningful alterations to their appearance (e.g. changing their age or pose) [137, 246, 283] to allow them to predict how the biometric model would cope with those changes in reality. The indicators proposed in Chapter 8 allowed a user to gain insights into the current device confidence level and anticipate an upcoming re-authentication. Matching the development of device confidence to events in their life or actively changing their behavior to see the impact can help users build a mental model of how the system will react to changes in the future.

Context Information

Finally, we observed, that not only the users themselves but also their context and surroundings can play a significant role in the performance of a biometric system. In Chapter 7 we explored how an interface could look like, that leverages contextual information to suggest the use of an appropriate authentication mechanism. The advantage of displaying the rationale behind this suggestion (instead of e.g. automatically switching when possible) is again the opportunity for the user to understand which factors impact the biometric system and anticipate them in the future.

> Where possible, biometric interfaces should leverage user-centered information like personalized performance metrics, predictive information, or context information instead of global performance metrics.

12.1.4 When to Present Information

We introduced biometric interfaces as any point where users come in touch with a biometric system. However, not all of those interactions are equally suited for presenting information. In Chapter 5 we found three opportunities for presenting information: one-time (e.g. at enrollment), continuous, or event-based. Here we summarize our findings from exploring those options and our insights as to which information is suitable in those cases.

In Chapter 6 we explored a method to provide information at *enrollment* time. The main goal at this point is to provide users with the necessary information to make an informed decision for or against using a biometric system or which parameters to choose if they are available. Our approach can enable such comparisons by showing results for different recognition thresholds or training samples. A *continuous* indicator can be useful to get insights into the current system status (see Chapter 5), in particular as continuous biometric systems are designed to work in the background and a user may need reassurance that it is still running. Potential reasons for *event-based* interventions can be (upcoming) re-authentications or contextual information relevant to the authentication process. Our work (see Chapter 7) showed, that such contextual information was well received and users followed the suggestions. However, it is also important that perceived reasons match the system explanation to not hamper trust in the system. We also found that users should not be interrupted in their interaction without need, in particular for important tasks (see Chapter 8). Short-term indicators can offer a compromise for necessary interruptions to allow users to finish their current tasks.

One point to consider in all of those cases is to make certain, that only legitimate users get access to this information. Similar to the legitimate user, an attacker could leverage feedback to gain insights into the model or expected inputs and use them to facilitate an attack. As such, information should only be presented after successful authentication.

> Information can be displayed one-time, continuous, or event-based. Displaying event-based information should consider the trade-off between importance and interrupting the user. Information should only be displayed in a secure environment (i.e. after authentication) so that it does not facilitate attacks.

12.1.5 Degrees of Involvement

In this thesis, we often assumed, that users would want to better understand a system and gain more agency over it. However, this is not necessarily the case and many users may be contempt as long as their biometric system works. Designers of biometric interfaces should thus consider how much users should be involved and how much they want to be involved. They should offer solutions for both cases. Using generated samples to illustrate system performance is a good example of this (see Chapter 6). Presenting just sample cases and their respective ratings by the biometric model does not cause any additional effort from the user. In case they are interested, the same approach can be used to build an interactive application that allows for active exploration. We also explored embedding an interface into a game to increase motivation to engage with it and – in the best case – achieve a learning effect the user may not even have looked for in the first place (see Chapter 10). We found our design effective in appealing to the groups of players it was designed for. Finally, an approach like context-based authentication method switches as proposed in Chapter 7 could be automated for users that only want their system to work but still display information about the switch to give explanations to users who are interested. Depending on the used methods this can also preserve the original interaction, for example triggering authentication through touching the fingerprint sensor but using face recognition. Reflecting on those options, it becomes clear, that gaining information or agency should be easy and offer the option to explore more if users are interested. This is particularly true as authentication in itself is already considered a secondary task (in the way of the main interaction goal of the user).

> Biometric interfaces should allow for and encourage involvement to support user literacy and agency but should always expect that users may not want to engage.

12.1.6 Interplay of Literacy and Agency

Throughout this thesis, we often put a stronger focus on either supporting user literacy or user agency. However, the two aspects are often not clearly distinguishable and should not be seen as separate aspects in practice. Biometrics are based on complex machine learning models, so giving users control over aspects like model parameters only makes sense if they understand what they do. Similarly, literacy by itself can be valuable, but options to act on this knowledge are limited as long as users have no agency to take control.

This is also reflected in our work. In Chapter 7 we built an interface to inform users about the impact of contextual factors like moisture on using fingerprint recognition. However, we combined this with a nudge to take control and switch to a different authentication method. In Chapter 11 we proposed the idea to use electromagnets as a way for users to modify their keystrokes without having to actively control their typing. However, the exerted force on the user's finger is still just a nudge, leaving users with full control if they want to follow it. When they decide against it, this nudge still serves as a reminder that typing could be modified and how it would be done. As a final example, we designed *Imitation Game* as a game to support users in gaining control over their typing behavior and subsequently over keystroke dynamic biometric systems (see Chapter 10). However, to learn this skill users need to first understand which aspects of their typing are relevant and how small changes in their typing influence the sensor readings.

> Biometric literacy and agency are interlinked and biometric interfaces should be designed to facilitate both.

12.1.7 Summary: How to Approach Designing Biometric Interfaces

Here, we summarize our insights into the design of user-centered biometric interfaces in the form of a set of questions and considerations for designers to follow. They are based on the user-centered approach we took in this work and thus emphasize user needs as the main driver for design. Our recommendations are based on and extend the design space we established in Chapter 5.

Why is a biometric interface needed? The first consideration should be about the (user) problem or need that should be addressed. This also includes the question of whether it is a purely technical problem or if user insights into the solution are beneficial to them (e.g. to build a mental model or anticipate similar situations).

Who is the target audience? This question is aimed at better understanding how a solution needs to be presented. For end-users, complex terminology should be avoided in favor of examples. For developers, precise terminology may be helpful. This question is also aimed at finding out, how users can be motivated to interact with a biometric interface, e.g. in the form of a game or an incentive structure.

What information can be leveraged? Where possible, personalized information should be preferred over global performance metrics (e.g. how does the system work for my personal situation rather than how good is it in general).

When should the interface be presented? We found enrollment, authentication, and event-based interventions to be possible opportunities to show biometric interfaces. The exact choice depends on the concrete use case. If event-based interfaces are needed, their timing should be well considered to avoid (unnecessary) user interruptions.

How should the interaction look like? Biometric interfaces can be purely informative but should in general contain an option for users to take action. However, this step should also consider, that authentication is a secondary task and users may not be interested in an interaction.

Which method is suited for evaluation? Following a user-centered design approach also means, that interface designs should be evaluated with users. If possible, users should interact with a real interface in the field. Mocking functionality can be a viable way of making this possible.

12.2 Reflection on Methods Used

The representation of risk and external validity of security studies are a general challenge of usable security research [80](see Section 2.5.3). Here we reflect on our choice of research methods made throughout this thesis.

12.2.1 Reflection on Mocking Interactions

Testing biometric interfaces is a difficult task. On one hand, many models currently used are proprietary and thus cannot be modified or used in studies. Other approaches are still in the research phase and thus not widely available. On the other hand, having study participants use self-developed or open-source solutions can put them at risk as they may actually rely on the security of those solutions. The first responsibility always has to be to keep participants and their data secure. We thus often relied on adding an interface on top of existing authentication mechanisms (e.g. a PIN) and mocking (parts of) the functionality of the biometrics systems [65]. Here we reflect on our choices and their implications for our results.

Fully Mocked Interactions

For our investigation of user perception of different biometric mechanisms (Chapter 4), no actual biometric system was used but all decisions were triggered by a human experimenter. However, we designed the environment in a way to give the impression of a real setup, e.g. by showing the pose estimation of a Kinect camera and adding electronics to the door handle. This approach was very effective in convincing users they had interacted with a real system. However, while we simulated error cases, they could not reflect the behavior of a real system and thus limit our findings in that direction. Similarly, the device confidence level of our Authenticator app (Chapter 8) was completely mocked as well. However, here we could not rely on the human judgment of an experimenter to trigger re-authentications as the study was conducted in the wild. While the approach was successful in testing our indicator designs, participants also noticed mismatches with their intuition which was reflected in their ratings. Overall, we find mocked interactions an effective tool for exploring interactions. However, results cannot directly be generalized, as a mocked system can only approximate real behavior.

Supporting Mocked Interactions with Real Data

When nudging users to adapt their authentication method to context factors (Chapter 7) we used a *semi-mocked* approach, i.e. we used real context data like the local weather when it was available but chose a random explanation when this was not the case. We believe this to be an improvement over fully mocked approaches and received no feedback indicating participants noticed some of the given reasons being random. However, participants perceived suggested switches as significantly less appropriate when they were given an explanation which we believe to be caused by a mismatch between the given information and the users' perceived actual context. Overall, the use of semi-mocked interactions can be a good compromise between exploring an unavailable system and simulating real behavior.

Exploring Real Interactions detached from the Biometric System

Finally, we also explored interactions detached from their respective biometric model. An example of this approach is our implementation to explore modifications to typing behavior without involving a real recognition model described in Part IV. As such we did not need to account for participants' security but in turn, gained insights about modifications of typing only on a more fundamental level instead of observing it when interacting with a real biometric system. This approach can be useful to fully explore an interaction but may be limited in its generalizability to use in a different context.

> (Partially) mocking interactions and systems can be an effective tool to explore biometric interfaces. When explanations are given, great care should be taken to avoid mismatches between the presented information and users' perceived reality.

12.2.2 Reflection on taking Studies to the Wild

In this thesis, we conducted both studies in the lab and in the wild but always gave priority to testing under realistic conditions when possible. In Chapters 9 and 10 we explored an approach to extend a study originally designed for the lab to be viable in a remote setting using gamification. Here we reflect on the differences observed from this direct comparison but also our general observations with regards to the different study types.

The probably most prominent difference is an increase in effort when conducting studies in the wild. The design has to account for different user setups and the lack of a researcher present who can spot potential errors or answer questions that may arise. Prototypes thus require extensive testing and iteration. However, even with all precautions, it is hard to foresee all possible cases, making in the wild studies also more prone to errors. That said, the benefit of conducting studies in the wild is gaining better insights into the way users would actually use a system. As an example, we observed that lab participants of our Imitation Game study (Chapter 10) were better at modifying their behavior and reported higher motivation. There are different possible explanations for this, but we believe one main point to be observation bias, leading to remote participants interacting with the system more like they actually would while lab participants had a bias to "perform" better. While this is a weakness we also see this as the strength of lab studies: they can reveal the best-case outcome of using an intervention and thus serve as a goal to strive for and to compare against. While it is less ecologically valid, their results have more internal validity as researchers have more control over the environment and can thus exclude many factors that can impact the outcome of an in-the-wild study.

> Lab studies can reveal the potential of a biometric interface but should be complemented with in-the-wild studies to understand real usage and shortcomings.

12.3 Insights and Use Cases Beyond Biometric Interfaces

This work was mainly concerned with the design, implementation, and evaluation of concepts for biometric interfaces for end-users. Here we discuss, how our results and approaches can be useful beyond this specific use case and target group.

The main focus of this thesis was on end-users and thus our work was aimed at designers of biometric interfaces for end-users. To this end, we uncovered design considerations and user preferences, and suggested new and extended interfaces to convey this information and enable control for users of biometric systems. However, we see value in our results beyond this group. Our insights on user perception and misconceptions with regard to biometrics can be useful for *educators* to give a more nuanced picture of biometrics and their use. *Developers* of new biometric methods can consider the user preferences uncovered in our work and use our method for finding challenging samples for decision-making models (see Chapter 6) as a user-centered approach to inspect their models and to generate new training data to improve them. We believe this approach to also extend to other domains, though small adjustments may be needed to reflect relevant samples for the specific use case. We further introduced and showed the viability of adding explicit behavior modifications as an additional security layer for password systems. For *providers* of biometric systems, we hope this work illustrates the value of giving users control over the model (parameters) as well as the need for clear communication during enrollment to enable informed use of their systems.

Finally, other *researchers* can follow our approach for converting a security study to a game guided by a player-type model [25] to take their work to the wild. Our magnetic keyboard can be leveraged beyond security to convey both feedback and feed-forward information, which can be useful in many areas like text correction or learning to type.

Thinking more broadly, physiological and behavioral measurements become increasingly available through different devices. Some examples include the wide range of physiological and behavioral measurements modern smartwatches provide, tracking of user behavior online or in games, or gaze and motion tracking in VR environments. Possible applications include but are not limited to general health monitoring, adaptive interfaces, or personalized suggestions (e.g. in the form of a running coach). While the context is different, similar questions to the ones found in this thesis may arise: Why was a decision made, how do external factors impact the system, or what options exist to take control over the collected data and its use? We thus believe, that both our design approach and discussed considerations (see Section 12.1) can be a valuable starting point for exploring interfaces for such applications.

13
Conclusion and Future Research Directions

We conclude this thesis with a summary of the contributions made. We further outline directions for future research building on our work before closing with some final remarks.

13.1 Summary of Contributions

With this thesis, we contribute to answering three overarching research questions: what are user needs with regard to biometrics, how can they be supported to acquire literacy about biometric systems and how can interfaces be designed to enable and extend user agency over their biometric systems? Here we summarize our contributions to those questions.

13.1.1 What are user needs and how can they be addressed through the design of biometric interfaces?

We conducted three studies to understand user needs and preferences as well as to uncover design opportunities for biometric interfaces. In Chapter 3 we assessed the use and perception of biometrics in two surveys four years apart. We found that participants struggled with explaining what biometrics were but could name examples. They were worried about changes in their physiology and behavior affecting biometric systems. Despite the increased adoption of biometric methods – in particular face recognition –, we observed no clear difference in biometric literacy between the two surveys. In Chapter 4 we compared user perception of a key, gait recognition, and a palm vein scanner to unlock doors in a Wizard-of-Oz lab study. We found that participants liked the convenience of the biometric methods, but were still concerned about them and valued the agency that the key gave them. In Chapter 5 we contribute a design space on communicating the security of biometric systems derived from a focus group. We establish the dimensions of input, output, and purpose as guiding elements in the design of security indicators.

Together those chapters lay the foundation on which we built our investigation of how to increase user literacy and agency through the design of biometric interfaces.

13.1.2 How can users be supported to acquire biometric literacy through biometric interfaces?

To answer this question we explored points where users come in contact with biometric systems, namely enrollment, authentication, and re-authentication. In Chapter 6 we contribute a method to generate challenging samples that can be used to explore the weaknesses of a biometric model (here face recognition). We propose to generate such samples for single users to give them an impression of the individual performance of the model on their data. In Chapter 7 we propose and evaluate the use of context information to nudge users to choose the appropriate authentication mechanism. At the same time, this approach can contribute to users' knowledge of context factors influencing their biometric system. In Chapter 8 we contribute the design and evaluation of indicators to communicate the system state of a biometric model as well as upcoming re-authentications.

Together, those contributions can help users gain personalized insights into the performance of a biometric system, understand and react to contextual factors influencing it, and anticipate model decisions through introspection into its state.

13.1.3 How can biometric interfaces be leveraged to extend user agency?

Instead of a broad approach, we decided to answer this question with a single biometric system and explore more opportunities in the process. In Chapter 9 we show in a lab study, that users are able to take active control over their typing behavior. This gives users agency over when to be identified and can be used as an additional layer to increase the security of password authentication. In Chapter 10 we show, how this complex lab study can be transformed into a game to make training for this skill available and motivating to users in the wild. With Chapter 11 we explore the use of electromagnets to free users from having to actively control their behavior to achieve typing modifications in the first place. However, this approach still retains user control, as users are only nudged towards the desired modifications.

Together, those chapters show that it is possible to gain agency over a biometric method that does not inherently offer interfaces. We further contribute two methods to make this task more engaging and less effortful for the user.

13.2 Future Research Directions

In the previous sections, we reflected on the work done in this thesis. Looking ahead, we now outline some suggestions for future research directions building on and extending the work done in this thesis.

13.2.1 Extending Proposed Solutions

In this thesis we took a very broad approach, picking many different biometric approaches to illustrate how interfaces can be built for them. However, this also means that it remains an open question if they can be transferred to other biometrics, e.g. if our method for finding challenging samples (Chapter 6) can also be used to generate input for a voice recognition model or if it is limited to face generation. We investigated users' ability to modify their typing (Part IV), but it remains an open question if our findings and solutions transfer to e.g. modifying gait patterns. Future research can use our solutions and findings as a starting point to explore effects across biometric methods.

Similarly, our approaches can also be extended to give deeper insights or additional utility for the biometrics we designed them for. We did not explore how generated samples can be leveraged in real user interfaces to inform users. They also have further potential to be used for training purposes. We proposed to leverage context information to nudge users to switch their authentication mechanism but did not explore the option to automatically make such switches. We explored electromagnets to influence typing behavior but there may be other options that do not require active user involvement, e.g. the use of music. Section 12.3

highlights more potential applications beyond biometrics and end-users as a target group and we illustrate more potential extensions in the respective chapters throughout this thesis.

13.2.2 Real World Deployment

As previously mentioned, real-world deployments for biometric interfaces are challenging as many solutions are proprietary and users' security has to be the first priority. Also, the focus in this thesis was less on the design of actual interfaces but mostly on which elements they should include and when and how to present them. A natural next step would be to try and involve providers of biometric systems to integrate elements of our research into their products. This step would thus also open up the opportunity to design and compare concrete (user) interfaces and test their impact in the wild. This thesis could only provide theoretical insights on whether users would use the proposed solutions but such an approach would enable real insights into their use and perception. Involving providers would also be the only way to enable actual choice. Many current approaches do not allow users to configure them or pick from different models, leaving them only with the choice of either using or not using a biometric method. Collaboration with providers can open up the opportunity to find options that are suitable to be customized by end-users (e.g. adapting a threshold to adjust biometrics to users' personal preference between security and convenience) and which settings should be globally optimized instead.

13.2.3 Coping with Changes

A major challenge for biometric systems are changes over time which can appear both in physiology (e.g. quickly through changing hair styles or slowly through aging processes) but more prominently also in behavioral patterns. Throughout our work, we saw participants being concerned about the impact of such changes, but mainly focused on designs for coping with their consequences by improving the interaction with upcoming re-authentications and helping users predict how and when changes may impact their biometric system. Technical approaches for improving biometric models over time already exist, but it remains unclear how to include users in this continuous learning process. Future research could also further investigate the influence of temporal factors on biometric systems and similar to our work leverage this to educate users and give them agency over when and where data should be used for retraining.

13.2.4 Multi-Biometric Systems

In this work, we focused on a single biometric at a time. However, in practice using multiple different biometric features is possible [143] and can enable better performance [132] or reliability, as strengths and weaknesses of different approaches may cover each other. However, it is unclear, how findings for single-biometric interfaces can transfer to multiple biometrics, opening up new questions for future work. How to communicate the contribution and utility of different biometrics in the ensemble? How can users get a choice in which features to use and when? How can multiple biometrics be used to reduce active re-training and re-authentications? As such, we see the exploration of multiple biometrics (or combinations with other authentication approaches at that) as a natural extension of our work.

13.2.5 Addressing Privacy Concerns of Biometrics

In our studies, we found many users to be concerned about biometric data being recorded or misused (Chapter 3). This points to a more general problem where behavioral biometrics require constant data collection and it is often not clear how this data is stored and used. In Part IV we made contributions to this field by showing how user behavior can be changed to gain control over if and when to be recognized by a biometric system. However, there are many more opportunities for future work to explore how to communicate privacy implications for biometrics and facilitate user choice. What data is used and how can this be made transparent to the user? How to show when data is collected and what options exist to give users control over this process? One related goal could be to find new or adapt existing approaches to use fewer data or only capture data on demand, i.e., when it is needed for authentication.

13.3 Closing Remarks

Biometrics are nowadays the most used authentication mechanisms for mobile devices and it stands to be expected that their relevance will further increase with improving algorithms and sensors in the future. Previous research on the topic was mainly concerned with technical aspects. With our work we lay the foundation for a more user-centered approach proposing to educate users about the shortcomings and consequences of using biometrics and enabling more nuanced user control over their authentication. With the increasing amount of data protected through biometrics we invite other researchers to take our work as a starting point for further continuous research in this direction with a focus on users of biometric systems to shape the secure and informed use of this technology in the future.

VI

APPENDIX

A

Appendix for Chapter 3: Understanding Use and Perception of Biometrics

Questionnaire

In this section, we list the full set of questions asked in our survey. Wherever an _ appears it was replaced by the type of biometric (physiological/behavioral) in the corresponding part of the survey.

Questions in Part *B* (Demographics)

Please answer the following short questions about yourself.

1. How old are you?
2. What is your gender?
3. What is your profession?
4. How would you rate your technical knowledge? (1 (no knowledge) to 5 (very proficient))

Questions in Part *C* (Biometric Methods)

In the following, you will see several questions about biometric methods.

Are you familiar with the concept of (physiological) biometric methods?

Yes Please give a short definition/explanation. Please give a short explanation of how it works

No Please think about what it could be and answer with your thoughts

Questions in Part D (Briefing)

With this definition [see Section 3.2.2] in mind, please answer the following questions:

1. Please name all physiological biometric methods that you know/have heard of.

2. Please name all behavioral biometric methods that you know/have heard of.
3. Do you use biometric methods in your everyday life?

   Yes Please indicate all of them; why do you use them.

   No why not?
4. Beyond the methods that you know about, which physiological characteristics could you imagine to be used for biometrics? Also, give an application in case you have one in mind.
5. Beyond the methods that you know about, which behavioral characteristics could you imagine to be used for biometrics? Also, give an application in case you have one in mind.

Questions in Part *E* (Interlude)

Please answer the following questions about your authentication behavior:

1. Do you own and regularly use a smartphone?
2. What operating system does your smartphone use (e.g. Android, IOS, Windows).
3. What (primary) authentication scheme do you use (i.e. how do you usually unlock your device)? Please select only one option to indicate your primary authentication scheme.
4. In case you use a biometric method, what is your fallback authentication scheme? Fallback refers to the method that you have to enter in case your primary authentication method fails or in some cases on a regular bases (commonly 3 days)

Likert Statements in Part *F* (Biometric Perception)

Please rate the following statements about physiological biometric methods. There are no right or wrong answers. This is only about your perception.

1. Compared to a pin/password, using _ biometrics makes authentication faster.
2. _ biometrics methods are reliable.
3. _ biometric methods are easy to use.
4. Performance of _ biometric methods (security, errors,...) is equal for all users.

5. Compared to a pin/password, using _ biometrics makes authentication more secure.
6. _ biometrics are well suited to protect my personal data.
7. _ biometrics can be faked.
8. I have concerns about my privacy when using _ biometrics.
9. I am concerned that someone might hack my device/account when using _ biometrics.
10. I am concerned that I might have no access to my device/account when using _ biometrics.

Questions in Part *G* (Performance & Security)

Please answer the following questions about your perception of _ biometrics.

1. Do you think someone else could access your device/account if you protect it with _ biometrics?
2. Do you think changes in your _ have an impact on _ biometric systems?
3. What do/would you do, in case your _ biometric system does not work (i.e. you are unable to authenticate with it)?
4. What do you think would happen if someone hacked your _ biometric system (i.e. what would be the practical consequences)? What would you do?
5. If you were to attack a _ biometric system, how would you do it?
6. What do you think can be done to make _ biometric systems more secure?

A.0.1 Questions in Part *H* (Conclusion)

Thank you for your participation. Please answer these final questions and enter your email address below in case you want to participate in the raffle.

1. Are there any questions that you would have liked to answer differently after completing the questionnaire (but were not able to do so because going back was not allowed)? If so: which ones and what changed?
2. Are there any other remarks that you would like to make?
3. Please indicate your email address in case you want to participate in the raffle. Your address will not be associated with your answers and will be deleted after the winners were determined.

Codebook

Themes and groups as they resulted from our thematic analysis. We give counts of unique participants mentioning groups divided by the iteration of the survey (1: 2019, 2:2023).

Theme	Group	1	2	All
Definition	Features Mentioned	45	19	64
	Correct or Related Action Mentioned	32	15	47
	Definition by Example	24	12	36
	Missing Knowledge	48	34	82
	No Answer	53	42	95
Usage of Biometrics	Examples of Usage of Biometric Methods	29	27	56
	Devices/Usecases for which Biometric Methods	19	23	42
	Reasons for using Biometric Methods	4	7	11
	Examples of Avoidance of Biometric Methods	1	3	4
	Reasons for not using Biometric Methods	14	5	19
Attacks & Challenges	Non-Malicious: Generic	18	23	41
	Non-Malicious: Physical	20	24	44
	Non-Malicious: Behavioural	2	7	9
	Attack: Software based	13	10	23
	Attack: Force	7	9	16
	Attack: Immitation/Replay	36	36	72
	Attack: Other	11	5	16
	Perceived Weakness	29	20	49
	No Impact	9	4	13
Consequences	Action: Fall back to other Method	50	43	93
	Action: Damage Control	24	21	45
	Action: Seek Support	24	10	34
	Action: Other	5	3	8
	Damage: Access and Missuse	17	18	35
	Damage: Loss	17	11	28
	Damage: None	5	6	11
	Other Consequence	12	3	15
Future Suggestions & Improvements	Novel Biometrics	37	34	71
	Future Applications	7	5	12
	Improvements	27	30	57

B

Appendix for Chapter 9: Exploring Intentional Keystroke Control

Calculating Entropy of Modified Passwords

For a random password with no modifications of length n on the alphabet Σ we calculate entropy E as:

$$E_0 = log_2(|\Sigma|^n)$$

For one modification we choose a password first and then add a single modification at a random location. There are 7 possible modifications (assuming that one manifestation of each feature would be the default (e.g., pressing keys in the centre). Finally we exclude the single case where a flight time would be applied to the first character (as it does not have a preceding character to measure flight time from). This yields:

$$E_1 = log_2(|\Sigma|^n \cdot (7n - 1))$$

Analogous, we calculate the entropy for two modifications by choosing a password first and then either applying two modifications on one character (15 options) or two single modifications; again excluding cases where a flight time modification would be applied to the first character.

$$E_2 = log_2(|\Sigma|^n \cdot \underbrace{((15n - 6)}_{\text{2 on one}} + \underbrace{(\frac{7n \cdot 7(n-1)}{2} - 7(n-1))))}_{\text{2 single}}$$

We calculate entropy for three modifications analogously, taking into account the possibility of three modifications on one character (line 1), two modifications on one character combined with a single modification (line 2) and three single modifications (line 3):

$$\begin{aligned}E_3 = log_2(&|\Sigma|^n \cdot ((13n - 9)\\ &+ (15n \cdot 7(n-1) - 57(n-1))\\ &+ (\frac{7n \cdot 7(n-1) \cdot 7(n-2)}{6} - \frac{7(n-1) \cdot 7(n-2)}{2}))\end{aligned}$$

C

Appendix for Chapter 10: Imitation Game

Questionnaire Results

Here we show the full results of our statistical analysis of Likert statements. All statements were rated on a 7-point Likert scale (1=strongly disagree, 7=strongly agree). However, for calculating the UEQ-S scores we subtracted 4, as indicated [238].

This Table shows the differences between *settings*, i.e. the results of Mann-Whitney tests comparing questionnaire answers of remote and lab participants (only from the first session). Subscripts indicate the setting (L: lab, R: remote).

question	n_L	n_R	p	U_L	U_R	$mean_L$	$mean_L$
authentication: Using passwords for authentication is secure	24	24	0.236	255	321	5.208	4.833
authentication: Using typing behavior for authentication is secure	24	24	0.248	255	321	4.083	4.417
authentication: Using passwords together with typing behavior is secure	24	24	0.374	272.5	303.5	5.375	5.292
authentication: I would use a system that uses my typing behavior for authentication	24	24	0.187	245	331	3.417	3.917
attributes: Controling flight time was	23	24	0.213	239	313	3.609	4.000
attributes: Controling hold time was	24	23	0.368	260	292	4.250	4.391
attributes: Controling touch offset/position was	24	24	0.008	172.5	403.5	4.208	5.333
attributes: Controling pressure was	23	17	0.395	185.5	205.5	5.087	5.059
tasks: I was able to adjust to the specified behavior	24	24	0.166	242	334	4.708	5.083
tasks: I was successful in completing the tasks	24	24	0.364	271.5	304.5	5.167	5.458
tasks: The tasks were difficult for me	24	24	0.071	218.5	357.5	4.833	4.167
tasks: I can influence my own typing behavior	24	24	0.027	198.5	377.5	4.583	5.417
tasks: I would also play this game outside a study setting	24	24	0.215	250	326	3.250	2.917
tasks: I improved at producing the specified behavior	24	24	0.323	266	310	4.542	4.875
liked: the character attributes	23	24	0.261	246	306	4.783	4.708
liked: the story	24	24	0.002	146.5	429.5	5.333	3.708
liked: the dialogues	24	24	0.004	160.5	415.5	5.083	3.708
liked: the typing visualisation	24	24	0.182	245	331	5.000	4.833
liked: the colored feedback (training)	23	23	0.254	234.5	294.5	4.957	4.739
liked: the mission scores	24	24	0.156	239.5	336.5	5.167	4.750
liked: the high score	22	23	0.225	220	286	4.636	4.348
liked: creating challenges	20	22	0.281	197	243	4.600	4.409
liked: taking challenges	22	23	0.125	203	303	4.773	4.304
motivated: the character attributes	24	24	0.300	262.5	313.5	4.083	3.792
motivated: the story	24	24	0.029	196.5	379.5	4.583	3.500
motivated: the dialogues	24	24	0.054	210.5	365.5	4.458	3.583

motivated: the typing visualisation	24	24	0.321	265.5	310.5	4.708	4.542
motivated: the colored feedback (training)	23	24	0.365	259.5	292.5	4.522	4.458
motivated: the mission scores	24	24	0.190	246	330	5.125	4.875
motivated: the high score	23	23	0.113	210	319	4.696	4.130
motivated: creating challenges	22	22	0.034	166	318	4.500	3.818
motivated: taking challenges	22	23	0.035	174.5	331.5	4.545	3.696
skipped: I skipped dialogues	24	24	0.055	211	365	3.833	4.792
UEQ: UEQ pragmatic	24	24	0.072	217	359	1.229	0.844
UEQ: UEQ hedonic	24	24	0.002	148	428	1.125	0.073
UEQ: UEQ overall	24	24	0.002	149.5	426.5	1.177	0.458

This Table shows the difference between *sessions* in our study, i.e. the results of Wilcoxon tests comparing questionnaire answers of lab participants in the first and second session (indicated in subscript)

question	n	p	Z	mean$_1$	mean$_1$
authentication: Using passwords for authentication is secure	24	0.049	28	5.208	5.667
authentication: Using typing behavior for authentication is secure	24	0.662	52.5	4.083	3.917
authentication: Using passwords together with typing behavior is secure	24	0.830	42.5	5.375	5.333
authentication: I would use a system that uses my typing behavior for authentication	24	0.693	68.5	3.417	3.542
attributes: Controling flight time was	23	0.228	45	3.609	4.000
attributes: Controling hold time was	23	0.163	54	4.250	3.696
attributes: Controling touch offset/position was	24	0.252	34.5	4.208	3.833
attributes: Controling pressure was	23	0.035	24.5	5.087	4.417
tasks: I was able to adjust to the specified behavior	24	0.033	16.5	4.708	5.250
tasks: I was successful in completing the tasks	24	0.018	7	5.167	5.792
tasks: The tasks were difficult for me	24	0.051	48	4.833	4.208
tasks: I can influence my own typing behavior	24	0.234	20	4.583	5.000
tasks: I would also play this game outside a study setting	24	0.368	51	3.250	3.042
tasks: I improved at producing the specified behavior	24	0.102	32	4.542	5.000
liked: the character attributes	22	0.299	48.5	4.783	5.130
liked: the story	22	0.600	44.5	5.333	5.591
liked: the dialogues	23	0.357	38	5.083	5.478
liked: the typing visualisation	24	0.044	30	5.000	5.625
liked: the colored feedback (training)	23	0.272	41	4.957	5.348
liked: the mission scores	24	0.172	42	5.167	5.583
liked: the high score	21	0.271	30	4.636	5.000
liked: creating challenges	19	0.586	27	4.600	4.857
liked: taking challenges	21	0.604	32.5	4.773	4.591
motivated: the character attributes	23	0.149	35	4.083	4.565
motivated: the story	23	1.000	52.5	4.583	4.565
motivated: the dialogues	23	0.751	47.5	4.458	4.565
motivated: the typing visualisation	24	0.275	54	4.708	5.000
motivated: the colored feedback (training)	22	0.748	41	4.522	4.636
motivated: the mission scores	24	0.938	66.5	5.125	5.125
motivated: the high score	22	0.849	49.5	4.696	4.727
motivated: creating challenges	21	0.307	36.5	4.500	4.048

motivated: taking challenges	21	0.392	39	4.545	4.182
skipped: I skipped dialogues	24	0.002	11.5	3.833	5.750
UEQ: UEQ pragmatic	24	0.267	93	1.229	1.427
UEQ: UEQ hedonic	24	0.442	93.5	1.125	1.042
UEQ: UEQ overall	24	0.958	114	1.177	1.234

Typing Data Results

This table shows our statistical tests on participants *ability to modify* their behavior (e.g. pressing significantly longer when presented with a flight time modification). We used paired sample t-tests for normally distributed data as indicated by Shapiro-Wilk tests. If the normality assumption was violated ($p<0.05$), we used the Wilcoxon signed rank tests instead. Mean$_d$ describes the mean for default (i.e. unmodified) behavior and mean$_m$ for modified.

measure	setting	n	test	p	Z or t	mean$_d$	mean$_m$
flighttime	lab	24	Wilcoxon	0.000	0.000	404.250	926.845
holdtime	lab	24	Wilcoxon	0.000	1.000	90.069	371.655
pressure	lab	24	Wilcoxon	0.178	102.000	0.182	0.185
X offset from bottom target	lab	24	t-test	0.242	1.201	0.008	−0.005
X offset from left target	lab	24	t-test	0.000	5.061	0.008	−0.109
X offset from right target	lab	24	t-test	0.016	−2.593	0.008	0.082
X offset from top target	lab	24	t-test	0.707	−0.380	0.008	0.013
Y offset from bottom target	lab	24	t-test	0.000	−8.818	0.165	0.420
Y offset from left target	lab	24	t-test	0.069	1.906	0.165	0.142
Y offset from right target	lab	24	Wilcoxon	0.546	128.000	0.165	0.179
Y offset from top target	lab	24	t-test	0.000	9.875	0.165	−0.232
flighttime	remote	24	Wilcoxon	0.000	0.000	259.607	910.313
holdtime	remote	24	t-test	0.000	−12.239	91.060	451.301
X offset from bottom target	remote	24	t-test	0.769	0.297	−0.028	−0.031
X offset from left target	remote	24	Wilcoxon	0.900	145.000	−0.028	−0.043
X offset from right target	remote	24	t-test	0.466	0.741	−0.028	−0.047
X offset from top target	remote	24	t-test	0.036	−2.233	−0.028	−0.002
Y offset from bottom target	remote	24	t-test	0.000	−8.285	0.118	0.385
Y offset from left target	remote	24	t-test	0.714	0.372	0.118	0.113
Y offset from right target	remote	24	t-test	0.466	−0.741	0.118	0.131
Y offset from top target	remote	24	t-test	0.000	7.205	0.118	−0.159

This table shows our statistical tests on participants' *deviations* from the expected modification value (e.g. how closely did a participant produce an expected hold time). We used paired sample t-tests for normally distributed data as indicated by Shapiro-Wilk tests. If the normality assumption was violated ($p<0.05$), we used the Wilcoxon signed rank tests instead. Mean$_d$ describes the mean for default (i.e. unmodified) behavior and mean$_m$ for modified.

deviation	setting	n	test	p	Z or t	mean$_d$	mean$_m$
flighttime	lab	24	Wilcoxon	0.000	22.000	197.887	342.726
holdtime	lab	24	Wilcoxon	0.000	0.000	27.293	200.225
pressure	lab	24	Wilcoxon	0.000	0.000	0.025	0.216
offset from bottom target	lab	24	t-test	0.337	0.980	0.270	0.251
offset from left target	lab	24	t-test	0.092	−1.758	0.270	0.298
offset from right target	lab	24	t-test	0.001	−3.832	0.270	0.341
offset from top target	lab	24	Wilcoxon	0.197	104.000	0.270	0.339
flighttime	remote	24	Wilcoxon	0.000	0.000	134.724	339.692
holdtime	remote	24	Wilcoxon	0.000	0.000	25.305	242.687
offset from bottom target	remote	24	t-test	0.002	−3.439	0.256	0.391
offset from left target	remote	24	t-test	0.000	−9.163	0.256	0.444
offset from right target	remote	24	t-test	0.000	−10.860	0.256	0.519
offset from top target	remote	24	t-test	0.000	−6.867	0.256	0.594

D

Appendix for Chapter 11: Supporting Key Targeting using Electromagnets

Schematics of the electronics

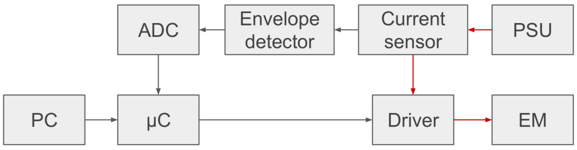


Figure D.1: Overview of the components of our prototype and their interactions. Red lines indicate a high current. The power provided by the power supply unit (PSU) is monitored by the current sensor. Hence, the current sensor measures the current that drivers and electromagnet (EM) are consuming and sends the measured signal to the envelope detector. The signal's envelope continues to the analog-to-digital converter (ADC). The digital data is then processed by the microcontroller (μC), which communicates with the motor drivers to control the EM.

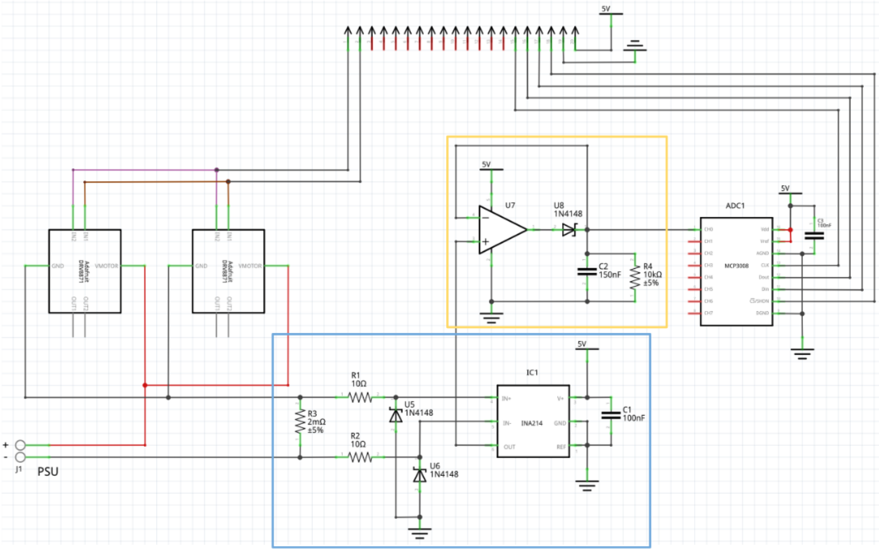


Figure D.2: Schematic of our circuit except for the microcontroller. The current sensor is marked blue, and the envelope detector yellow. The ADC is shared between EMs.

BIBLIOGRAPHY

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space?. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4432–4441.

[2] Yomna Abdelrahman, Mohamed Khamis, Stefan Schneegass, and Florian Alt. 2017. Stay cool! understanding thermal attacks on mobile-based user authentication. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3751–3763.

[3] Piotr D Adamczyk and Brian P Bailey. 2004. If not now, when?: the effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 271–278.

[4] Anne Adams and Martina Angela Sasse. 1999. Users Are Not the Enemy. *Commun. ACM* 42, 12 (dec 1999), 40–46. DOI:http://dx.doi.org/10.1145/322796.322806

[5] Lalit Agarwal, Hassan Khan, and Urs Hengartner. 2016. Ask me again but don't annoy me: Evaluating re-authentication strategies for smartphones. In *Symposium on Usable Privacy and Security (SOUPS).*

[6] Divyansh Aggarwal, Jiayu Zhou, and Anil K Jain. 2021. Fedface: Collaborative learning of face recognition model. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 1–8.

[7] Sohaib Ahmad and Benjamin Fuller. 2020. Resist: Reconstruction of irises from templates. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 1–10.

[8] Thamer Alhussain, Rayed AlGhamdi, Salem Alkhalaf, and Osama Alfarraj. 2013. Users' Perceptions of Mobile Phone Security: A Survey Study in the Kingdom of Saudi Arabia. *international journal of computer theory and engineering* 5, 5 (2013), 793.

[9] Florian Alt, Andreas Bulling, Gino Gravanis, and Daniel Buschek. 2015. GravitySpot: guiding users in front of public displays using on-screen visual cues. In *Proceedings*

of the 28th Annual ACM Symposium on User Interface Software & Technology. ACM, 47–56.

[10] Norman G Altman. 1968. Palm print identification system. U.S. Patent US3581282A. (3 December 1968).

[11] Selay Arkün Kocadere and Şeyma Çağlar Özhan. 2018. Gamification from Player Type Perspective: A Case Study. *Educational Technology & Society* 21 (07 2018), 1436–4522.

[12] Parul Arora, Madasu Hanmandlu, and Smriti Srivastava. 2015. Gait based authentication using gait information image features. *Pattern Recognition Letters* 68 (2015), 336–342.

[13] Sunpreet S Arora, Kai Cao, Anil K Jain, and Nicholas G Paulter. 2016. Design and fabrication of 3D fingerprint targets. *IEEE Transactions on Information Forensics and Security* 11, 10 (2016), 2284–2297.

[14] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, and others. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[15] Farzaneh Asgharpour, Debin Liu, and L. Jean Camp. 2007. Mental Models of Security Risks. In *Financial Cryptography and Data Security*, Sven Dietrich and Rachna Dhamija (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 367–377.

[16] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2018. Synthesizing robust adversarial examples. In *International conference on machine learning*. PMLR, 284–293.

[17] Adam J. Aviv, Katherine Gibson, Evan Mossop, Matt Blaze, and Jonathan M. Smith. 2010. Smudge Attacks on Smartphone Touch Screens. In *Proceedings of the 4th USENIX Conference on Offensive Technologies (WOOT'10)*. USENIX Association, Berkeley, CA, USA, 1–7. http://dl.acm.org/citation.cfm?id=1925004.1925009

[18] Brian P Bailey, Joseph A Konstan, and John V Carlis. 2001. The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface.. In *Interact*, Vol. 1. 593–601.

[19] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. 2021. Towards causal benchmarking of bias in face analysis algorithms. In *Deep Learning-Based Face Analytics*. Springer, 327–359.

[20] Michal Balazia and Petr Sojka. 2018. Gait recognition from motion capture data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 1s (2018), 22.

[21] Matthias Baldauf, Sebastian Steiner, Mohamed Khamis, and Sarah-Kristin Thiel. 2019. Investigating the User Experience of Smartphone Authentication Schemes-The Role of the Mobile Context. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. `https://hdl.handle.net/10125/59918`

[22] Tyler Baldwin and Joyce Chai. 2012. Towards Online Adaptation and Personalization of Key-target Resizing for Mobile Devices. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces (IUI '12)*. ACM, New York, NY, USA, 11–20. `DOI:http://dx.doi.org/10.1145/2166966.2166969`

[23] Lucas Ballard, Daniel Lopresti, and Fabian Monrose. 2007. Forgery quality and its implications for behavioral biometric security. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 37, 5 (2007), 1107–1118.

[24] Jakob E Bardram, Rasmus E Kjær, and Michael Ø Pedersen. 2003. Context-aware user authentication–supporting proximity-based login in pervasive computing. In *International Conference on Ubiquitous Computing*. Springer, 107–123.

[25] Richard Bartle. 1996. Hearts, clubs, diamonds, spades: Players who suit MUDs. *Journal of MUD research* 1, 1 (1996), 19.

[26] Lewis Bell, Jay Lees, Will Smith, Charlie Harding, Ben Lee, and Daniel Bennett. 2020. PauseBoard: A Force-Feedback Keyboard for Unintrusively Encouraging Regular Typing Breaks. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[27] Adrien Bennetot, Jean-Luc Laurent, Raja Chatila, and Natalia Díaz-Rodríguez. 2019. Towards explainable neural-symbolic visual reasoning. *arXiv preprint arXiv:1909.09065* (2019).

[28] Joanna Bergstrom-Lehtovirta and Antti Oulasvirta. 2014. Modeling the Functional Area of the Thumb on Mobile Touchscreen Surfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 1991–2000. `DOI:http://dx.doi.org/10.1145/2556288.2557354`

[29] Rasekhar Bhagavatula, Blase Ur, Kevin Iacovino, Su Mon Kywe, Lorrie Faith Cranor, and Marios Savvides. 2015. Biometric authentication on iphone and android: Usability, perceptions, and influences on adoption. (2015). `https://doi.org/10.14722/usec.2015.23003`

[30] B Bharathi and P Bindhu Shamily. 2020. A review on iris recognition system for person identification. *International Journal of Computational Biology and Drug Design* 13, 3 (2020), 316–331.

[31] Robert Biddle, Sonia Chiasson, and Paul C Van Oorschot. 2012. Graphical passwords: Learning from the first twelve years. *ACM Computing Surveys (CSUR)* 44, 4 (2012), 19.

[32] Jonathan Bishop. 2014. *Gamification for Human Factors Integration: Social, Education, and Psychological Issues: Social, Education, and Psychological Issues*. IGI Global.

[33] Serena Booth, Yilun Zhou, Ankit Shah, and Julie Shah. 2021. Bayes-trex: a bayesian sampling approach to model transparency by example. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11423–11432.

[34] Virginia Braun and Victoria Clarke. 2021. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy Research* 21, 1 (2021), 37–47.

[35] Arnaud Buchoux and Nathan L Clarke. 2008. Deployment of keystroke analysis on a smartphone. (2008).

[36] Oliver Buckley and Jason RC Nurse. 2019. The language of biometrics: Analysing public perceptions. *Journal of Information Security and Applications* 47 (2019), 112–119.

[37] Andreas Bulling, Florian Alt, and Albrecht Schmidt. 2012. Increasing the security of gaze-based cued-recall graphical passwords using saliency masks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3011–3020.

[38] Ulrich Burgbacher and Klaus Hinrichs. 2014. An Implicit Author Verification System for Text Messages Based on Gesture Typing Biometrics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2951–2954. DOI:http://dx.doi.org/10.1145/2556288.2557346

[39] Attaullah Buriro, Bruno Crispo, Filippo Del Frari, and Konrad Wrona. 2015. Touchstroke: smartphone user authentication based on touch-typing biometrics. In *International Conference on Image Analysis and Processing*. Springer, 27–34.

[40] Daniel Buschek and Florian Alt. 2015. TouchML: A Machine Learning Toolkit for Modelling Spatial Touch Targeting Behaviour. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA.

[41] Daniel Buschek, Benjamin Bisinger, and Florian Alt. 2018. ResearchIME: A Mobile Keyboard Application for Studying Free Typing Behaviour in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 255, 14 pages. DOI:http://dx.doi.org/10.1145/3173574.3173829

[42] Daniel Buschek, Alexander De Luca, and Florian Alt. 2015. Improving Accuracy, Applicability and Usability of Keystroke Biometrics on Mobile Touchscreen Devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1393–1402. DOI:http://dx.doi.org/10.1145/2702123.2702252

[43] Daniel Buschek, Alexander De Luca, and Florian Alt. 2016a. Evaluating the Influence of Targets and Hand Postures on Touch-based Behavioural Biometrics. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1349–1361. DOI:http://dx.doi.org/10.1145/2858036.2858165

[44] Daniel Buschek, Fabian Hartmann, Emanuel von Zezschwitz, Alexander De Luca, and Florian Alt. 2016b. SnapApp: Reducing Authentication Overhead with a Time-Constrained Fast Unlock Option. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3736–3747. DOI:http://dx.doi.org/10.1145/2858036.2858164

[45] Daniel Buschek, Mariam Hassib, and Florian Alt. 2018. Personal mobile messaging in context: Chat augmentations for expressiveness and awareness. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 4 (2018), 1–33.

[46] Karoline Busse, Sabrina Amft, Daniel Hecker, and Emanuel von Zezschwitz. 2019. "Get a Free Item Pack with Every Activation!" Do Incentives Increase the Adoption Rates of Two-Factor Authentication? *i-com* 18, 3 (2019), 217–236.

[47] Sookeun Byun and Sang-Eun Byun. 2013. Exploring perceptions toward biometric technology in service encounters: a comparison of current users and potential adopters. *Behaviour & Information Technology* 32, 3 (2013), 217–230.

[48] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 46–56.

[49] P Campisi, E Maiorana, M Lo Bosco, and A Neri. 2009. User authentication using keystroke dynamics for cellular phones. *IET Signal Processing* 3, 4 (2009), 333–341.

[50] Kai Cao and Anil K Jain. 2014. Learning fingerprint reconstruction: From minutiae to image. *IEEE Transactions on information forensics and security* 10, 1 (2014), 104–117.

[51] Tom Carter, Sue Ann Seah, Benjamin Long, Bruce Drinkwater, and Sriram Subramanian. 2013. UltraHaptics: Multi-Point Mid-Air Haptic Feedback for Touch Surfaces. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13)*. Association for Computing Machinery, New York, NY, USA, 505–514. DOI:http://dx.doi.org/10.1145/2501988.2502018

[52] Darryl Charles, Michael Mcneill, Moira Mcalister, Michaela Black, Adrian Moore, Karl Stringer, Julian Kücklich, and Aphra Kerr. 2005. Player-centred game design: Player modelling and adaptive digital games. *Proceedings of DiGRA 2005 Conference: Changing Views - Worlds in Play* (01 2005).

[53] Robert Chen, Roger Levy, and Tiwalayo Eisape. 2021. On factors influencing typing time: Insights from a viral online typing game. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 43.

[54] Sungzoon Cho and Seongseob Hwang. 2005. Artificial Rhythms and Cues for Keystroke Dynamics Based Authentication. In *Advances in Biometrics*, David Zhang and Anil K Jain (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 626–632.

[55] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[56] Tarang Chugh, Sunpreet S Arora, Anil K Jain, and Nicholas G Paulter. 2017. Benchmarking fingerprint minutiae extractors. In *2017 International conference of the biometrics special interest group (BIOSIG)*. IEEE, 1–8.

[57] Nathan L Clarke and Steven M Furnell. 2007. Authenticating mobile phone users using keystroke analysis. *International Journal of Information Security* 6, 1 (2007), 1–14.

[58] Nathan L Clarke, Sevasti Karatzouni, and Steven M Furnell. 2009. Flexible and transparent user authentication for mobile devices. In *IFIP International Information Security Conference*. Springer, 1–12.

[59] Laurent Colbois, Tiago de Freitas Pereira, and Sébastien Marcel. 2021. On the use of automatically generated synthetic image datasets for benchmarking face recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 1–8.

[60] Gennaro Costagliola, Mattia De Rosa, Vittorio Fuccella, and Fabrizio Torre. 2012. TypeJump: A typing game for KeyScretch. In *2012 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 241–242.

[61] Lorrie Faith Cranor. 2008. A framework for reasoning about the human in the loop. In *Proceedings of the 1st Conference on Usability, Psychology, and Security*. 1–15.

[62] Heather Crawford and Karen Renaud. 2014. Understanding user perceptions of transparent authentication on a mobile device. *Journal of Trust Management* 1, 1 (2014), 7.

[63] Heather Crawford, Karen Renaud, and Tim Storer. 2013. A framework for continuous, transparent mobile device authentication. *Computers & Security* 39 (2013), 127–136.

[64] James E Cutting and Lynn T Kozlowski. 1977. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the psychonomic society* 9, 5 (1977), 353–356.

[65] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*. 193–200.

[66] Houde Dai, Wanan Yang, Xuke Xia, Shijian Su, and Kui Ma. 2016. A three-axis magnetic sensor array system for permanent magnet tracking. In *2016 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 476–480.

[67] Hai Dang, Lukas Mecke, and Daniel Buschek. 2022. GANSlider: How Users Control Generative Models for Images using Multiple Sliders with and without Feedforward Information. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[68] Antitza Dantcheva, Petros Elia, and Arun Ross. 2015. What else does your biometric data reveal? A survey on soft biometrics. *IEEE Transactions on Information Forensics and Security* 11, 3 (2015), 441–467.

[69] Xavier de Carné de Carnavalet and Mohammad Mannan. 2014. From Very Weak to Very Strong: Analyzing Password-Strength Meters. In *NDSS*, Vol. 14. 23–26.

[70] Alexander De Luca, Bernhard Frauendienst, Max Maurer, and Doris Hausen. 2010. On the Design of a "Moody" Keyboard. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems (DIS '10)*. Association for Computing Machinery, New York, NY, USA, 236–239. DOI:http://dx.doi.org/10.1145/1858171.1858213

[71] Alexander De Luca, Alina Hang, Frederik Brudy, Christian Lindner, and Heinrich Hussmann. 2012. Touch me once and i know it's you!: implicit authentication based on touch screen patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 987–996.

[72] Alexander De Luca, Emanuel Von Zezschwitz, Ngo Dieu Huong Nguyen, Max-Emanuel Maurer, Elisa Rubegni, Marcello Paolo Scipioni, and Marc Langheinrich. 2013. Back-of-device authentication on smartphones. In *Proceedings of the sigchi conference on human factors in computing systems*. 2389–2398.

[73] Maria De Marsico, Alfredo Petrosino, and Stefano Ricciardi. 2016. Iris recognition through machine learning techniques: A survey. *Pattern Recognition Letters* 82 (2016), 106–115.

[74] Debayan Deb, Jianbang Zhang, and Anil K Jain. 2019. Advfaces: Adversarial face synthesis. *arXiv preprint arXiv:1908.05008* (2019).

[75] Emily Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, and Andrew Zaldivar. 2019. Image counterfactual sensitivity analysis for detecting unintended bias. *arXiv preprint arXiv:1906.06439* (2019).

[76] Mohammad Omar Derawi, Claudia Nickel, Patrick Bours, and Christoph Busch. 2010. Unobtrusive user-authentication on mobile phones using biometric gait recognition. In *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2010 Sixth International Conference on*. IEEE, 306–311.

[77] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From Game Design Elements to Gamefulness: Defining "Gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (MindTrek '11)*. Association for Computing Machinery, New York, NY, USA, 9–15. DOI:http://dx.doi.org/10.1145/2181037.2181040

[78] Anind K. Dey, Raffay Hamid, Chris Beckmann, Ian Li, and Daniel Hsu. 2004. A CAPpella: Programming by Demonstration of Context-aware Applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 33–40. DOI:http://dx.doi.org/10.1145/985692.985697

[79] Rhea Diamond and Susan Carey. 1986. Why faces are and are not special: an effect of expertise. *Journal of experimental psychology: general* 115, 2 (1986), 107.

[80] Verena Distler, Matthias Fassl, Hana Habib, Katharina Krombholz, Gabriele Lenzini, Carine Lallemand, Lorrie Faith Cranor, and Vincent Koenig. 2021. A systematic literature review of empirical methods and risk representation in usable privacy and security research. *ACM Transactions on Computer-Human Interaction (TOCHI)* 28, 6 (2021), 1–50.

[81] Benjamin Draffin, Jiang Zhu, and Joy Zhang. 2014. KeySens: Passive User Authentication through Micro-behavior Modeling of Soft Keyboard Interaction. In *Mobile Computing, Applications, and Services*, Gérard Memmi and Ulf Blanke (Eds.). Springer International Publishing, Cham, 184–201.

[82] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. 2020. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society* 1, 2 (2020), 89–103.

[83] Serge Egelman, Jennifer King, Robert C Miller, Nick Ragouzis, and Erika Shehan. 2007. Security user studies: methodologies and best practices. In *CHI'07 extended abstracts on Human factors in computing systems*. 2833–2836.

[84] Malin Eiband, Mohamed Khamis, Emanuel von Zezschwitz, Heinrich Hussmann, and Florian Alt. 2017. 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)*. Association for Computing Machinery, New York, NY, USA, 4254–4265. DOI:http://dx.doi.org/10.1145/3025453.3025636

[85] Elakkiya Ellavarason, Richard Guest, Farzin Deravi, Raul Sanchez-Riello, and Barbara Corsetti. 2020. Touch-dynamics based behavioural biometrics on mobile devices–a review from a usability and performance perspective. *ACM Computing Surveys (CSUR)* 53, 6 (2020), 1–36.

[86] Stephen J Elliott, Sarah A Massie, and Mathias J Sutton. 2007. The perception of biometric technology: A survey. In *Automatic Identification Advanced Technologies, 2007 IEEE Workshop on*. IEEE, 259–264.

[87] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. 2018. GANSynth: Adversarial Neural Audio Synthesis. In *International Conference on Learning Representations*.

[88] Joshua J Engelsma, Kai Cao, and Anil K Jain. 2019. Learning a fixed-length fingerprint representation. *IEEE transactions on pattern analysis and machine intelligence* 43, 6 (2019), 1981–1997.

[89] Joshua J Engelsma, Debayan Deb, Kai Cao, Anjoo Bhatnagar, Prem S Sudhish, and Anil K Jain. 2021. Infant-ID: Fingerprints for global good. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 7 (2021), 3543–3559.

[90] Clayton Epp, Michael Lippold, and Regan L Mandryk. 2011. Identifying emotional states using keystroke dynamics. In *Proceedings of the sigchi conference on human factors in computing systems*. 715–724.

[91] Michael Fairhurst and Márjory Da Costa-Abreu. 2011. Using keystroke dynamics for gender identification in social network environment. In *4th International Conference on Imaging for Crime Detection and Prevention 2011 (ICDP 2011)*. IET, 1–6.

[92] Adrienne Porter Felt, Robert W Reeder, Alex Ainslie, Helen Harris, Max Walker, Christopher Thompson, Mustafa Embre Acer, Elisabeth Morant, and Sunny Consolvo. 2016. Rethinking Connection Security Indicators. In *Proc. of SOUPS'16*. 1–14.

[93] Tao Feng, Ziyi Liu, Kyeong-An Kwon, Weidong Shi, Bogdan Carbunar, Yifei Jiang, and Nhung Nguyen. 2012. Continuous mobile authentication using touchscreen gestures. In *Homeland Security (HST), 2012 IEEE Conference on Technologies for*. Citeseer, 451–456.

[94] Bill Ferguson. 2015. Reality Is Broken by J. McGonigal. *Games for health journal* 1 (07 2015), 77–8. DOI:http://dx.doi.org/10.1089/g4h.2012.1013

[95] Lauren S Ferro, Steffen P Walz, and Stefan Greuter. 2013. Towards personalised, gamified systems: an investigation into game design, personality and player typologies. In *Proceedings of the 9th Australasian conference on interactive entertainment: Matters of life and death*. 1–6.

[96] Joel E Fischer, Chris Greenhalgh, and Steve Benford. 2011. Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*. ACM, 181–190.

[97] Andrew Dathan Frankel and Muthucumaru Maheswaran. 2009. Feasibility of a Socially Aware Authentication Scheme. In *2009 6th IEEE Consumer Communications and Networking Conference*. 1–6. DOI:http://dx.doi.org/10.1109/CCNC.2009.4784910

[98] Christie Franks and Russell G Smith. 2020. Identity crime and misuse in Australia: Results of the 2019 online survey. (2020).

[99] Christie Franks and Russell G Smith. 2021. *Changing perceptions of biometric technologies*. Australian Institute of Criminology.

[100] Sara Freitas and Fotis Liarokapis. 2011. *Serious Games: A New Paradigm for Education?* 9–23. DOI:http://dx.doi.org/10.1007/978-1-4471-2161-9_2

[101] Fujitsu. 2018. Fujitsu Begins Large-Scale Internal Deployment of Palm Vein Authentication to Accelerate Workstyle Transformation. *Fujitsu Press Release* (2018). http://www.fujitsu.com/global/about/resources/news/press-releases/2018/0118-01.html.

[102] Tracy Fullerton. 2014. *Game design workshop: a playcentric approach to creating innovative games*. CRC press.

[103] Steven M Furnell and Konstantinos Evangelatos. 2007. Public awareness and perceptions of biometrics. *Computer Fraud & Security* 2007, 1 (2007), 8–13.

[104] Davrondzhon Gafurov. 2007. A survey of biometric gait recognition: Approaches, security and challenges. In *Annual Norwegian computer science conference*. Annual Norwegian Computer Science Conference Norway, 19–21.

[105] Davrondzhon Gafurov, Einar Snekkenes, and Tor Erik Buvarp. 2006. Robustness of biometric gait authentication against impersonation attack. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*. Springer, 479–488.

[106] Mikel Galar, Joaquín Derrac, Daniel Peralta, Isaac Triguero, Daniel Paternain, Carlos Lopez-Molina, Salvador García, José M Benítez, Miguel Pagola, Edurne Barrenechea, and others. 2015. A survey of fingerprint classification Part I: Taxonomies on feature extraction methods and learning models. *Knowledge-based systems* 81 (2015), 76–97.

[107] Simson Garfinkel and Heather Richter Lipford. 2014. *Usable security: History, themes, and challenges*. Morgan & Claypool Publishers.

[108] Nina Gerber, Paul Gerber, and Melanie Volkamer. 2018. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & security* 77 (2018), 226–261.

[109] Romain Giot and Christophe Rosenberger. 2012. A new soft biometric approach for keystroke dynamics based on gender recognition. *International Journal of Information Technology and Management* 11, 1-2 (2012), 35–49.

[110] Cristiano Giuffrida, Kamil Majdanik, Mauro Conti, and Herbert Bos. 2014. I sensed it was you: authenticating mobile users with sensor-enhanced keystroke dynamics. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 92–111.

[111] Mayank Goel, Leah Findlater, and Jacob Wobbrock. 2012. WalkType: Using Accelerometer Data to Accomodate Situational Impairments in Mobile Touch Screen Text Entry. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 2687–2696. DOI:http://dx.doi.org/10.1145/2207676.2208662

[112] Mayank Goel, Alex Jansen, Travis Mandel, Shwetak N. Patel, and Jacob O. Wobbrock. 2013. ContextType: Using Hand Posture Information to Improve Mobile Touch Screen Text Entry. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2795–2798. DOI: http://dx.doi.org/10.1145/2470654.2481386

[113] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014a. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[114] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014b. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[115] Joshua Goodman, Gina Venolia, Keith Steury, and Chauncey Parker. 2002. Language Modeling for Soft Keyboards. In *Proceedings of the 7th International Conference on Intelligent User Interfaces (IUI '02)*. ACM, New York, NY, USA, 194–195. DOI: http://dx.doi.org/10.1145/502716.502753

[116] Patrick J Grother, Patrick J Grother, Mei Ngan, and K Hanaoka. 2014. *Face recognition vendor test (FRVT).* US Department of Commerce, National Institute of Standards and Technology.

[117] David Gueorguiev, Anis Kaci, Michel Amberg, Frédéric Giraud, and Betty Lemaire-Semail. 2018. Travelling ultrasonic wave enhances keyclick sensation. In *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*. Springer, 302–312.

[118] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[119] Asela Gunawardana, Tim Paek, and Christopher Meek. 2010. Usability Guided Key-target Resizing for Soft Keyboards. In *Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI '10)*. ACM, New York, NY, USA, 111–118. DOI:http://dx.doi.org/10.1145/1719970.1719986

[120] Marian Harbach, Alexander De Luca, and Serge Egelman. 2016. The Anatomy of Smartphone Unlocking: A Field Study of Android Lock Screens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4806–4817. DOI:http://dx.doi.org/10.1145/2858036.2858267

[121] Marian Harbach, Emanuel Von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. 2014. It's a hard lock life: A field study of smartphone (un) locking behavior and risk perception. In *Symposium on usable privacy and security (SOUPS)*. 213–230.

[122] Gunnar Harboe and Elaine M. Huang. 2015. Real-World Affinity Diagramming Practices: Bridging the Paper-Digital Gap. In *Proc. 33rd Annual ACM Conf. Human Factors in Computing Systems*. ACM, New York, NY, USA, 95–104. DOI:http://dx.doi.org/10.1145/2702123.2702561

[123] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GANspace: Discovering Interpretable GAN Controls. *Advances in neural information processing systems* 33 (2020), 9841–9850.

[124] Chris Harrison and Scott E. Hudson. 2009. Texture Displays: A Passive Approach to Tactile Presentation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. Association for Computing Machinery, New York, NY, USA, 2261–2264. DOI:http://dx.doi.org/10.1145/1518701.1519047

[125] Eiji Hayashi, Sauvik Das, Shahriyar Amini, Jason Hong, and Ian Oakley. 2013. CASA: Context-aware Scalable Authentication. In *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS '13)*. ACM, New York, NY, USA, Article 3, 10 pages. DOI:http://dx.doi.org/10.1145/2501604.2501607

[126] Eiji Hayashi and Jason Hong. 2011. A Diary Study of Password Usage in Daily Life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2627–2630. DOI:http://dx.doi.org/10.1145/1978942.1979326

[127] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15262–15271.

[128] Niels Henze, Enrico Rukzio, and Susanne Boll. 2012. Observational and experimental investigation of typing behaviour using virtual keyboards for mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2659–2668.

[129] Andreas Hinderks, Martin Schrepp, and Jörg Thomaschewski. 2018. A Benchmark for the Short Version of the User Experience Questionnaire.. In *WEBIST*. 373–377.

[130] Jean Hitchings. 1995. Deficiencies of the traditional approach to information security and the requirements for a new methodology. *Computers & Security* 14, 5 (1995), 377–383.

[131] Alexander Hoffmann, Daniel Spelmezan, and Jan Borchers. 2009. TypeRight: a keyboard with tactile error prevention. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2265–2268.

[132] Lin Hong, Anil K Jain, and Sharath Pankanti. 1999. Can multibiometrics improve performance?. In *Proceedings AutoID*, Vol. 99. Citeseer, 59–64.

[133] Seongseob Hwang, Sungzoon Cho, and Sunghoon Park. 2009a. Keystroke dynamics-based authentication for mobile devices. *Computers & Security* 28, 1–2 (2009), 85–93. DOI:http://dx.doi.org/10.1016/j.cose.2008.10.002

[134] Seongseob Hwang, Hyoung joo Lee, and Sungzoon Cho. 2009b. Improving authentication accuracy using artificial rhythms and cues for keystroke dynamics-based authentication. *Expert Systems with Applications* 36, 7 (2009), 10649 – 10656. DOI:http://dx.doi.org/https://doi.org/10.1016/j.eswa.2009.02.075

[135] Ada Lovelace Institute. 2019. Beyond Face Value: Public Attitudes to Facial Recognition Technology. (2019).

[136] ISO/IEC 2382-37 2012. *Biometric Vocabulary*. Technical Report. International Organization for Standardization(ISO).

[137] Ali Jahanian, Lucy Chai, and Phillip Isola. 2019. On the”steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171* (2019).

[138] Anil K Jain, Sunpreet S Arora, Kai Cao, Lacey Best-Rowden, and Anjoo Bhatnagar. 2016. Fingerprint recognition of young children. *IEEE Transactions on Information Forensics and Security* 12, 7 (2016), 1501–1514.

[139] Anil K Jain, Ruud Bolle, and Sharath Pankanti. 1996. *Introduction to biometrics*. Springer.

[140] Anil K Jain, Debayan Deb, and Joshua J Engelsma. 2021. Biometrics: Trust, but verify. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4, 3 (2021), 303–323.

[141] Anil K Jain, Lin Hong, Sharath Pankanti, and Ruud Bolle. 1997. An identity-authentication system using fingerprints. *Proc. IEEE* 85, 9 (1997), 1365–1388.

[142] Anil K Jain, S Prabhakar, and Lin Hong. 1999. A multichannel approach to fingerprint classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21, 4 (1999), 348–359. DOI:http://dx.doi.org/10.1109/34.761265

[143] Anil K Jain and Arun Ross. 2004. Multibiometric systems. *Commun. ACM* 47, 1 (2004), 34–40.

[144] Anil K Jain, Arun Ross, and Salil Prabhakar. 2004. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology* 14, 1 (2004), 4–20.

[145] Nayna Jain, Karthik Nandakumar, Nalini Ratha, Sharath Pankanti, and Uttam Kumar. 2021. Efficient CNN building blocks for encrypted data. *arXiv preprint arXiv:2102.00319* (2021).

[146] Markus Jakobsson, Elaine Shi, Philippe Golle, and Richard Chow. 2009. Implicit Authentication for Mobile Devices. In *Proceedings of the 4th USENIX Conference on Hot Topics in Security (HotSec’09)*. USENIX Association, Berkeley, CA, USA, 9–9. http://dl.acm.org/citation.cfm?id=1855628.1855637

[147] Andre Kalamandeen, Adin Scannell, Eyal de Lara, Anmol Sheth, and Anthony LaMarca. 2010. Ensemble: Cooperative Proximity-based Authentication. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services (MobiSys ’10)*. ACM, New York, NY, USA, 331–344. DOI:http://dx.doi.org/10.1145/1814433.1814466

[148] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. 2023. Scaling up GANs for Text-to-Image Synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[149] Sevasti Karatzouni and Nathan L Clarke. 2007. Keystroke Analysis for Thumb-based Keyboards on Mobile Devices. In *New Approaches for Security, Privacy and Trust in Complex Environments*, Hein Venter, Mariki Eloff, Les Labuschagne, Jan Eloff, and Rossouw von Solms (Eds.). Springer US, Boston, MA, 253–263.

[150] Sevasti Karatzouni, Steven M Furnell, Nathan L Clarke, and Reinhardt A Botha. 2007. Perceptions of user authentication on mobile devices. In *Proceedings of the ISOneWorld Conference*. Citeseer, 11–13.

[151] Amy K Karlson, AJ Brush, and Stuart Schechter. 2009. Can i borrow your phone?: understanding concerns when sharing mobile phones. In *Proc. of CHI'09*. ACM, 1647–1650.

[152] Marcus Karnan, Muthuramalingam Akila, and Nishara Krishnaraj. 2011. Biometric personal authentication using keystroke dynamics: A review. *Applied soft computing* 11, 2 (2011), 1565–1573.

[153] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[154] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[155] Mohamed Khamis, Florian Alt, Mariam Hassib, Emanuel von Zezschwitz, Regina Hasholzner, and Andreas Bulling. 2016. GazeTouchPass: Multimodal Authentication Using Gaze and Touch on Mobile Devices. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 2156–2164. DOI:http://dx.doi.org/10.1145/2851581.2892314

[156] Hassan Khan, Aaron Atwater, and Urs Hengartner. 2014. Itus: an implicit authentication framework for android. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 507–518.

[157] Hassan Khan, Urs Hengartner, and Daniel Vogel. 2015. Usability and security perceptions of implicit authentication: convenient, secure, sometimes annoying. In *Eleventh Symposium on Usable Privacy and Security (SOUPS 2015)*. 225–239.

[158] Hassan Khan, Urs Hengartner, and Daniel Vogel. 2016. Targeted mimicry attacks on touch input based implicit authentication schemes. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 387–398.

[159] Hassan Khan, Urs Hengartner, and Daniel Vogel. 2018. Augmented Reality-based Mimicry Attacks on Behaviour-Based Smartphone Authentication. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 41–53.

[160] Hassan Khan, Urs Hengartner, and Daniel Vogel. 2020. Mimicry Attacks on Smartphone Keystroke Authentication. *ACM Transactions on Privacy and Security* 23 (02 2020), 1–34. DOI:http://dx.doi.org/10.1145/3372420

[161] Durk P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems* 31 (2018).

[162] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[163] René F. Kizilcec. 2016. How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2390–2395. DOI:http://dx.doi.org/10.1145/2858036.2858402

[164] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on information forensics and security* 7, 6 (2012), 1789–1801.

[165] Saranga Komanduri, Richard Shay, Lorrie Faith Cranor, Cormac Herley, and Stuart E Schechter. 2014. Telepathwords: Preventing Weak Passwords by Reading Users' Minds. In *USENIX Security Symposium*. 591–606.

[166] Adams Kong, David Zhang, and Mohamed Kamel. 2009. A survey of palmprint recognition. *pattern recognition* 42, 7 (2009), 1408–1418.

[167] Joey Lee and Jessica Hammer. 2011. Gamification in Education: What, How, Why Bother? *Academic Exchange Quarterly* 15 (01 2011), 1–5.

[168] Lingjun Li, Xinxin Zhao, and Guoliang Xue. 2013. Unobservable re-authentication for smartphones.. In *NDSS*, Vol. 56. 57–59.

[169] Yuk L Li and Padmaja Ramadas. 2012. Context aware biometric authentication. (Aug. 28 2012). US Patent 8,255,698.

[170] Zhaoyuan Ma, Darren Edge, Leah Findlater, and Hong Z Tan. 2015. Haptic keyclick feedback improves typing speed and reduces typing errors on a flat keyboard. In *2015 IEEE World Haptics Conference (WHC)*. IEEE, 220–227.

[171] Ahmed Mahfouz, Ildar Muslukhov, and Konstantin Beznosov. 2016. Android users in the wild: Their authentication and usage behavior. *Pervasive and Mobile Computing* 32 (2016), 50 – 61. DOI:http://dx.doi.org/https://doi.org/10.1016/j.pmcj.2016.06.017 Mobile Security, Privacy and Forensics.

[172] Guangcan Mai, Kai Cao, Pong C Yuen, and Anil K Jain. 2018. On the reconstruction of face images from deep face templates. *IEEE transactions on pattern analysis and machine intelligence* 41, 5 (2018), 1188–1202.

[173] Emanuele Maiorana, Patrizio Campisi, Noelia González-Carballo, and Alessandro Neri. 2011. Keystroke Dynamics Authentication for Mobile Phones. In *Proceedings of the 2011 ACM Symposium on Applied Computing (SAC '11)*. ACM, New York, NY, USA, 21–26. DOI:http://dx.doi.org/10.1145/1982185.1982190

[174] Davide Maltoni, Dario Maio, Anil K Jain, Salil Prabhakar, and others. 2009. *Handbook of fingerprint recognition*. Vol. 2. Springer.

[175] Emanuela Marasco and Arun Ross. 2014. A survey on antispoofing schemes for fingerprint recognition systems. *ACM Computing Surveys (CSUR)* 47, 2 (2014), 1–36.

[176] Sven Mayer, Huy Viet Le, and Niels Henze. 2017. Estimating the Finger Orientation on Capacitive Touchscreens Using Convolutional Neural Networks. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces (ISS '17)*. ACM, New York, NY, USA, 220–229. DOI:http://dx.doi.org/10.1145/3132272.3134130

[177] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 72 (nov 2019), 23 pages. DOI:http://dx.doi.org/10.1145/3359174

[178] Daniel C McFarlane. 2002. Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human-Computer Interaction* 17, 1 (2002), 63–139.

[179] Jane McGonigal. 2011. *Reality is Broken: Why Games Make Us Better and How They Can Change the World*. Vol. 22.

[180] Lukas Mecke, Daniel Buschek, Uwe Gruenefeld, and Florian Alt. 2024. Exploring the Lands Between: A Method for Finding Differences between AI-Decisions and Human Ratings through Generated Samples. *arXiv preprint arXiv:2409.12801* (2024).

[181] Lukas Mecke, Daniel Buschek, Mathias Kiermeier, Sarah Prange, and Florian Alt. 2019a. Exploring intentional behaviour modifications for password typing on mobile touchscreen devices. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. 303–317.

[182] Lukas Mecke, Sarah Delgado Rodriguez, Daniel Buschek, Sarah Prange, and Florian Alt. 2019b. Communicating Device Confidence Level and Upcoming Re-Authentications in Continuous Authentication Systems on Mobile Devices. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. 289–301.

[183] Lukas Mecke, Ken Pfeuffer, Sarah Prange, and Florian Alt. 2018a. Open sesame! user perception of physical, biometric, and behavioural authentication concepts to open doors. In *Proceedings of the 17th international conference on mobile and ubiquitous multimedia*. 153–159.

[184] Lukas Mecke, Sarah Prange, Daniel Buschek, and Florian Alt. 2018b. A Design Space for Security Indicators for Behavioural Biometrics on Mobile Touchscreen Devices. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, LBW003.

[185] Lukas Mecke, Ismael Prieto Romero, Sarah Delgado Rodriguez, and Florian Alt. 2023. Exploring the Use of Electromagnets to Influence Key Targeting on Physical Keyboards. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–8.

[186] Lukas Mecke, Alia Saad, Sarah Prange, Uwe Gruenefeld, Stefan Schneegass, and Florian Alt. 2024. Do They Understand What They Are Using? – Assessing Perception and Usage of Biometrics. *arXiv preprint arXiv:2410.12661* (2024).

[187] William Melicher, Darya Kurilova, Sean M Segreti, Pranshu Kalvani, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L Mazurek. 2016. Usability and security of text passwords on mobile devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 527–539.

[188] Nicholas Micallef, Mike Just, Lynne Baillie, Martin Halvey, and Hilmi Güneş Kayacik. 2015. Why aren't users using protection? investigating the usability of smartphone locking. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 284–294.

[189] Laurent Mignonneau and Christa Sommerer. 2005. Nano-Scape: experiencing aspects of nanotechnology through a magnetic force-feedback interface. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*. 200–203.

[190] B Miller. 1988. Everything you need to know about biometric identification. *Personal Identification News 1988 Biometric Industry Directory* (1988).

[191] Viktor Miruchna, Robert Walter, David Lindlbauer, Maren Lehmann, Regine Von Klitzing, and Jörg Müller. 2015. Geltouch: Localized tactile feedback through thin, programmable gel. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 3–10.

[192] Kenrick Mock, Bogdan Hoanca, Justin Weaver, and Mikal Milton. 2012. Real-time continuous iris recognition for authentication using an eye tracker. In *Proceedings of the 2012 ACM conference on Computer and communications security*. ACM, 1007–1009.

[193] John V Monaco, Md Liakat Ali, and Charles C Tappert. 2015. Spoofing key-press latencies with a generative keystroke dynamics model. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 1–8.

[194] Jocelyn Monnoyer, Emmanuelle Diaz, Christophe Bourdin, and Michaël Wiertlewski. 2016. Ultrasonic friction modulation while pressing induces a tactile feedback. In *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*. Springer, 171–179.

[195] Fabian Monrose, Michael K. Reiter, and Susanne Wetzel. 1999. Password Hardening Based on Keystroke Dynamics. In *Proceedings of the 6th ACM Conference on Computer and Communications Security (CCS '99)*. ACM, New York, NY, USA, 73–82. DOI:http://dx.doi.org/10.1145/319709.319720

[196] Fabian Monrose and Aviel Rubin. 1997. Authentication via Keystroke Dynamics. In *Proceedings of the 4th ACM Conference on Computer and Communications Security (CCS '97)*. ACM, New York, NY, USA, 48–56. DOI:http://dx.doi.org/10.1145/266420.266434

[197] George Musumba and Henry Nyongesa. 2013. Context awareness in mobile computing: A review. *International Journal of Machine Learning and Applications* 2, 1 (2013), 5. DOI:http://dx.doi.org/10.4102/ijmla.v2i1.5

[198] Jakob Nielsen. 1994. Enhancing the Explanatory Power of Usability Heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. Association for Computing Machinery, New York, NY, USA, 152–158. DOI:http://dx.doi.org/10.1145/191666.191729

[199] Lawrence O'Gorman. 2003. Comparing passwords, tokens, and biometrics for user authentication. *Proc. IEEE* 91, 12 (Dec 2003), 2021–2040. DOI:http://dx.doi.org/10.1109/JPROC.2003.819611

[200] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature visualization. *Distill* 2, 11 (2017), e7.

[201] Gian Pangaro, Dan Maynes-Aminzade, and Hiroshi Ishii. 2002. The actuated workbench: computer-controlled actuation in tabletop tangible interfaces. In *Proceedings of the 15th annual ACM symposium on User interface software and technology*. 181–190.

[202] Keunwoo Park, Daehwa Kim, Seongkook Heo, and Geehyuk Lee. 2020. MagTouch: Robust Finger Identification for a Smartwatch Using a Magnet Ring and a Built-in Magnetometer. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[203] Ken Pfeuffer, Matthias J Geiger, Sarah Prange, Lukas Mecke, Daniel Buschek, and Florian Alt. 2019. Behavioural biometrics in VR: Identifying people from body motion

and relations in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[204] P Jonathon Phillips, Patrick J Flynn, and Kevin W Bowyer. 2017. Lessons from collecting a million biometric samples. *Image and Vision Computing* 58 (2017), 96–107.

[205] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. 2020. Adversarial Latent Autoencoders. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 14104–14113.

[206] David MW Powers. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (2020).

[207] Sarah Prange, Daniel Buschek, and Florian Alt. 2018. An Exploratory Study on Correlations of Hand Size and Mobile Touch Interactions. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia (MUM 2018)*. ACM, New York, NY, USA, 279–283. DOI:http://dx.doi.org/10.1145/3282894.3282924

[208] Sarah Prange, Lukas Mecke, Alice Nguyen, Mohamed Khamis, and Florian Alt. 2020. Don't Use Fingerprint, it's Raining! How People Use and Perceive Context-Aware Selection of Mobile Authentication. In *Proceedings of the international conference on advanced visual interfaces*. 1–5.

[209] Pearl Pu and Li Chen. 2006. Trust Building with Explanation Interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces (IUI '06)*. ACM, New York, NY, USA, 93–100. DOI:http://dx.doi.org/10.1145/1111449.1111475

[210] Maja Pusara and Carla E. Brodley. 2004. User Re-Authentication via Mouse Movements. In *Proceedings of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security (VizSEC/DMSEC '04)*. Association for Computing Machinery, New York, NY, USA, 1–8. DOI:http://dx.doi.org/10.1145/1029208.1029210

[211] Khandaker A Rahman, Kiran S Balagani, and Vir V Phoha. 2013. Snoop-forge-replay attacks on continuous verification with keystrokes. *IEEE Transactions on Information Forensics and Security* 8, 3 (2013), 528–541.

[212] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[213] Sanka Rasnayaka and Terence Sim. 2018. Who wants continuous authentication on mobile devices?. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 1–9.

[214] Nataasha Raul, Radha Shankarmani, and Padmaja Joshi. 2020. A comprehensive review of keystroke dynamics-based authentication mechanism. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2019, Volume 2*. Springer, 149–162.

[215] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558* (2020).

[216] Danilo Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *International conference on machine learning*. PMLR, 1530–1538.

[217] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[218] Oriana Riva, Chuan Qin, Karin Strauss, and Dimitrios Lymberopoulos. 2012. Progressive Authentication: Deciding when to Authenticate on Mobile Phones. In *Proceedings of the 21st USENIX Conference on Security Symposium (Security'12)*. USENIX Association, Berkeley, CA, USA, 15–15. `http://dl.acm.org/citation.cfm?id=2362793.2362808`

[219] Chris Roberts. 2007. Biometric attack vectors and defences. *Computers & Security* 26, 1 (2007), 14 – 25.

[220] Paul Rodrigues and C Anton Rytting. 2012. Typing Race Games as a Method to Create Spelling Error Corpora.. In *LREC*. 3019–3024.

[221] Joseph Romanowski, Kirsanov Charles, Patricia Jasso, Shreyansh Shah, and Hugh W Eng. 2016. A Biometric Security Acceptability and Ease-of-Use Study on a Palm Vein Scanner. *Proceedings of Student-Faculty Research Day, CSIS, Pace University* (2016).

[222] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[223] Arun Ross, Sudipta Banerjee, and Anurag Chowdhury. 2022. Deducing health cues from biometric data. *Computer Vision and Image Understanding* 221 (2022), 103438. DOI:`http://dx.doi.org/https://doi.org/10.1016/j.cviu.2022.103438`

[224] Manuel Rudolph and Reinhard Schwarz. 2012. A critical survey of security indicator approaches. In *Proc. of ARES'12*. IEEE, 291–300.

[225] Alia Saad, Uwe Gruenefeld, Lukas Mecke, Marion Koelle, Florian Alt, and Stefan Schneegass. 2022. Mask removal isn't always convenient in public!–The Impact of the Covid-19 Pandemic on Device Usage and User Authentication. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[226] Hataichanok Saevanee, Nathan L Clarke, and Steven M Furnell. 2012. Multi-modal behavioural biometric authentication for mobile devices. In *IFIP International Information Security Conference*. Springer, 465–474.

[227] Jerome H Saltzer and Michael D Schroeder. 1975. The protection of information in computer systems. *Proc. IEEE* 63, 9 (1975), 1278–1308.

[228] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *CoRR* abs/1708.08296 (2017). http://arxiv.org/abs/1708.08296

[229] S Sanderson and JH Erbetta. 2000. Authentication for secure environments based on iris scanning technology. (2000).

[230] Martina Angela Sasse, Sacha Brostoff, and Dirk Weirich. 2001. Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security. *BT technology journal* 19, 3 (2001), 122–131.

[231] Martina Angela Sasse and Ivan Flechais. 2005. Usable security: Why do we need it? How do we get it? O'Reilly.

[232] Grégory Savioz, Miroslav Markovic, and Yves Perriard. 2011. Towards multi-finger haptic devices: A computer keyboard with adjustable force feedback. In *2011 International Conference on Electrical Machines and Systems*. IEEE, 1–6.

[233] Grégory Savioz and Yves Perriard. 2009. A miniature short stroke linear actuator and its position control for a haptic key. In *2009 IEEE Energy Conversion Congress and Exposition*. IEEE, 2441–2446.

[234] Florian Schaub, Ruben Deyhle, and Michael Weber. 2012. Password Entry Usability and Shoulder Surfing Susceptibility on Different Smartphone Platforms. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia (MUM '12)*. ACM, New York, NY, USA, Article 13, 10 pages. DOI:http://dx.doi.org/10.1145/2406367.2406384

[235] B. Schilit, N. Adams, and R. Want. 1994. Context-Aware Computing Applications. In *1994 First Workshop on Mobile Computing Systems and Applications*. 85–90. DOI:http://dx.doi.org/10.1109/WMCSA.1994.16

[236] Albrecht Schmidt, Michael Beigl, and Hans-W Gellersen. 1999. There is more to context than location. *Computers & Graphics* 23, 6 (1999), 893 – 901. DOI:http://dx.doi.org/https://doi.org/10.1016/S0097-8493(99)00120-X

[237] Stefan Schneegass, Frank Steimle, Andreas Bulling, Florian Alt, and Albrecht Schmidt. 2014. SmudgeSafe: Geometric Image Transformations for Smudge-Resistant User Authentication. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 775–786. DOI:http://dx.doi.org/10.1145/2632048.2636090

[238] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108.* (2017).

[239] Tobias Seitz and Heinrich Hussmann. 2017. PASDJO: quantifying password strength perceptions with an online game. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*. 117–125.

[240] Tobias Seitz, Emanuel von Zezschwitz, Stefanie Meitner, and Heinrich Hussmann. 2016. Influencing self-selected passwords through suggestions and the decoy effect. In *Proceedings of the 1st European Workshop on Usable Security*. 1–2.

[241] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[242] Alireza Sepas-Moghaddam and Ali Etemad. 2022. Deep gait recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence* 45, 1 (2022), 264–284.

[243] Sefik Ilkin Serengil and Alper Ozpinar. 2020. LightFace: A Hybrid Deep Face Recognition Framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 23–27. DOI:http://dx.doi.org/10.1109/ASYU50717.2020.9259802

[244] Abdul Serwadda and Vir V Phoha. 2013. Examining a large keystroke biometrics dataset for statistical-attack openings. *ACM Transactions on Information and System Security (TISSEC)* 16, 2 (2013), 8.

[245] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. 2020. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*. 1589–1604.

[246] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the Latent Space of GANs for Semantic Face Editing. In *CVPR*.

[247] Elaine Shi, Yuan Niu, Markus Jakobsson, and Richard Chow. 2010. Implicit authentication through learning user behavior. In *International Conference on Information Security*. Springer, 99–113.

[248] Yichun Shi and Anil K Jain. 2019. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6902–6911.

[249] Kye Shimizu, Naoto Ienaga, Kazuma Takada, Maki Sugimoto, and Shunichi Kasahara. 2022. Human Latent Metrics: Perceptual and Cognitive Response Correlates to Distance in GAN Latent Space for Facial Images. In *ACM Symposium on Applied Perception 2022*. 1–10.

[250] Hanul Sieger, Niklas Kirschnick, and Sebastian Möller. 2010. Poster: User preferences for biometric authentication methods and graded security on mobile phones. In *Symposium on usability, privacy, and security (SOUPS)*. Citeseer.

[251] Jasvinder Pal Singh, Sanjeev Jain, Sakshi Arora, and Uday Pratap Singh. 2018. Vision-based gait recognition: A survey. *Ieee Access* 6 (2018), 70497–70527.

[252] Dawn Xiaodong Song, David Wagner, and Xuqing Tian. 2001. Timing Analysis of Keystrokes and Timing Attacks on {SSH}. In *10th USENIX Security Symposium (USENIX Security 01)*.

[253] Zhexuan Song and Jesus Molina. 2011. Method and apparatus for context-aware authentication. (Dec. 22 2011). US Patent App. 12/816,966.

[254] Sebastijan Sprager and Matjaz B Juric. 2015. Inertial sensor-based gait recognition: A review. *Sensors* 15, 9 (2015), 22089–22127.

[255] Andrew Stapleton. 2004. Serious games: Serious opportunities. *Australian Game Developers Conference* (01 2004).

[256] Deian Stefan, Xiaokui Shu, and Danfeng (Daphne) Yao. 2012. Robustness of Keystroke-dynamics Based Biometrics Against Synthetic Forgeries. *Computers & Security* 31, 1 (Feb. 2012), 109–121. DOI:http://dx.doi.org/10.1016/j.cose.2011.10.001

[257] Abby Stylianou, Richard Souvenir, and Robert Pless. 2019. Visualizing deep similarity networks. In *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2029–2037.

[258] Ioannis Stylios, Spyros Kokolakis, Andreas Skalkos, and Sotirios Chatzis. 2022. BioGames: a new paradigm and a behavioral biometrics collection tool for research purposes. *Information & Computer Security* 30, 2 (2022), 243–254.

[259] Kalaivani Sundararajan and Damon L Woodard. 2018. Deep learning for biometrics: A survey. *ACM Computing Surveys (CSUR)* 51, 3 (2018), 1–34.

[260] Ryuichi Tachibana and Mamoru Komachi. 2016. Analysis of english spelling errors in a word-typing game. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 385–390.

[261] Murat Taskiran, Nihan Kahraman, and Cigdem Eroglu Erdem. 2020. Face recognition: Past, present and future (a review). *Digital Signal Processing* 106 (2020), 102809.

[262] Pin Shen Teh, Andrew Beng Jin Teoh, and Shigang Yue. 2013. A Survey of Keystroke Dynamics Biometrics. *The Scientific World Journal* 2013 (2013). DOI:http://dx.doi.org/10.1155/2013/408280

[263] Pin Shen Teh, Ning Zhang, Andrew Beng Jin Teoh, and Ke Chen. 2016. A Survey on Touch Dynamics Authentication in Mobile Devices. *Computers & Security* 59, C (2016), 210–235. DOI:http://dx.doi.org/10.1016/j.cose.2016.03.003

[264] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. 2021. A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society* 3, 1 (2021), 16–30.

[265] Chee Meng Tey, Payas Gupta, and Debin Gao. 2013. I can be you: Questioning the use of keystroke dynamics as biometrics. In *Annual Network and Distributed System Security Symposium 20th NDSS*. Research Collection School Of Information Systems, 1–6.

[266] Christian Tiefenau, Maximilian Häring, Eva Gerlitz, and Emanuel von Zezschwitz. 2019a. Making Privacy Graspable: Can we Nudge Users to use Privacy Enhancing Techniques? *arXiv preprint arXiv:1911.07701* (2019).

[267] Christian Tiefenau, Maximilian Häring, Mohamed Khamis, and Emanuel von Zezschwitz. 2019b. " Please enter your PIN"–On the Risk of Bypass Attacks on Biometric Authentication on Mobile Devices. *arXiv preprint arXiv:1911.07692* (2019).

[268] Patrick Tinsley, Adam Czajka, and Patrick Flynn. 2021. This face does not exist... but it might be yours! identity leakage in generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1320–1328.

[269] AS Tolba, AH El-Baz, and AA El-Harby. 2006. Face recognition: A literature review. *International Journal of Signal Processing* 2, 2 (2006), 88–103.

[270] Blase Ur, Felicia Alfieri, Maung Aung, Lujo Bauer, Nicolas Christin, Jessica Colnago, Lorrie Faith Cranor, Henry Dixon, Pardis Emami Naeini, Hana Habib, and others. 2017. Design and evaluation of a data-driven password meter. In *Proc. of CHI 2017*. ACM, 3775–3786.

[271] Blase Ur, Jonathan Bees, Sean M Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2016. Do Users' Perceptions of Password Security Match Reality?. In *Proc. of CHI'16*. ACM, 3748–3760.

[272] Blase Ur, Patrick Gage Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L Mazurek, Timothy Passaro, Richard Shay, Timothy Vidas, Lujo Bauer, and others. 2012. How does your password measure up? The effect of strength meters on password creation. In *USENIX Security Symposium*. 65–80.

[273] Arash Vahdat and Jan Kautz. 2020. NVAE: A deep hierarchical variational autoencoder. *Advances in neural information processing systems* 33 (2020), 19667–19679.

[274] Radu-Daniel Vatavu, Lisa Anthony, and Quincy Brown. 2015. Child or Adult? Inferring Smartphone Users' Age Group from Touch Measurements Alone. In *Human-Computer Interaction – INTERACT 2015*, Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler (Eds.). Springer International Publishing, Cham, 1–9.

[275] Esteban Vazquez-Fernandez and Daniel Gonzalez-Jimenez. 2016. Face recognition for authentication on mobile devices. *Image and Vision Computing* 55 (2016), 31–33. DOI:http://dx.doi.org/https://doi.org/10.1016/j.imavis.2016.03.018

[276] Ruben Vera-Rodriguez, John S.D. Mason, Julian Fierrez, and Javier Ortega-Garcia. 2013. Comparative Analysis and Fusion of Spatiotemporal Information for Footstep Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 4 (2013), 823–834. DOI:http://dx.doi.org/10.1109/TPAMI.2012.164

[277] Luca Vigano and Daniele Magazzeni. 2020. Explainable security. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 293–300.

[278] Emanuel Von Zezschwitz, Alexander De Luca, Bruno Brunkow, and Heinrich Hussmann. 2015. Swipin: Fast and secure pin-entry on smartphones. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*. 1403–1406.

[279] Emanuel von Zezschwitz, Alexander De Luca, Philipp Janssen, and Heinrich Hussmann. 2015. Easy to Draw, but Hard to Trace?: On the Observability of Grid-based (Un)Lock Patterns. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2339–2342. DOI:http://dx.doi.org/10.1145/2702123.2702202

[280] Emanuel von Zezschwitz, Malin Eiband, Daniel Buschek, Sascha Oberhuber, Alexander De Luca, Florian Alt, and Heinrich Hussmann. 2016. On Quantifying the Effective Password Space of Grid-based Unlock Gestures. In *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia (MUM '16)*. ACM, New York, NY, USA, 201–212. DOI:http://dx.doi.org/10.1145/3012709.3012729

[281] Emanuel von Zezschwitz, Anton Koslow, Alexander De Luca, and Heinrich Hussmann. 2013. Making Graphic-based Authentication Secure Against Smudge Attacks.

In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13)*. ACM, New York, NY, USA, 277–286. DOI:http://dx.doi.org/10.1145/2449396.2449432

[282] Changsheng Wan, Li Wang, and Vir V Phoha. 2018. A survey on gait recognition. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–35.

[283] Rui Wang, Jian Chen, Gang Yu, Li Sun, Changqian Yu, Changxin Gao, and Nong Sang. 2021. Attribute-specific control units in stylegan for fine-grained image manipulation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 926–934.

[284] Rick Wash. 2010. Folk Models of Home Computer Security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security (SOUPS '10)*. Association for Computing Machinery, New York, NY, USA, Article 11, 16 pages. DOI:http://dx.doi.org/10.1145/1837110.1837125

[285] J Wayman. 2000. A definition of biometrics. *National Biometric Test Center Collected Works* 1, 2 (2000), 21–23.

[286] Malte Weiss, Chat Wacharamanotham, Simon Voelker, and Jan Borchers. 2011. FingerFlux: Near-Surface Haptic Feedback on Tabletops. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. Association for Computing Machinery, New York, NY, USA, 615–620. DOI:http://dx.doi.org/10.1145/2047196.2047277

[287] John Williamson and Roderick Murray-Smith. 2012. Rewarding the Original: Explorations in Joint User-sensor Motion Spaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1717–1726. DOI:http://dx.doi.org/10.1145/2207676.2208301

[288] Jacob O. Wobbrock. 2010. Measures of Text Entry Performance. In *Text Entry Systems: Mobility, Accessibility, Universality*. Morgan Kaufmann, Chapter 3, 47 – 74.

[289] Jacob O Wobbrock and Julie A Kientz. 2016. Research contributions in human-computer interaction. *interactions* 23, 3 (2016), 38–44.

[290] Adam Wójtowicz and Krzysztof Joachimiak. 2016. Model for Adaptable Context-based Biometric Authentication for Mobile Devices. *Personal Ubiquitous Comput.* 20, 2 (April 2016), 195–207. DOI:http://dx.doi.org/10.1007/s00779-016-0905-0

[291] Min Wu, Robert C Miller, and Simson L Garfinkel. 2006. Do security toolbars actually prevent phishing attacks?. In *Proc. of CHI'06*. ACM, 601–610.

[292] Hui Xu, Yangfan Zhou, and Michael R. Lyu. 2014. Towards Continuous and Passive Authentication via Touch Biometrics: An Experimental Study on Smartphones. In *Symposium On Usable Privacy and Security (SOUPS 2014)*. USENIX Association, Menlo Park, CA, 187–198. https://www.usenix.org/conference/soups2014/proceedings/presentation/xu

[293] Weitao Xu, Yiran Shen, Yongtuo Zhang, Neil Bergmann, and Wen Hu. 2017. Gait-watch: A context-aware authentication system for smart watch based on gait recognition. In *Proceedings of the Second International Conference on Internet-of-Things Design and Implementation*. ACM, 59–70.

[294] Neil Yager and Ted Dunstone. 2008. The biometric menagerie. *IEEE transactions on pattern analysis and machine intelligence* 32, 2 (2008), 220–230.

[295] Junichi Yamaoka and Yasuaki Kakehi. 2013. DePENd: Augmented Handwriting System Using Ferromagnetism of a Ballpoint Pen. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13)*. Association for Computing Machinery, New York, NY, USA, 203–210. DOI:http://dx.doi.org/10.1145/2501988.2502017

[296] Roman V Yampolskiy and Venu Govindaraju. 2008. Behavioural biometrics: a survey and classification. *International Journal of Biometrics* 1, 1 (2008), 81–113.

[297] Wencheng Yang, Song Wang, Jiankun Hu, Guanglou Zheng, and Craig Valli. 2019. Security and accuracy of fingerprint-based biometrics: A review. *Symmetry* 11, 2 (2019), 141.

[298] Adrienne Yapo and Joseph Weiss. 2018. Ethical implications of bias in machine learning. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

[299] Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu. 2019. Towards interpretable face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9348–9357.

[300] Ying Yin, Tom Yu Ouyang, Kurt Partridge, and Shumin Zhai. 2013. Making Touchscreen Keyboards Adaptive to Keys, Hand Postures, and Individuals: A Hierarchical Spatial Backoff Model Approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2775–2784. DOI:http://dx.doi.org/10.1145/2470654.2481384

[301] Soweon Yoon, Jianjiang Feng, and Anil K Jain. 2012. Altered fingerprints: Analysis and detection. *IEEE transactions on pattern analysis and machine intelligence* 34, 3 (2012), 451–464.

[302] Rongyu Yu, Burak Kizilkaya, Zhen Meng, Emma Li, Guodong Zhao, and Muhammad Imran. 2023. Robot Mimicry Attack on Keystroke-Dynamics User Identification

and Authentication System. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9879–9884.

[303] Saira Zahid, Muhammad Shahzad, Syed Ali Khayam, and Muddassar Farooq. 2009. Keystroke-Based User Identification on Smart Phones. In *LNCS*, Vol. 5758. 224–243. DOI:http://dx.doi.org/10.1007/978-3-642-04342-0_12

[304] Juan Jose Zarate, Thomas Langerak, Bernhard Thomaszewski, and Otmar Hilliges. 2020. Contact-free Nonplanar Haptics with a Spherical Electromagnet. In *2020 IEEE Haptics Symposium (HAPTICS)*. IEEE, 698–704.

[305] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[306] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

[307] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. 2003. Face recognition: A literature survey. *ACM computing surveys (CSUR)* 35, 4 (2003), 399–458.

[308] Dexing Zhong, Xuefeng Du, and Kuncai Zhong. 2019. Decade progress of palmprint recognition: A brief survey. *Neurocomputing* 328 (2019), 16–28.

[309] Gabe Zichermann and Christopher Cunningham. 2011. *Gamification by design: Implementing game mechanics in web and mobile apps*. " O'Reilly Media, Inc.".

[310] Mary Ellen Zurko and Richard T Simon. 1996. User-centered security. In *Proceedings of the 1996 workshop on New security paradigms*. 27–33.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig und ohne unerlaubte Beihilfe angefertigt wurde.

16. Oktober 2024

Lukas Benjamin Mecke