USENIX Association

Proceedings of the Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)

August 10–11, 2020

© 2020 by The USENIX Association

All Rights Reserved

This volume is published as a collective work. Rights to individual papers remain with the author or the author's employer. Permission is granted for the noncommercial reproduction of the complete work for educational or research purposes. Permission is granted to print, primarily for one person's exclusive use, a single copy of these Proceedings. USENIX acknowledges all trademarks herein.

ISBN 978-1-939133-16-8

Cover Image created by freevector.com and distributed under the Creative Commons Attribution-ShareAlike 4.0 license (https://creativecommons.org/licenses/by-sa/4.0/).

Symposium Organizers

General Chair
Heather Richter Lipford, *University of North Carolina at Charlotte*

Vice General Chair
Sonia Chiasson, *Carleton University*

Technical Papers Co-Chairs
Joe Calandrino, *Federal Trade Commission*
Michelle Mazurek, *University of Maryland*

Technical Papers Committee
Adam Aviv, *The George Washington University*
Lujo Bauer, *Carnegie Mellon University*
Cristian Bravo-Lillo, *Ciberseguridad Humana, Chile*
Lynne Coventry, *Northumbria University*
Lorrie Cranor, *Carnegie Mellon University*
Heather Crawford, *Florida Institute of Technology*
Serge Egelman, *International Computer Science Institute (ICSI) and University of California, Berkeley*
Sascha Fahl, *Leibniz University Hannover*
Carrie Gates, *Bank of America*
Maximilian Golla, *Max Planck Institute for Cyber Security and Privacy*
Julie Haney, *National Institute of Standards and Technology (NIST)*
Marian Harbach, *Google*
Jun Ho Huh, *Samsung Research*
Maritza Johnson, *Google*
Apu Kapadia, *Indiana University Bloomington*
Jennifer King, *Stanford University*
Katharina Krombholz, *CISPA Helmholtz Center for Information Security*
Janne Lindqvist, *Aalto University and Rutgers University*
Abigail Marsh, *Macalester College*
Susan E. McGregor, *Tow Center for Digital Journalism, Columbia Journalism School*
Robert Reeder, *Google*
Irwin Reyes, *Two Six Labs*
Scott Ruoti, *University of Tennessee, Knoxville*
Florian Schaub, *University of Michigan*
Kent Seamons, *Brigham Young University*
Manya Sleeper, *Google*
Elizabeth Stobert, *Carleton University*
Jose Miguel Such, *King's College London*
Blase Ur, *University of Chicago*
Emanuel von Zezschwitz, *Google*
Rick Wash, *Michigan State University*
Josephine Wolff, *Fletcher School, Tufts University*
Mary Ellen Zurko, *MIT Lincoln Laboratory*

Lightning Talks and Demos Co-Chairs
Yousra Javed, *Illinois State University*
Blase Ur, *University of Chicago*

Karat Award Chair
Rick Wash, *Michigan State University*

Posters Co-Chairs
Heather Crawford, *Florida Institute of Technology*
Maximilian Golla, *Max Planck Institute of Cyber Security and Privacy*

Tutorials and Workshops Co-Chairs
Katharina Krombholz, *CISPA Saarland*
Daniel Zappala, *Brigham Young University*

Mentoring Co-Chairs
Adam Aviv, *The George Washington University*
Daniela Napoli, *Carleton University*
Daniel Votipka, *University of Maryland*

Publicity Co-Chairs
Nalin Asanka Gamagedara Arachchilage, *University of New South Wales*
Yaxing Yao, *Syracuse University*

Email List Chair
Lorrie Cranor, *Carnegie Mellon University*

Accessibility Chair
Rich Williams, *USENIX Association*

USENIX Liaison
Casey Henderson, *USENIX Association*

Extended Program Committee

Yasemin Acar
Pam Briggs
Sunny Consolvo
Patrick Kelley
Bart Knijnenburg
Matthew Smith
Yang Wang

External Reviewers

Yasmin Acar
Tousif Ahmed
David Barrera
Merlin Chlosta
Camille Cobb
Shaneé Dawkins
Florian Farke
Matthias Fassl
Clement Fung
Shakthidhar Reddy Gopavaram
Roberto Hoyle
Tyler Kaczmarek
Bart Knijnenburg
Philipp Markert
Ian Oakley
Kopo M. Ramokapane
Theodor Schnitzler
Hyejin Shin
Bich Vu
Simon S. Woo

Message from the SOUPS 2020 Program Co-Chairs

Welcome to SOUPS 2020!

Although SOUPS turns 16 during a uniquely challenging year, we are happy to see that the conference continues to thrive. We are pleased this year to present a program that covers a broad range of topics within usable privacy and security. Technical paper presentations form the core of the SOUPS program, but the conference also includes workshops, posters, lightning talks, mentorship, and a keynote. This year, SOUPS received 131 valid paper submissions, accepting 26 (20% acceptance rate).

In 2016, SOUPS became an independent conference body. For the last four years, we have partnered with USENIX for hosting and administrative support, a move that has enabled continued growth for the conference. We thank all the members of the USENIX staff for their work in organizing SOUPS and supporting our community. We particularly appreciate their assistance and flexibility this year, including managing the shift to a virtual event.

In 2018, we were co-located with the USENIX Security Symposium for the first time, and we are continuing that co-location for 2020 and 2021. Co-locating the two conferences allows for interactions and shared ideas between SOUPS and USENIX Security attendees. We have found this beneficial for both conferences and look forward to repeating it virtually this year and in-person in Vancouver, BC, next year.

SOUPS relies on a range of volunteers for all of its activities. Steering Committee members provide oversight and guidance and are elected for three-year terms. Organizing Committee members help determine the conference content for a particular year, often serving two-year terms to facilitate the transition of knowledge. Technical Papers Committee members are chosen by the Technical Papers Co-Chairs each year. This year, the pandemic created sudden and unexpected obstacles for many of us, including members of the Technical Papers Committee. As a result, other members of the committee and external reviewers—including an Extended Technical Papers Committee comprised of experienced community members—stepped in to assist with reviews. SOUPS is a product of the hard work by all the SOUPS Organizers, the SOUPS Steering Committee, the technical paper reviewers, the workshop organizers, the poster jury, and the USENIX staff. We thank each and every one of you for your contributions to SOUPS 2020.

Heather is serving her second and final year as General Chair of SOUPS and Chair of the Steering Committee. In 2021 and 2022, Sonia will serve as General Chair. If you are interested in helping with SOUPS 2021 in any way, please contact Sonia.

We thank our sponsors. SOUPS would not be possible without their generous support. Please visit our website to view the recipients of the SOUPS 2020 awards. Congratulations to all of the recipients for their outstanding work.

Heather Richter Lipford, *University of North Carolina at Charlotte*
General Chair

Sonia Chiasson, *Carleton University*
Vice General Chair

Michelle Mazurek, *University of Maryland*
Technical Papers Co-Chair

Joe Calandrino, *Federal Trade Commission*
Technical Papers Co-Chair

Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)

August 10–11, 2020

Monday, August 10

Authentication

Passworld: A Serious Game to Promote Password Awareness and Diversity in an Enterprise **1**
Gokul Chettoor Jayakrishnan, Gangadhara Reddy Sirigireddy, Sukanya Vaddepalli, Vijayanand Banahatti, and Sachin Premsukh Lodha, *TCS Research, Tata Consultancy Services Limited, Pune, India;* Sankalp Suneel Pandit, *Former employee of TCS Research, Tata Consultancy Services Limited, Pune, India*

"You still use the password after all" – Exploring FIDO2 Security Keys in a Small Company **19**
Florian M. Farke, *Ruhr University Bochum;* Lennart Lorenz, *tracekey solutions GmbH;* Theodor Schnitzler, Philipp Markert, and Markus Dürmuth, *Ruhr University Bochum*

Knock, Knock. Who's There? On the Security of LG's Knock Codes **37**
Raina Samuel, *New Jersey Institute of Technology;* Philipp Markert, *Ruhr University Bochum;* Adam J. Aviv, *The George Washington University;* Iulian Neamtiu, *New Jersey Institute of Technology*

An Empirical Study of Wireless Carrier Authentication for SIM Swaps **61**
Kevin Lee, Benjamin Kaiser, Jonathan Mayer, and Arvind Narayanan, *Princeton University*

Appreciating User Needs and Concerns

Lessons Learnt from Comparing WhatsApp Privacy Concerns Across Saudi and Indian Populations **81**
Jayati Dev, *Indiana University;* Pablo Moriano, *Oak Ridge National Laboratory;* L. Jean Camp, *Indiana University*

Realizing Choice: Online Safeguards for Couples Adapting to Cognitive Challenges **99**
Nora McDonald, Alison Larsen, and Allison Battisti, *University of Maryland, Baltimore County;* Galina Madjaroff, *University of Maryland;* Aaron Massey and Helena Mentis, *University of Maryland, Baltimore County*

Blind and Human: Exploring More Usable Audio CAPTCHA Designs **111**
Valerie Fanelle, Sepideh Karimi, Aditi Shah, Bharath Subramanian, and Sauvik Das, *Georgia Institute of Technology*

Usable Sexurity: Studying People's Concerns and Strategies When Sexting **127**
Christine Geeng, Jevan Hutson, and Franziska Roesner, *University of Washington*

Towards Understanding Privacy and Trust in Online Reporting of Sexual Assault **145**
Borke Obada-Obieh, *University of British Columbia;* Lucrezia Spagnolo, *Vesta Social Innovation Technologies;* Konstantin Beznosov, *University of British Columbia*

Implementing Security and Privacy: Design and Development

Replication: On the Ecological Validity of Online Security Developer Studies: Exploring Deception in a Password-Storage Study with Freelancers **165**
Anastasia Danilova, Alena Naiakshina, and Johanna Deuter, *University of Bonn;* Matthew Smith, *University of Bonn, Fraunhofer FKIE*

Innovation Inaction or In Action? The Role of User Experience in the Security and Privacy Design of Smart Home Cameras **185**
George Chalhoub and Ivan Flechais, *University of Oxford;* Norbert Nthala, *Michigan State University;* Ruba Abu-Salma, *University College London (UCL) & Inria*

An Ethnographic Understanding of Software (In)Security and a Co-Creation Model to Improve Secure Software Development **205**
Hernan Palombo, Armin Ziaie Tabari, Daniel Lende, Jay Ligatti, and Xinming Ou, *University of South Florida*

Why Can't Johnny Fix Vulnerabilities: A Usability Evaluation of Static Analysis Tools for Security **221**
Justin Smith, *Lafayette College;* Lisa Nguyen Quang Do and Emerson Murphy-Hill, *Google*

Tuesday, August 11

Implementing Security and Privacy: Operation, Administration, Vulnerability Analysis

Security, Availability, and Multiple Information Sources: Exploring Update Behavior of System Administrators . . 239
Christian Tiefenau and Maximilian Häring, *University of Bonn;* Katharina Krombholz, *CISPA Helmholtz Center for Information Security;* Emanuel von Zezschwitz, *University of Bonn, Fraunhofer FKIE*

An investigation of phishing awareness and education over time: When and how to best remind users 259
Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, and Reyhan Duezguen, *SECUSO - Security, Usability, Society, Karlsruhe Institute of Technology;* Bettina Lofthouse, *Landesamt für Geoinformation und Landesvermessung Niedersachsen;* Tatiana von Landesberger, *Interactive Graphics Systems Group, Technische Universität Darmstadt;* Melanie Volkamer, *SECUSO - Security, Usability, Society, Karlsruhe Institute of Technology*

An Exploratory Study of Hardware Reverse Engineering — Technical and Cognitive Processes 285
Steffen Becker, Carina Wiesen, and Nils Albartus, *Ruhr University Bochum, Max Planck Institute for Cybersecurity and Privacy;* Nikol Rummel, *Ruhr University Bochum;* Christof Paar, *Max Planck Institute for Cybersecurity and Privacy*

Industry Responses to the European Directive on Security of Network and Information Systems (NIS): Understanding policy implementation practices across critical infrastructures 301
Ola Aleksandra Michalec, Dirk van der Linden, Sveta Milyaeva, and Awais Rashid, *University of Bristol*

"You've Got Your Nice List of Bugs, Now What?" Vulnerability Discovery and Management Processes in the Wild . .319
Noura Alomar, *University of California, Berkeley;* Primal Wijesekera, *University of California, Berkeley, and International Computer Science Institute (ICSI);* Edward Qiu, *University of California, Berkeley;* Serge Egelman, *University of California, Berkeley, and International Computer Science Institute (ICSI)*

Applications and User Perceptions

User Mental Models of Cryptocurrency Systems - A Grounded Theory Approach 341
Alexandra Mai and Katharina Pfeffer, *SBA Research;* Matthias Gusenbauer, *Tokyo Institute of Technology, SBA Research;* Edgar Weippl, *University of Vienna;* Katharina Krombholz, *CISPA Helmholtz Center for Information Security*

Cloudy with a Chance of Misconceptions: Exploring Users' Perceptions and Expectations of Security and Privacy in Cloud Office Suites 359
Dominik Wermke, Nicolas Huaman, Christian Stransky, Niklas Busch, Yasemin Acar, and Sascha Fahl, *Leibniz University Hannover*

From Intent to Action: Nudging Users Towards Secure Mobile Payments 379
Peter Story, Daniel Smullen, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh, *Carnegie Mellon University;* Florian Schaub, *University of Michigan*

Do Privacy and Security Matter to Everyone? Quantifying and Clustering User-Centric Considerations About Smart Home Device Adoption 417
Natã M. Barbosa, Zhuohao Zhang, and Yang Wang, *University of Illinois at Urbana-Champaign*

Privacy, Tracking, and Measurement

Enhancing Privacy through the Visual Design of Privacy Notices: Exploring the Interplay of Curiosity, Control and Affect 437
Agnieszka Kitkowska, *Karlstad University;* Mark Warner, *Northumbria University;* Yefim Shulman, *Tel Aviv University;* Erik Wästlund and Leonardo A. Martucci, *Karlstad University*

"I Have a Narrow Thought Process": Constraints on Explanations Connecting Inferences and Self-Perceptions . . 457
Emilee Rader, Samantha Hautea, and Anjali Munasinghe, *Michigan State University*

Replication: Why We Still Can't Browse in Peace: On the Uniqueness and Reidentifiability of Web Browsing Histories 489
Sarah Bird, Ilana Segall, and Martin Lopatka, *Mozilla*

How Risky Are Real Users' IFTTT Applets? 505
Camille Cobb and Milijana Surbatovich, *Carnegie Mellon University;* Anna Kawakami, *Wellesley College;* Mahmood Sharif, *NortonLifeLock;* Lujo Bauer, *Carnegie Mellon University;* Anupam Das, *North Carolina State University;* Limin Jia, *Carnegie Mellon University*

Passworld: A Serious Game to Promote Password Awareness and Diversity in an Enterprise

Gokul Chettoor Jayakrishnan, Gangadhara Reddy Sirigireddy, Sukanya Vaddepalli, Vijayanand Banahatti, Sachin Premsukh Lodha, Sankalp Suneel Pandit[1]
TCS Research, Tata Consultancy Services Limited, Pune, India ([1]former employee)
{gokul.cj, gangadhara.sirigireddy, sukanya.vaddepalli, vijayanand.banahatti, sachin.lodha}@tcs.com, pandit.sankalp123@gmail.com

Abstract

Usage of weak passwords for authentication within an organization can be exploited during cyberattacks leading to unauthorized account access, denial of service, data and identity theft, sabotage etc. Such attacks could bring financial and reputational losses apart from legal consequences. Organizational password policies came into being in an attempt to encourage users to create more complex and diverse passwords. However, it has been observed that people show similar behavior in adopting those policies and end up creating passwords with similar patterns. Security training has been found to be a popular mechanism in an enterprise setting, of which, game-based trainings have shown positive impact with an added advantage of being immersive. In this paper, we present a serious game-based training on creating password security awareness among enterprise users. The training involves promoting understanding among users about various common password heuristics during password creation. This study focuses on two research questions: 1) Can a game-based password awareness training teach participants about the various password heuristics? 2) Can such a training improve the organizational password diversity? With a participation of 4,906 employees from our enterprise in the study, we were able to observe effects of game-based training on password awareness. We also found insights during the study to show that users created diverse passwords.

1. Introduction

Despite advancements in user authentication methods, the decades old system of username-password combination is still prevalent [54, 57]. Even the method of two-factor authentication generally consists of passwords as one of its factors [3, 40]. The human element involved in password creation is one of the major factors affecting password strength [35]. Studies show that people are more likely to use weaker and easily memorable passwords because of the lack of knowledge in creating stronger passwords [20] or due to the limitations in memorizing passwords [56]. This makes it easier for attackers to crack the passwords, resulting in security breaches and loss of personal and confidential information. Organizations have had their fair share of difficulties in dealing with password breaches [12, 30, 45, 46]. In an enterprise scenario, attackers often target individuals to steal passwords and gain access to organizational data [6]. Verizon's Data Breach Investigation Report (DBIR) [52] states that 63% of data breaches worldwide happened due to weak or stolen passwords or usage of default passwords. Organizational password policies such as setting a password expiration period could also lead to adoption of insecure methods like writing down the passwords [19] or using newer passwords with minimal difference from the old ones. For example, appending symbols and/or digits at the end [37] could result in weak and similar passwords. These fall under insecure password practices and avoiding them will help create better passwords.

Password meters are found to be useful in creating stronger passwords [51]. However, most of them provide basic feedback [13, 51] and many of them also rate passwords inconsistently [39]. There is a need to create awareness among users when it comes to creating better passwords. An advanced data-driven password meter by Ur, et al. [48] that uses several heuristics to score passwords showed compelling results, but with limitations.

Among various training methods, text-based training like reading documents has been said to be monotonous [10]. However, interactive games are found to be helpful and immersive [8, 41, 53] when it comes to training in cybersecurity. Considering this, we intended to provide an interactive training experience to the employees of our

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2020. August 9 -- 11, 2020, Virtual Conference.

organization by following a game-based password awareness methodology. We utilized the heuristics from the previous study [48] to create a comprehensive training method in order to teach password practices. Heuristics, in the context of this study, denote the techniques or practices that have to be satisfied to improve the overall password strength. For example, a heuristic like "Not more than three consecutive repeating characters" will be satisfied if the created password does not have more than three consecutive repeating characters. We conducted the game-based experiment as a part of our organization's annual information security awareness week. The online 2D web game titled Passworld focuses on enterprise password training, with little compromise on the fun element of gameplay. This paper details the experiment, results, and observations obtained from our study.

2. Related Work

There are several previous studies relating to user behaviors, tools for password security, and game-based cybersecurity awareness training.

2.1. Password Security: User Behaviors

When it comes to user behavior, studies [2, 50] show that users assess various cybersecurity threats by themselves and their assessment often ends up being inaccurate, mainly because they are not knowledgeable in this regard. The usage of multiple online accounts leads to password reuse [2]. This, combined with insecure creation practices, make passwords all the more vulnerable. The users' cognitive ability to simultaneously deal with multiple passwords is limited [23]. A recent study on users' perceptions on password strength [50] shows that many users' idea of strong passwords, like adding digits and symbols at certain positions, is exploited by today's password cracking tools. Another study [49] shows that even with the awareness about importance of creating strong passwords, many users create passwords that are easy to guess. While many users did not know how to create strong passwords, many believed that they were creating strong passwords, but in fact, those passwords were just predictable [49]. As per Habib et al. [20], in workplaces with a password expiration policy, many users tend to use password creation strategies that are very predictable and with minimal modifications. A study with adaptive password blacklisting policies [38] was found to have good results. However, the study also advises using a good password-composition policy to augment the structure-based adaptive, blacklist system. Certain tools exist to help users create better passwords.

2.2. Existing Tools for Password Security

Proactive password checking is a method of checking a user's password to see if it satisfies certain criteria to consider it as secure [23, 56]. Password strength meters implement this by checking several criteria for testing the strength of a password. A recent study with a data driven password meter [48] showed positive results, which were attributed to the feedback mechanism implemented into the tool. However, the study also concluded that the overall effect of the meter was not strong and some participants did not trust suggestions from a computer. Studies on users' perceptions on password systems and behaviors [2, 55, 57] have recommended the need to provide training and instructions on constructing usable and secure passwords, and to provide adequate feedback during training to enhance their knowledge.

2.3. Existing Password and Security Awareness Trainings

A report by security firm LastPass in 2018 [24] revealed that only 19% of users create strong passwords at work and 62% of users use the same passwords for work and personal accounts. It implies that knowing an employee's personal password makes it easier for hackers to guess their work password [24]. This suggests the importance of making enterprise users aware of the various password heuristics that could help them create secure passwords for both personal and work accounts. Many enterprise-training methods like educational documents, mail-based embedded training [29] etc. are available. However, these are considered passive training methods that are found to be less engaging and monotonous [10]. On the contrary, interactive games facilitate participation between the teaching agent and learner [32, 53].

2.4. Interactive Training: Serious Games

Games with a purpose, titled serious games, perform an additional task of creating awareness apart from entertainment [5]. Serious games have been used to spread cybersecurity awareness with successful results. Previous studies with games like Anti Phishing Phil [41], CyberCIEGE [44], Phishy[8], GAP [47], Cyberaware [18], PASDJO[39], and Control-Alt-Hack [14] have shown that games are not only effective in training cybersecurity concepts, but are also more engaging, compared to the traditional methods like reading documents.

Considering the existing password awareness games, PASDJO [39] trains users on measuring the strengths of on-screen passwords through the method of inspection. GAP [47] focuses on training users on predictable positions of uppercase letters, digits, and symbols. While these game-based studies provided good results in terms of user's understanding when it came to insecure password

practices, they focused on a very small set of password heuristics. We tried to extend this study to a larger audience. We followed a game-based approach for password training that focused on a set of 16 password heuristics. To bring the game more in line with the latest findings and studies, we used the results from the study by Ur, et al. [48] and a set of password recommendations from a study by Vu, et al. on improving password security and memorability [54]. These were analyzed in light of the organizational password complexity requirements, and were organized to form a set of heuristics to be taught through the game. The tool used in a previous study [38], suggests modifications to passwords entered by users in order to make them compliant with policies. Our intent was to test the user understanding by letting them apply their learnings about the heuristics, to their passwords without any additional suggestions on modifications to the entered password. The reports from our organization's Corporate Security Team (CST), which takes care of awareness training, suggested from their data that employees participate in mandatory training modules because of the condition that they must finish it. This condition could provide a sense of urgency to complete, rather than understand the course content with a curiosity to learn. So, as suggested by CST, we decided on a non-mandatory, gamified way of training to create awareness about passwords among participants. The training was conducted in collaboration with CST, which also took care of all the clearances and ethics review. The game was made part of the normal workday.

3. Passworld Game

We developed an online, web-based game to spread awareness about various password heuristics among the employees of our organization. Our unique game titled "Passworld" follows certain basic principles to ensure that an optimal learning outcome is obtained. We followed a set of established theories and principles during the design phase of our game.

Experiential Gaming Model: Games have been shown to be most successful and engaging when they provide a flow experience [11] to players [26]. An optimal flow experience, combined with experiential learning and feedback, termed as the Experiential Gaming model, is found to maximize the impact of a training game [26]. Passworld tries to incorporate a similar learn and reflect methodology with emphasis on immersion and enjoyment by providing learning tips, timely feedback, and an easy-to-learn interface, similar to the ones in the classic and widely famous games of the 80s and 90s like Super Mario Bros. [34] and Adventure Island [21].

Bloom's Taxonomy: The game also follows all the six levels of Bloom's Taxonomy in the cognitive domain [27, 42] to maximize the learning outcome. The game instructions, learning tips, and basic information on password heuristics form the first level *Knowledge* in the Bloom's hierarchy. The game presents these learning materials in a fun and entertaining way to help with *Comprehension*. Players have to apply this knowledge while creating passwords, which satisfies *Application* level in the hierarchy. Further, players *Analyze* each password heuristic and *Synthesize* their own knowledge by adding up pieces of information obtained from the game levels. The instant and delayed feedback in the game helps them to *Evaluate* their choices, thereby helping in reflection, a learning science principle [15].

Password also follows the **Conceptual-Procedural** gaming principle [41] to ensure a better learning outcome. This principle states that conceptual and procedural knowledge augments one another in an iterative process [25]. This, along with game design patterns like integration pattern, cognition pattern, and presentation pattern [5], helped us decide a flow and overall design for the game. We also ensured that the Six "I" Framework of Serious Game Design [4] that focuses on Identity, Immersion, Interactivity, Increased Complexity, Informed Teaching, and Instructional Content, is followed.

3.1. Game Description

Passworld is a 2D, single-player, horizontal scrolling platformer game [9] falling in the Action-Adventure [33] genre, developed in the Unity3D game engine. We tried to keep the game as lightweight as possible for compatibility with all the browser versions and machines.

3.1.1. Game Design Choices

We decided on providing a positive gameplay experience by intertwining password awareness with a platformer-based game, where both intrinsic motivation of fun element of gameplay as well as extrinsic motivation like rewards and benefits merge [17]. Passworld was made to have a gameplay experience similar to that of certain classic games [21, 34] to create a positive mindset in players, through a feeling of nostalgia [7]. We chose a horizontal, jungle-based platformer game considering the non-monotonous gameplay factor, game time, interactions and visuals, and the analogy that reflects real-life scenario of password attacks. We tried to relate the real world to our game through the following design choices:

Open-interconnected world: The real cyber world is always open and interconnected. The chances of cyber-attacks and password breaches are high, if we do not follow proper security measures. We wanted to create a similar scenario within the game. Therefore, we chose a jungle

environment, which is a wide-open scenario. If the user is not careful, the chances of being attacked by animals are much higher.

Digital assets need protection where passwords play an important role. Similarly, the game focuses on storing important ancient artifacts using secure gates.

Know thy enemy: In order to create secure passwords, one must know the weaknesses of passwords that are exploited by attackers. This is where the game introduces various animals. The animals in the game provide tips about several password heuristics. They also check the user created passwords to see if heuristics are satisfied, and if not, they attack. We used the password heuristics from a previous study [48]. While [48] was based on using the heuristics to check and analyze user created passwords, and provide suggestions for improvement, we used the heuristics to teach users about individual password requirements.

Prepare to defend: In real life, we can create passwords using all available character classes. We converted this into a game resource, where players have to gather the different character classes to create their passwords (gates). The resources are the raw materials used to create the gates, just like L, U, D, and S character classes are used to create a password. Typical enterprise password policies mandate usage of all character classes while creating passwords.

Build a strong defense: Passworld teaches users about password heuristics and requires them to apply their learnings to create strong and memorable passwords that satisfy all the heuristics.

3.1.2. Game Mechanics

The game storyline is based in a fictional world. Learning experiences have found to be enhanced using story-based agents [41]. The gameplay of a level starts with a pre-test, then the game, followed by a post-test. The game consists of two levels, Level 1 and Level 2, and each level consists of different sequential stages for pre-test, gameplay, password creation, distraction task, password recall, and post-test (cf. Figure 1). Based on our design choices, we framed our game story and various gameplay elements, as follows:

Jungle environment: The protagonist, Soma, is an archaeologist who is in search of two ancient artifacts that were lost years ago in a land called "Passworld". Soma has to travel two days and two nights through the jungles of Passworld to find them (cf. Figure 2). The two days are represented by Level 1 and Level 2 gameplays.

Securely storing the artifact: Since the ancient artifacts are precious, Soma has to store them after collecting, to protect them from being stolen. This is done by creating strong gates (analogous to secure passwords) around Soma's camp. This happens in the two password creation stages (cf. Figure 4), represented by two nights.

Learning the password heuristics: In the two main levels, the players can interact with oncoming animals during gameplay to learn about various password heuristics (cf. Figure 3). Every oncoming animal will raise the curiosity of the player by showing basic heuristic details as a riddle (E.g. Fox in. Figure 2). The player can choose to "*know more*" about a particular heuristic by clicking the animal's heuristic text. This will pop up a detailed description of the heuristic and certain statistics associated with it (cf. Figure 3). If these heuristics are not satisfied during each password creation stage, the corresponding animal will attack the password gate (cf. Figure 5), and enter the camp. This also signifies how a password meter checks for various heuristics [48]. While in [48], the users are not required to satisfy all the password heuristics in an entered password, our game has this requirement as we wanted to teach all the available heuristics to the users and tell them that every single one of them is important.

Resource gathering: The resources for creating these gates are obtained throughout the journey, in the form of tablets with character classes mentioned as L, U, D, and S (Lowercase, Uppercase characters, Digits, and Symbols) (cf. Figure 7). In real life, these character classes are required for creating a password.

Creation of password gates: Once the player collects the artifacts, stores them using secure password gates (cf. Figure 4), they complete one full day in the game. We introduced two activities post each password creation stage that act as distraction tasks. Distraction tasks [31] distract the players for a brief period after password entry, to encourage them to create memorable passwords. Our tasks are two mini activities that ask the players to arrange certain items correctly (using drag and drop) to a) Ignite a campfire b) Cook food (cf. Figure 8). This step is added to promote awareness about the importance of creating memorable passwords. To continue the journey further on the next day, the player has to unlock the gate using the same password (cf. Figure 4). This password recall stage is where password memorability is tested.

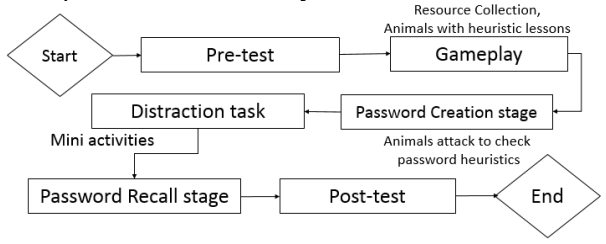


Figure 1: Passworld Game Level Flow

If players fail a level, it can be replayed again. The game did not have timers as these might have created unwanted sense of urgency that could have limited gameplay experience [43]. Passworld used simple controls with arrow keys for navigation, jumping, and mouse clicks for selections.

3.2. Implementation of Game-based Training

The game implements the use of feedback and instructions to promote learning throughout.



Figure 2: Passworld Level 1 Gameplay

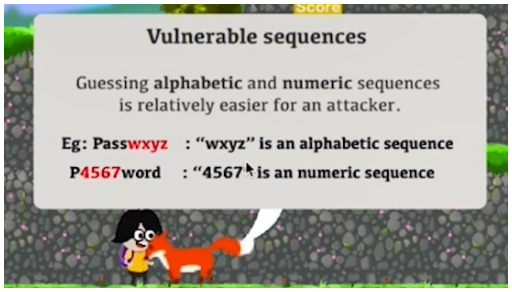


Figure 3: Heuristic description received from animal



Figure 4: Password creation (Gate Creation) stage. The password recall stage is similar in appearance, but without the options to view password heuristics (yellow animal icons).



Figure 5: Animal (fox) approaching to attack the gate.

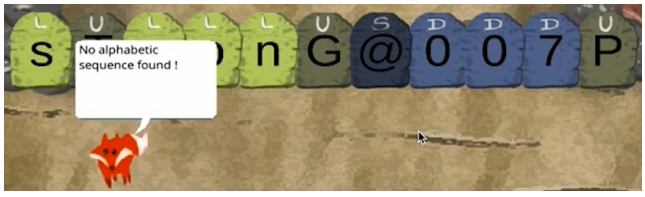


Figure 6: Animal returning as the corresponding heuristic (no alphabetic sequences found) was satisfied.



Figure 7: Game Resources and artifacts- L, U, D, S; the numbers on the tablets correspond to the count of each collected resource. The artifacts (one for each Level) depict important information that the player should store securely.



Figure 8: Distraction Task, Level 2

The feedback methodology was added to promote self-learning and reflection [**15**]. The game implements the following methods to promote learning.

Instructions: The oncoming animals provide instructions on various heuristics to the players. We added different animals to provide a visual identity to each heuristic, to make it more memorable. The same heuristic information is also available during password creation stages to help users learn the password creation strategies. Therefore, players who feel reading the information during gameplay is disruptive can read it at a further time, while creating the password. The heuristics are taught one at a time, but at the end of the level, the created password should incorporate all these heuristics. Even though a strong password does not need all these heuristics to be satisfied [48], we did this in order to teach and make the users understand that every heuristic is important.

Feedback: As soon as the player enters a password during the password creation stage, immediate feedback is received, indicating to them the potential vulnerabilities within the password entered. The study on adaptive password blacklisting policies [38] introduces an interface to provide the users with suggestions on modifying passwords to conform to the policies. We used our study to make this process voluntary. We did not provide suggestions to the users, but only feedback on if they satisfied certain password heuristics or not.

3.3. Game Data Recorded

Data captured and stored in the form of game data included the demographic information, pre-test, post-test, and feedback survey responses, various time stamps, gameplay

Password Heuristics	Heuristic Description (and Corresponding Animal in Game)	Levels taught	Identi-fier	Clus-ter
Length	Password length must be more than 8 **(Default policy)**	1,2	H1	
Lowercase present	At least 1 lowercase character must be present **(Default policy)**	1,2	H2	
Uppercase present	At least 1 uppercase character must be present **(Default policy)**	1,2	H3	C1
Digit present	At least 1 digit must be present **(Default policy)**	1,2	H4	
Symbol present	At least 1 symbol must be present **(Default policy)**	1,2	H5	
Repeated characters	Not more than 3 consecutive repeating characters (e.g. "honeeeey" has 4 repeating 'e's,). **(Animal: Raccoon)**	1,2	H6	
Duplicated characters	Number of duplicated characters should not be more than 50% of total character count in password **(Animal: Porcupine)**	1,2	H7	
Repeated sections	3 or more repeating set of characters (e.g. honeyhoney honey, honeyyenohhoney) should not be present **(Animal: Fox)**	1,2	H8	C2
Alphabetic sequences	Not more than 3 consecutive Alphabetic or numerical sequences should be present(e.g. 12345, ghiJkLm etc.) **(Animal: Fox)**	1,2	H9	
Predictable positions of:	Checking the predictable positions of symbols, digits and uppercase characters **(Animal: Monkey)**	2	H11	
Symbols	Symbols should not be present just at the end (e.g. passwor@) or password should not be having a common "letters-symbols-digits" format (e.g. pasS@921)			
Digits	Digits present in the beginning or at the end or password having all digits is a common pattern, should be avoided	2	H12	
Uppercase Characters	UPPERCASE order present in the beginning or all characters in password being uppercase, is a common pattern, should be avoided	2	H13	C3
Predictable structure	Password should not fall under a set of common password structures (e.g. LSLLLDLD) **(Animal: Leopard)**	2	H14	
Keyboard patterns	4+ equally spaced keys in password (e.g. QWERTY) **(Animal: Hyena)** should not be present (US-English language keyboards).	2	H10	
Date formats	DDMMYY – with and without delimiters and years (1900-2049), in all formats including month names should not be present. **(Animal: Bear)**	2	H15	C4
Blacklists	Password should not contain common organization related words anywhere in it. **(Animal: Snake)**	1,2	H16	C5

Table 1: Password Heuristics Checked Within the Game

data, heuristics viewed by the player, password structures entered, level attempt counts, heuristics (failed and successful) and password creation, recall attempts. The game converts passwords entered by users into their respective character structures and stores in the database. Password structure [38] is an ordered sequence that captures the password's composition using four character classes. These classes are L, U, D, and S, for lowercase and uppercase characters, digits, and symbols respectively. For example, a password like "P@ssw0rd" will only be stored as "USLLLDLL" instead of its plain text for analysis.

3.4. Password Heuristics

The game trains in a set of 16 password heuristics, with each heuristic being tagged to a particular animal (as shown in Table 1). A previous study by Ur, et al. [48] found them to be effective in increasing password strength. These password heuristics, categorized as two sets based on increasing complexity, were added to the two game levels. The first level has basic password requirements like length (H1), presence of character classes (H2-H5), alphabetic sequences (H9) etc. of which, H1-H5 were part of our organization's default password policies. The second level focuses on the heuristics from the first level along with new heuristics that check formatting, repeated sections in passwords, date formats etc. A set of common words related to the organization (classified as "blacklisted" passwords) were added as a check as well. We also compared user created password structures with over 2,124 structures obtained from the previous study [48]. These

heuristics were also clustered based on the common characteristics they possess, as C1 having basic password heuristics, C2 with character sequences, C3 with predictable positions, C4 having certain patterns, and C5 with the blacklists. We have taught these heuristics through the game's main levels, and let the players incorporate these heuristics while creating the passwords in password creation stages, thus letting the players demonstrate what they have learnt.

3.4.1. How User Created Passwords are Checked

The game checks users' passwords through the following steps:

a. As soon as a player uses the resources (L, U, D, and S) to form a password gate, default checks for length and presence of all character classes are done. If any of them is not satisfied, the game shows appropriate error messages to the player instantly.

b. Once the password satisfies the basic criteria, the password heuristics evaluation begins.

c. For each heuristic, an animal approaches the gate (cf. Figure 5). If the corresponding heuristic is satisfied, the animal leaves (cf. Figure 6); else, it attacks the gate and enters the camp resulting in a penalty as loss of life. This process repeats until all level heuristics are satisfied (level cleared) or when all life is lost (level failed). After this, the player continues to the next level or goes back to the start of the level respectively.

4. Study Design

The goals of our study were to find the effectiveness of a game-based enterprise password awareness training on various password heuristics, and to identify if such a training could be beneficial to enterprise password diversity. Previous studies [8, 41] show that game-based methods have shown better results than text-based means, when it comes to cybersecurity training. We utilized this result to test if games could help in password awareness training, and we measured this using pre and post-tests along with the game. The following sections show our study procedure and evaluation results.

4.1. Participant Demographics

Game participants were employees of our organization. They were recruited for the study using mailers about the game. Interested participants clicked on the game URL within the mail to access the game. Though equipped with computer knowledge, the participants had varied understanding of gaming and password awareness. Passworld was online for one month and was played by 4,906 participants from around the globe. We selected a set of lucky winners from the participants who completed the game (20 people per day), and rewarded each of them using our organization's equivalent of virtual currency (with a monetary value of approximately $4). Table 2 shows the demographics data collected from the participants.

Criteria	Percentage (%)
Gender	
Female	43.15
Male	53.75
Others	0.12
No answer	2.98
CS/IT Education	
No	36.44
Yes	60.01
No answer	3.55
Age Group	
21-30	73.44
31-40	20.08
41-50	2.83
Above 50	0.45
No answer	3.20
Educational Degree	
Undergraduate Degree	75.30
Master's Degree	20.12
Doctorate	0.22
Others	1.22
No answer	3.14

Table 2: Participant Demographics

4.2. Procedure

We organized the study as a three-step methodology. Initially the participants had to answer a pre-test (step 1). This was followed by the actual gameplay (step 2) and then the post-test (step 3). The participants accessed the game using their respective devices and those participants who completed all the game levels from beginning to end were included in our evaluation. Only the first successful attempt of completion was used in our data analysis, even though many participants returned to play the game more than once. We measured the attempt count by tracking the "participant id" of participants, which was assigned based on their hashed email addresses. Evaluation on users' password knowledge improvement was done by analyzing their responses to the pre- and post-tests, and the password

structures entered by them during the password creation stage.

Each test question covered a password heuristic. The pre- and post-test questions followed a similar format asking players to select the relatively weaker password between two given choices. We created the password choices by picking suitable passwords from leaked databases [1, 58] and minimally modifying them to be able to test a particular heuristic, similar to the method followed in [50]. For example, to test H13 (predictable position of uppercase character) we chose the password "*brooklyn*" from the leaked database [58], and created the password pair comprising of "*Brooklyn*" and "*brooklYn*", of which the former is weaker as the uppercase character is at a very predictable position. This method was extended to the password pairs of other questions as well. Participants were also asked to provide their confidence ratings for every answer. The survey questions are provided as appendix.

5. Performance Evaluation

We evaluated the impact of our game-based training on users' password creation strategies. We were also interested in the changes in users' knowledge levels when it came to password practices by measuring correct answers given for pre and post-tests. We tried to answer our initial research questions through the study.

5.1. Can a Game-based Training Teach Password Heuristics to Participants?

We measured the effects of the game by analyzing the improvement in correct answers and confidence levels of the participants. The results are detailed in the following sections.

5.1.1. Participants' Correct Answers

We analyzed the participants' pre and post-level test responses. The test questions prompted users to choose the weaker between two given passwords. Passwords were shown such that the pairs focused on one heuristic each, and the password with the absence of that particular heuristic was considered weaker in that pair.

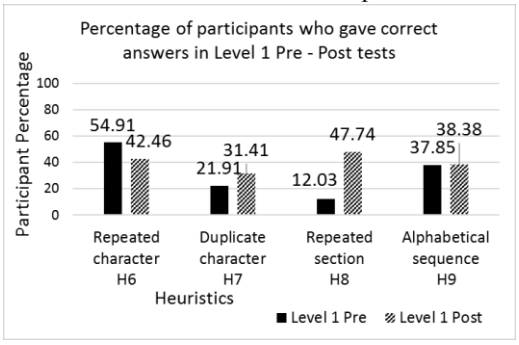


Figure 9: Level 1 - Participants' Correct Answers

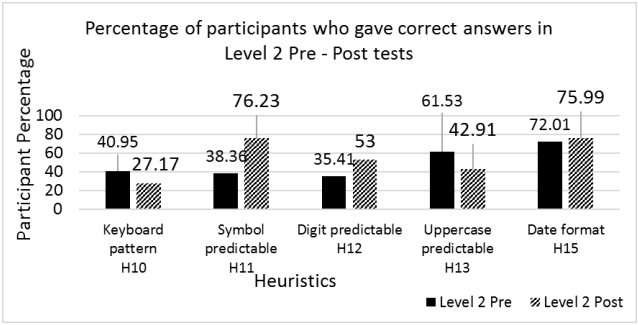


Figure 10: Level 2 - Participants' Correct Answers

The results show that more participants were able to provide correct answers to test questions after playing the game compared to before. There are however, exceptions for heuristics H6, H10, H13. The plots (cf. Figures 9, 10) show the participants' performance based on password heuristics in both the levels. H6 had a reduction of 12.45%, H10 had 13.78%, and H13 had 18.62% reduction in correct answers in post-level tests compared to pre-level tests. While we are unable to determine the cause for reduction in H6 now, the reduction in H10 could be attributed to the large possible combination of patterns possible.

The variety of keyboard patterns might turn out to be difficult for users to remember, which could mean that active feedback during password entry might be better for such heuristics. The predictable positions of uppercase characters (H13) had a decrease probably because the game time was not sufficient in unlearning the common practices of adding uppercase letters at the beginning or end. This may require multiple training sessions to unlearn. The users did show improvements in Level 2 while entering passwords by avoiding uppercase characters at the beginning. Questions on H1-H5 were not asked during the tests as these belonged to basic password requirements of our organization that the employees were aware of. Combining both levels, we asked the players 24 questions, 12 each in pre and post-level tests. The average number of correct answers increased from 5.96 (pre-test, *SD=2.3*) to 6.57 (post-test, *SD=2.69*). A statistically significant difference was observed with respect to the correct answers given by participants in the tests before and after the game (two-tailed paired t-test, $t(4905) = -19.35$, $p < .001$). Questions on H14 (predictable structures) were not asked separately in pre and post-tests as these would also fall under other heuristics. H14 was tested during password creation level; however, the passwords that satisfied the heuristics H1-H13 also satisfied H14. Averaging correct answer percentages of certain clusters (for both tests), we found an 8.32% increase in C2, 12.28% increase in C3, and 4.9% decrease in C4 (cf. Figure 11), showing that patterns

(especially keyboard patterns) is where participants need to gain more knowledge on.

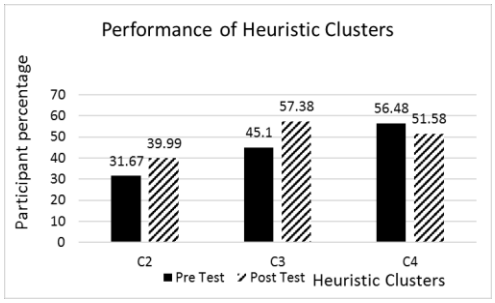


Figure 11: Performance of Heuristic Clusters

The previous study on GAP game [47] that tested password practices, similar to heuristics H11, H12, and H13, had shown better results with an average of 80.17%, 78.45%, and 87.93% participants correctly identifying password practices, H11, H12, and H13 respectively. Our game had a larger set of heuristics, with Level 1 requiring users to check ten heuristics and Level 2 having an additional six heuristics apart from Level 1 heuristics. This might have caused an overload of learning content in one go, as participants have shown better results for H13 in GAP [47] compared to Passworld. This could also show that while learning these heuristics is important, the manner to train users in them could be gradual. The users could be trained in an initial set of heuristics at first, and afterwards, the next set of training could be undergone.

5.1.2. Participants' Confidence Results

Participants were asked to rate their confidence level for their answers in both the tests. Each question had five levels of confidence (1: Not confident at all (Least Confident), 2: Not confident, 3: Neutral, 4: Confident, 5: Very confident). From Figures 12 and 13, we can see that the confidence level of players had a consistent increase. The average confidence rating increased from 4.39 (pre-test, variance =0.45) to 4.47 (post-test, variance =0.44). The analysis of the confidence levels shows the results to be statistically significant (Wilcoxon signed-rank test, $z = -18.87$, $p < .0001$).

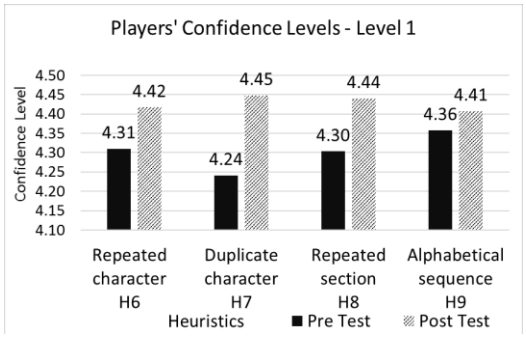


Figure 12: Level 1 - Participants' Confidence Levels

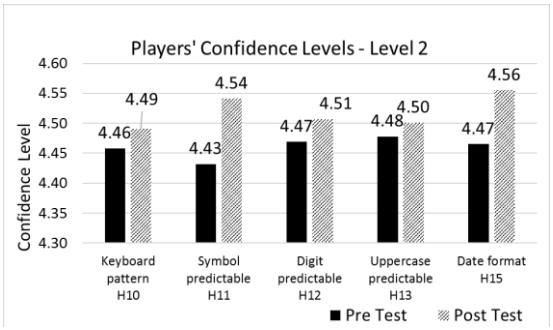


Figure 13: Level 2 - Participants' Confidence Levels

Considering confidence levels of clusters, average confidence level of C2 increased from 4.30 to 4.42, C3 increased from 4.45 to 4.51, and C3 increased from 4.46 to 4.52 (cf. Figure 14).

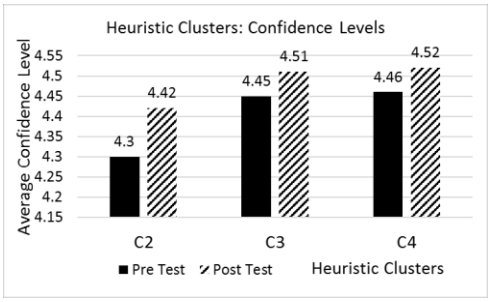


Figure 14: Confidence Levels of Heuristic Clusters

5.1.3 Demographic Analysis

Among the participants, the age group 21-30, and Bachelor's Degree holders showed a consistent increase in almost all heuristics. People with non-IT background (H9: 20% increase, H11: 104%) showed more improvements than the participants with IT background (H9: 5% increase, H11: 96%). H8 has better improvements in all demographics (with highest for Master's degree holders, 373% from pre to post). We also conducted a comparative study on gender differences vs. password practices. Petrie, et Al [36], also suggested a similar study. From our analysis, it was found that women did a better job of recalling created passwords, with 78.05% (Level 1) and 86.06% (Level 2) recalls matching their created passwords compared to men with 72.65% (Level 1) and 81.98% (Level 2). This supports the behavior reported in the study [36] where men expressed greater difficulty in remembering passwords.

5.1.4. Failure Count of Heuristics

Often, participants took several attempts before creating a password that satisfied each Level's heuristics. Counting all such attempts, for H1-H9 and H16, we calculated the overall number of times the participants failed for each heuristic before satisfying them. These heuristics appear in both levels 1 and 2. From Figure 15, we can see that the total number of failed attempts for each heuristic has come

down drastically from Level 1(*M= 283*) to Level 2 (*M= 106*). We can infer that once participants gained sufficient knowledge through game-play on various common password practices, they incorporated these learnings during password creation, thereby satisfying the heuristic checks. The number of failed attempts for blacklisted passwords decreased from 845 in Level 1 to 195 in Level 2 (~77% reduction), showing improvement. Alphabetical sequences (H9) however, showed a slight increase in failure rates. This could be attributed to the fact that there are a large number of possible sequences to avoid, and the participants were having difficulty in identifying them.

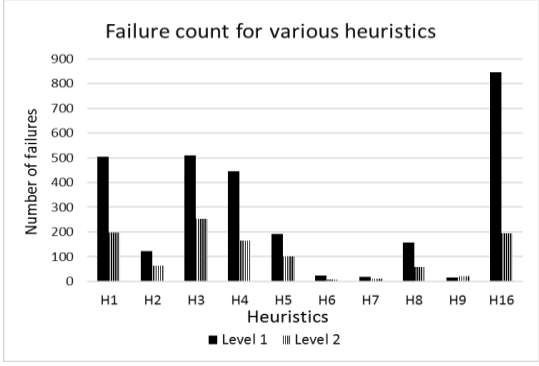


Figure 15: Heuristics vs Total Failures by Participants

5.2. Can Such a Training on Heuristics Improve the Participants' Password Diversity?

From the user responses, we found changes in password distribution when more heuristics were satisfied. We analyzed this using password structures. An interesting observation was about the use of common password structures. In the password creation stages, several participants followed similar structures while creating passwords (Table 3, Column 1). This shows that most of the players' ideas on secure passwords might end up being wrong, as shown in a previous study [49].

5.2.1. Level-wise Differences in Password Structures

Comparing the top 10 most popular password structures from Level 1 and Level 2, we were able to identify a major change in the number of occurrences of similar structures. Table 3 shows the comparison, where structures with common patterns (ULLLLSDDDD, for instance could be Abcde@1234) have shown drastic decrease in the number of occurrences. Level 2 structures are more spread out and with different password patterns, which suggest that participants created diverse passwords while satisfying more heuristics. It also brought down many common password structures, like those with common numerical sequences DDDD (e.g. 1234), uppercase characters (U) only in the beginning, etc.

Popular Password Structures with their Occurrences

Level 1	Level 2
ULLLLSDDDD(84)	LSLLLDLU(52)
ULLLSDDDD(64)	LSLLLDLLU(20)
ULLSDDDD(56)	LLLUSDLLL(10)
ULLLLLSDDDD(52)	ULLSULDDLL(8)
ULLLLLSDD(46)	LLLUSDDDUU(8)
LSLLLDLU(45)	UUUUSDDLL(6)
ULLLSDDD(44)	USLSUDLU(6)
ULLLLLLSDD(39)	SLLDLLLU(6)
ULLLLSDD(37)	LULLLSULLDDDL(5)
ULLLLSDDD(35)	ULSULDDDDL(5)

Table 3: Differences in occurrences of popular password structures (and their occurrences) from Level 1 compared to Level 2, after the initial analysis.

In Level 1, majority of the passwords start with an uppercase character, and the character does not appear anywhere else in the password. The game teaches this practice to be less secure, as one of the heuristics (H13) pointed out that use of uppercase in the beginning is a common practice. In the process of trying to satisfy more password heuristics into their structures, this trend was reduced. From the participant data, we found that the participants created 17,319 passwords that fell in 11,286 different structures. Considering only the password structures that satisfied the heuristics (4,906 in each level), there were 3,595 and 4,451 different structures respectively in Level 1 and Level 2.

Furthermore, 3,246 (66.16%) and 4,182 (85.24%) password structures were unique in Level 1 and Level 2 respectively, having only one occurrence. This trend showed that when more heuristics were to be satisfied in passwords, enterprise users could create passwords that were more diverse.

5.3. Other Game Data

For the password recall stage, data shows that 55.54% players matched the passwords in their first attempt for Level 1 and it increased to 71.70% for Level 2. Considering the players who recalled their passwords, in multiple attempts, the numbers come to 3,700 (75.41%) for Level 1 and 4,122 (84.01%) for Level 2. The participants who successfully recalled passwords for both levels comes to 3247(66.18%).

5.3.1. Player Involvement

Overall, 6,814 participants showed interest in playing the game, of which, 4,906 completed it. A decrease of 28%

could be because of network and proxy issues in certain locations, as noted from the participant feedback comments. Among the 4,906 participants, we calculated the player involvement by measuring number of resources and artifacts collected per level by the players (Table 4). The data shows that majority of players have had an immersive gameplay, with increasing number of collected artifacts and resources nearing the total number available. Here L, U, D, S correspond to character classes and A denotes the artifact collected.

		L	U	D	S	A
Level 1	Average	5.37	7.05	6.7	3.64	0.72
	Total	6	9	9	4	1
Level 2	Average	11.4	7.48	6.54	11.05	0.83
	Total	12	8	7	12	1

Table 4: Gameplay Resources collection data

5.4. Player Feedback

The general feedback by participants included comments like "informative and interesting", "Excellent game to deliver the message on the usage of strong passwords", "Fun and creative game. Nice idea." etc. While some participants requested for the game to be available as a permanent training method, some participants found the game time of over 15 minutes to be a bit longer. At the end of the post-test, we asked the participants to rate Passworld on a 5-point Likert scale [22] with respect to fun, education, and learning. As per the feedback data (cf. Figure 16), 93.50% participants agreed the game to be fun (*M = 4.42*, *SD = 0.69*), 93.85% (*M= 4.42, SD= 0.68*) found the game to be educational and 94.24% (*M= 4.48, SD = 0.68*) considered they have learned about secure password practices.

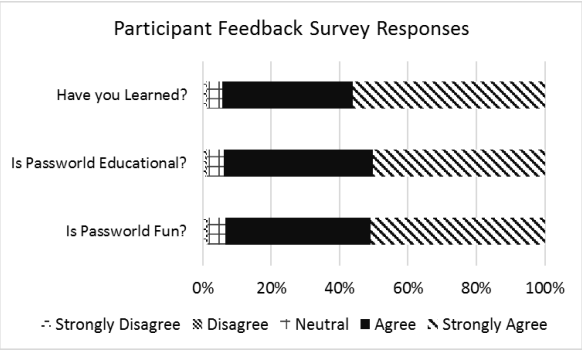


Figure 16: Participants' Feedback Data

6. Discussion

The Passworld game was successfully completed by 4,906 employees in our organization. The analysis of their gameplay data found the game to be effective and engaging. It also helped us find various insights on enterprise users' password practices. The insights are as follows.

a) Participants showed improvements in creating passwords that satisfied various heuristics like avoiding date formats, predictable positions of symbols and digits, character duplicates to name a few. Their new passwords satisfied more number of heuristics after users played the game levels, compared to before.

b) Prior to satisfying all game heuristics, participants created similar password structures, as seen from Column 1 of Table 3. After satisfying more heuristics, the participants were able to create more diverse passwords. Over 90% of passwords created after satisfying all the game heuristics resulted in unique password structures. This trend, when followed outside the game, would result in password diversity across organization.

c) It is to be noted that the diversity in user-created password structures post gameplay is not brought about by providing suggestions on modifications to the entered password, unlike in the previous study [38]. More likely, the users applied their learnings obtained from the game to satisfy more heuristics in their passwords. This resulted in more password diversity.

d) Comparing with the previous game GAP [47] which showed good results with a small set of heuristics, we found that certain heuristics like repeated characters (H6), keyboard patterns (H10), and predictable uppercase characters (H13) showed decrease in performance after game-play. H10 can be considered as a complex heuristic with several possible patterns that the user has to know of. H13 is such that addition of uppercase characters at the beginning of a password is a very common practice that might need multiple trainings to unlearn. This could suggest that these heuristics might either need a more in-depth training or an active feedback while passwords are being entered in real time. The results could also mean that training on a very large set of heuristics might have caused a learning overload that might have resulted in decrease in results in Level 2. This could be avoided by providing a gradual training on heuristics.

e) Considering the heuristic clusters, C4 had a decrease of around 5% showing that "identifying patterns" is where the participants struggled the most. Real-time suggestion to avoid keyboard patterns during password entry could be a method to reduce the users from including patterns into their passwords.

f) We received certain insights on changes to gameplay that include reducing the overall game length, reducing the time needed for animals to attack the gate, and distribution of heuristics into various levels.

6.1 Limitations

Our study has a certain number of limitations that have to be considered while analyzing the results. The participant demographics selected for the game included employees of our organization, who may know basics of cybersecurity or password practices. The game data of 4,906 participants shows positive responses in understanding password heuristics and creation strategies. However, to test the effectiveness of the game in a real-world scenario, we will have to monitor individual password practices throughout the organization to start with. This process could violate the privacy of individuals, and therefore has not been attempted by us. The results of a previous research [28] shows that the positive results of a smaller lab experiment were carried forward to a larger audience. We hope that our experimental results show a similar trend. Similarly, the passwords entered by the users within the game were most likely not their real passwords used in any accounts. While a research study [16] suggests that the passwords implemented by users in a study could resemble real life passwords, the extent to which our participants would utilize similar passwords in both game and real life needs to be studied. An alternate form of experiment, e.g. a text-based condition or another control condition, has not been tested. While the study makes no claims of being better than a controlled experiment, our training shows that a fun oriented gameplay can help teach password heuristics to users. More studies need to be conducted in order to obtain conclusive evidence on the effectiveness of a game as a medium to train enterprise users on password awareness.

The game makes no claims about making all the user passwords memorable. It attempts to improve the understanding of people in creating stronger passwords than what they used to create prior to the game. While we observed reasonable password recall rates in our study, we do not have conclusive evidence on password memorability over long periods.

Passworld tries to offer a starting point in the area of game-based password security education. A previous research study [39] show that user behavior related to password usage can be influenced with positive reinforcement. The results of our study suggest that game-based training could also influence users' behaviors related to passwords. A solid understanding on password practices coupled with the use of a password meter could provide better security in terms of password strength.

7. Conclusions

The Passworld game was designed to provide awareness on various password heuristics to enterprise users. The main objectives of our study were to find 1) if a game-based training could teach users on password heuristics 2) if such a training on heuristics could improve organizational password diversity. We used the password heuristics from a previous study [48] for teaching, and we checked if the users satisfied every one of these heuristics during their password creation. Our intention was different from the previous study [48], where satisfying all the heuristics was not mandatory. We intended to teach the users about the importance of each heuristic, and wanted to see how many users successfully implemented what they learnt. We presented the results from our enterprise study with 4,906 participants. Even though our study was a standalone study, without a control condition, we found that after playing the game, the correctness and confidence levels of the participants have increased. The password structures created by the participants have shown more diversity post gameplay. This, along with the positive feedback, shows that the gameplay has helped the participants learn the concepts to implement diverse and memorable passwords. We believe that this trend, when followed in real life, would result in organizational password diversity.

We recommend the launch of such training methods in an organizational environment to ensure that users learn about password heuristics and incorporate them while creating passwords to promote diversity in password structures. This could be a deciding factor when it comes to organizational password security.

Recommendations for further study follow. First, the study could be carried out on a set of wider demographics, with different levels of understanding of security concepts and learning backgrounds. To reduce the information overload, we propose a gradual learning with one set of heuristics, followed by another set. Training on patterns (like alphabetic sequences, keyboard patterns etc.) could be done separately to lay emphasis on it. A methodology to evaluate users' password memorability over long periods could also be beneficial in proposing further learning goals. We aim to explore further areas of password and cybersecurity education through interactive gameplay experiences.

References

[1] A. Vance. 2010. "If your password is 123456, just make it hackme," New York Times, Retrieved February 19, 2020 from https://www.nytimes.com/2010/01/21/technology/21password.html.

[2] Adams, A., and Sasse, M.A. User are not the enemy. Commun. ACM 42, 12 (Dec. 1999), 40–46

[3] Aloul, Fadi, Syed Zahidi,l Wassim El-Hajj. "Two factor authentication using mobile phones." 2009 IEEE/ACS International Conference on Computer Systems and Applications. IEEE, 2009.

[4] Annetta, Leonard A. "The "I's" have it: A framework for serious educational game design." Review of General Psychology 14.2 (2010): 105.

[5] Bellotti, F., Ott, M., Arnab, S., Berta, R., de Freitas, S., Kiili, K. and De Gloria, A., 2011, October. Designing serious games for education: from pedagogical principles to game mechanisms. In Proceedings of the 5th European Conference on Games Based Learning. University of Athens, Greece (pp. 26-34).

[6] Buttyán, Levente. "Introduction to IT security." 2016.

[7] Cheung, Wing-Yee, et al. "Back to the future: Nostalgia increases optimism." Personality and Social Psychology Bulletin 39.11 (2013): 1484-1496.

[8] CJ, G., Pandit, S., Vaddepalli, S., Tupsamudre, H., Banahatti, V., & Lodha, S. (2018, October). Phishy-a serious game to train enterprise users on phishing awareness. In Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts (pp. 169-181).

[9] Compton, Kate, and Michael Mateas. "Procedural Level Design for Platform Games." AIIDE. 2006.

[10] Cone, Benjamin D., Cynthia E. Irvine, Michael F. Thompson, and Thuy D. Nguyen. "A video game for cyber security training and awareness." computers & security 26, no. 1 (2007): 63-72.

[11] Csikszentmihalyi, Mihaly, Sami Abuhamdeh, and Jeanne Nakamura. "Flow." Flow and the foundations of positive psychology. Springer, Dordrecht, 2014. 227-238.

[12] Daitch, Heidi. 2017. 2017 Data Breaches- The Worst So Far. Retrieved February 19, 2020 from https://www.identityforce.com/blog/2017-data-breaches

[13] De Carnavalet, Xavier De Carné, and Mohammad Mannan. "From Very Weak to Very Strong: Analyzing Password-Strength Meters." NDSS.Vol. 14. 2014.

[14] Denning, Tamara, Adam Lerner, Adam Shostack, and Tadayoshi Kohno. "Control-Alt-Hack: the design and evaluation of a card game for computer security awareness and education." In Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, pp. 915-928. ACM, 2013.

[15] Donovan, M. Suzanne, John D. Bransford, and James W. Pellegrino. How people learn: Bridging research and practice. National Academy Press, 2101 Constitution Avenue NW, Lockbox 285, Washington, DC 20055, 1999.

[16] Fahl, S., Harbach, M., Acar, Y., and Smith, M. On the ecological validity of a password study. In Proc. SOUPS (2013)

[17] GameDev.Net.2017. Intrinsic and Extrinsic Motivation. Retrieved May 29, 2020 from https://www.gamedev.net/tutorials/game-design/game-design-and-theory/intrinsic-and-extrinsic-motivation-r4713/

[18] Giannakas, Filippos, Georgios Kambourakis, and Stefanos Gritzalis. "Cyberaware: A mobile game-based app for cybersecurity education and awareness." Interactive Mobile Communication Technologies and Learning (IMCL), 2015 International Conference on. IEEE. 2015.

[19] Gordon, William J., Adam Fairhall, and Adam Landman. "Threats to Information Security—Public Health Implications." New England Journal of Medicine 377.8 (2017): 707-709.

[20] Habib, Hana, Pardis Emami Naeini, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. "User behaviors and attitudes under password expiration policies." In Fourteenth Symposium on Usable Privacy and Security ({SOUPS} 2018), pp. 13-30. 2018.

[21] Hudson Soft.1986.Adventure Island. Game [MSX, NES] (September 12, 1986) Hudson Soft., Tokyo, Japan.

[22] I. Elaine Allen and Christopher A. Seaman. 2007. Likert Scales and Data Analyses. July 2007. Retrieved February 19, 2020 from http://rube.asq.org/quality-progress/2007/07/statistics/likert-scales-and-data-analyses.html

[23] Ives, Blake, Kenneth R. Walsh, and Helmut Schneider. "The domino effect of password reuse." Communications of the ACM 47.4 (2004): 75-78.

[24] Jay Jay. 2018. Despite increased cyber-risk awareness, poor password hygiene still rules. Retrieved February 19, 2020 from https://www.scmagazineuk.com/despite-increased-cyber-risk-awareness-poor-password-hygiene-rules/article/1472772

[25] Johnson, B. R., and Koedinger, K. R. 2002. Comparing instructional strategies for integrating conceptual and procedural knowledge. In Proceedings of the Annual Meeting [of the] North American Chapter of the International Group for the Psychology of Mathematics Education, vol. 1–4, pp. 969–978.

[26] Kiili, Kristian. "Digital game-based learning: Towards an experiential gaming model." Internet and Higher Education 8 (2005): 13-24.

[27] Krathwohl, David R. "A revision of Bloom's taxonomy: An overview." Theory into practice 41.4 (2002): 212-218.

[28] Kumaraguru, Ponnurangam, et al. "Teaching Johnny not to fall for phish." ACM Transactions on Internet Technology (TOIT) 10.2 (2010): 1-31.

[29] Kumaraguru, Ponnurangam, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. "Protecting people from phishing: the design and evaluation of an embedded training email system." In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 905-914. ACM, 2007.

[30] Lee Mathews. 2017. File With 1.4 Billion Hacked And Leaked Passwords Found On The Dark Web. Retrieved February 19, 2020 from https://www.forbes.com/sites/leemathews/2017/12/11/billion-hacked-passwords-dark-web/#516d814221f2

[31] Liu, Ye, and Xiaolan Fu. "How does distraction task influence the interaction of working memory and long-term memory?." International Conference on Engineering Psychology and Cognitive Ergonomics. Springer, Berlin, Heidelberg, 2007.

[32] Moreno R, Mayer RE, Spires HA, Lester JC. The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? Cognition and instruction. 2001 Jun 1;19(2):177-213

[33] New York Film Academy.2015. Learning From The Best: Action-Adventure Games. Retrieved February 19,2020 from https://www.nyfa.edu/student-resources/learning-from-the-best-action-adventure-games/

[34] Nintendo.1985. Super Mario Bros. Game [NES] (September 13, 1985) Nintendo, Kyoto, Japan. Level/area: World 8-4

[35] Ofcom. Adults media use and attitudes. Technical report, Ofcom: UK Communications Regulator, 2016. https://www.ofcom.org.uk/__data/assets/pdf_file/0026/80828/2016-adults-media-use-and-attitudes.pdf (accessed January 2020).

[36] Petrie, Helen, and Burak Merdenyan. "Cultural and Gender Differences in Password Behaviors: Evidence from China, Turkey and the UK." Proceedings of the 9th Nordic Conference on Human-Computer Interaction. ACM, 2016.

[37] Rass, Stefan, and Sandra König. "Password Security as a Game of Entropies." Entropy 20.5 (2018): 312.

[38] Segreti, Sean M., William Melicher, Saranga Komanduri, Darya Melicher, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L. Mazurek. "Diversify to survive: making passwords stronger with adaptive policies." In Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017), pp. 1-12. 2017.

[39] Seitz, Tobias, and Heinrich Hussmann. "PASDJO: quantifying password strength perceptions with an online game." Proceedings of the 29th Australian Conference on Computer-Human Interaction. ACM, 2017.

[40] Selvarajan, Balamurugan. "Simple two-factor authentication." U.S. Patent Application No. 11/267,148.

[41] Sheng, Steve, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. "Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish." In Proceedings of the 3rd symposium on Usable privacy and security, pp. 88-99. ACM, 2007.

[42] Sherry, John, and Angela Pacheco. "Matching computer game genres to educational outcomes." Electronic Journal of Communication/La Revue Electronique de Communication16.1& (2006): 2.

[43] Sweetser, P., & Wyeth,P.(2005). GameFlow: a model for evaluating player enjoyment in games. Computers in Entertainment (CIE), 3(3), 3-3.

[44] Thompson, Michael, and Cynthia Irvine. "Active learning with the CyberCIEGE video game." (2011)

[45] Bradley, Tony. 2018. Security Experts Weigh In On Massive Data Breach Of 150 Million MyFitnessPal.Accounts Retrieved February 19, 2020 from https://www.forbes.com/sites/tonybradley/2018/03/30/security-experts-weigh-in-on-massive-data-breach-of-150-million-myfitnesspal-accounts/#209ddebd3bba

[46] Hunt, Troy. 2019. The 773 Million Record "Collection #1" Data Breach. Retrieved February 19, 2020 from https://www.troyhunt.com/the-773-million-record-collection-1-data-reach/

[47] Tupsamudre, Harshal, Rahul Wasnik, Shubhankar Biswas, Sankalp Pandit, Sukanya Vaddepalli, Aishwarya Shinde, C. J. Gokul, Vijayanand Banahatti, and Sachin Lodha. "GAP: A Game for Improving Awareness About Passwords." In Joint International Conference on Serious Games, pp. 66-78. Springer, Cham, 2018.

[48] Ur, Blase, Felicia Alfieri, Maung Aung, Lujo Bauer, Nicolas Christin, Jessica Colnago, Lorrie Faith Cranor et al. "Design and evaluation of a data-driven password meter." In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 3775-3786. ACM, 2017.

[49] Ur, Blase, Fumiko Noma, Jonathan Bees, Sean M. Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. "" I Added'!'at the End to Make

It Secure": Observing Password Creation in the Lab." In Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015), pp. 123-140. 2015.

[50] Ur, Blase, Jonathan Bees, Sean M. Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. "Do users' perceptions of password security match reality?." In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 3748-3760. ACM, 2016.

[51] Ur, Blase, Patrick Gage Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L. Mazurek, Timothy Passaro et al. "How does your password measure up? the effect of strength meters on password creation." In Presented as part of the 21st {USENIX} Security Symposium ({USENIX} Security 12), pp. 65-80. 2012.

[52] Verizon Enterprise Solutions. 2016. 2016 Data Breach Investigations Report Retrieved February 19, 2020 from https://enterprise.verizon.com/resources/reports/2016/dbir-2016-executive-summary.pdf

[53] Virvou, Maria, George Katsionis, and Konstantinos Manos. "Combining software games with education: Evaluation of its educational effectiveness." Journal of Educational Technology & Society 8.2 (2005).

[54] Vu, Kim-Phuong L., Robert W. Proctor, Abhilasha Bhargav-Spantzel, Bik-Lam Belin Tai, Joshua Cook, and E. Eugene Schultz. "Improving password security and memorability to protect personal and organizational information." International Journal of Human-Computer Studies 65, no. 8 (2007): 744-757.

[55] Weirich, Dirk, and Martina Angela Sasse. "Pretty good persuasion: a first step towards effective password security in the real world." Proceedings of the 2001 workshop on New security paradigms. ACM.

[56] Yan, Jianxin Jeff. "A note on proactive password checking." Proceedings of the 2001 workshop on New security paradigms. ACM, 2001.

[57] Zviran, Moshe, and William J. Haga. "Password security: an empirical study." Journal of Management Information Systems15.4 (1999): 161-185.

[58] Hunt, T. (2019). Have I Been Pwned. Retrieved from https://haveibeenpwned.com/Passwords

Appendices

A. Demographics Survey

We conducted an initial demographic survey of our participants. The questions asked are as follows:

a. Do you have background in CS/IT/Security?

a. Yes b. No

b. **What is the highest level of education you have completed?**

a. Bachelor's Degree

b. Master's Degree

c. Doctorate

d. Other

c. **What is your Gender**

a. Female

b. Male

c. Other

d. **How old are you?**

a. 21-30

b. 31-40

c. 41-50

d. Above 50

e. **Select your Nationality** (from the list of nations).

B. Game Tests

The questions asked during the pre- and post-tests were focused on the heuristics being taught in the game. Every question contained two password choices, with one satisfying the particular heuristic, while the other failed to do so. We created the password choices by picking suitable passwords from leaked databases [1, 58] and minimally modifying them to be able to test a particular heuristic (similar to the method followed in [50]). For example, we chose a leaked password "*Password*", (leaked 130,999 times as per [58]), and incorporated repeated characters (H6) by adding four consecutive "s" to create "*Passssword*". This password did not satisfy H6. The corresponding alternate choice for this password was

"*Paswwoorsdd*", which had a maximum of two repeated characters. The latter satisfied the requirements for the heuristic H6. In this pair, the former is considered the weaker password. Similarly, we chose other leaked passwords and incorporated the same methodology to create a list of password pairs for use in the pre- and post-tests.

The passwords given as choices were judged based on the particular heuristic being dominant. For example, considering passwords "*asdfghjkl*" and "*afhsgdljk*", despite having similar characteristics, the former is a very common **keyboard pattern** and the latter is not. This puts the former in "common passwords" criteria, making it weaker. Both the options provided were of equal length and had similar characteristics like order of character classes etc., except the heuristic being checked. Presence of multiple heuristics within a password would have caused ambiguity to the participants. Apart from judging based on heuristics, we also analyzed the password strengths using several available tools to ensure that our choice of answers are true to the maximum extend. Following online services were used to rate the passwords before being added to the pre and post-tests:

My1Login.2020. Take the Password Test. Retrieved June 02, 2020 from https://www.my1login.com/resources/password-strength-test/

Kaspersky.2020.Kaspersky Secure Password Check. Retrieved June 02, 2020 from https://pasword.kaspersky.com/in/

The level-wise test questions and the heuristic being checked (*in italics*) are provided below:

Which of the following passwords is weaker?

Level 1 Pre-Test

1. a) **Passsssword** *(Repeated characters, H6)*
   b) Paswwoorsdd
   c) Both are identical

2. a) welccoommlee
   b) **weeleecomeee** *(Duplicate Character, H7)*
   c) Both are identical

3. a) **Passtuvw@12rd** *(Alphabetic Sequence, H9)*
   b) Pstasuwv@12rd
   c) Both are identical

4. a) ac2ab1c12bacbca
   b) **abc12abc12abcabc** *(Repeated Section, H8)*
   c) Both are identical

Level 1 Post-Test

1. a) **seeseameee** *(Duplicate Character, H7)*
   b) seasamemesa
   c) Both are identical

2. a) **tcstcstcsadmintcs** *(Repeated Section, H8)*
   b) sctatdmcicnstntcs
   c) Both are identical

3. a) letmmelein
   b) **letmmmmmin** *(Repeated Character, H6)*
   c) Both are identical

4. a) **Nutriopqrst** *(Alphabetic Sequence, H9)*
   b) Nutriposqtr
   c) Both are identical

Level 2 Pre-Test

1. a) jo14n21ny
   b) **jonny1421** *(Digit Predictable, H12)*
   c) Both are identical

2. a) **Brooklyn** *(Uppercase predictable, H13)*
   b) brooklYn
   c) Both are identical

3. a) **RockWell@789** *(Symbol Predictable, H11)*
   b) R@ockwell789
   c) Both are identical

4. a) tomhjklinmpo

   b) **tomhijklmnop** *(Alphabetic Sequence, H9)*

   c) Both are identical

5. a) **qwertyuiop** *(Keyboard Pattern, H10)*

   b) qtrwyeuiop

   c) Both are identical

6. a) **pass@July2017** *(Date Format, H15)*

   b) pass@Ju1y2@17

   c) Both are identical

Level 2 Post-Test

1. a) **asdfghjkl** *(Keyboard Pattern, H10)*

   b) afhsgdljk

   c) Both are identical

2. a) shad7w00o

   b) **shadow007** *(Digit Predictable, H12)*

   c) Both are identical

3. a) **Me@Aug95** *(Date Format, H15)*

   b) Me@Au9g5

   c) Both are identical

4. a) **secure@234** *(Symbol Predictable, H11)*

   b) sec@23ure4

   c) Both are identical

5. a) **Can@da** *(Uppercase predictable, H13)*

   b) caN@da

   c) Both are identical

6. a) sutjerrywv

   b) **stuvwjerry** *(Alphabetic Sequence, H9)*

   c) Both are identical

For each of these test questions, there was another question asking the player's confidence level for their responses:

How confident are you about your selection:

a. Not at all confident

b. Not very confident

c. Neutral

d. Confident

e. Very confident

C. Game Feedback Survey

Please answer all these questions:

1) The game was fun

   a. Strongly agree

   b. Agree

   c. Neutral

   d. Disagree

   e. Strongly Disagree

2) The game was educational

   a. Strongly agree

   b. Agree

   c. Neutral

   d. Disagree

   e. Strongly Disagree

3) I learned how to create a secure password

   a. Strongly agree

   b. Agree

   c. Neutral

   d. Disagree

   e. Strongly Disagree

D. Information Regarding our Data and RockYou database:

In the game, the participants have created passwords that satisfy all the game heuristics in Level 2. These can be considered as passwords with lesser vulnerabilities.

Table A shows the list of top 10 structures from Level 2 and its corresponding occurrence in the RockYou database of over 32 million leaked passwords [1].

RockYou database consists of leaked passwords from various sources. The password structures that cleared Level 2 were not found in the leaked database. The most common passwords from RockYou database included structures with very little heuristics incorporated, like LLLLLL (12.23%), LLLLLLL (8.35%), LLLLLLLL (7.57%), DDDDDD (6.98%) etc. This shows that the leaked passwords possibly satisfied fewer heuristics than Level 2 passwords.

Password Structures (Level 2)	Occurrences (%)	
	Game Data	RockYou Data
LSLLLDLU	0.300	0.00003
LSLLLDLLU	0.115	0
LLLUSDLLL	0.057	0.000003
ULLSULDDLL	0.046	0
LLLUSDDDUU	0.046	0
UUUUSDDLL	0.035	0
USLSUDLU	0.035	0
SLLDLLLU	0.035	0.000003
LULLLSULLDDDL	0.029	0
ULSULDDDDL	0.029	0

Table A: Top 10 recurring password structures from the study vs. RockYou data

E. Animals in Game used to Teach and Check Password Heuristics

The game teaches password heuristics using oncoming animals, who provide heuristic-related information to the players, at the same time, check if heuristics are satisfied during password creation. Given below are the various animals that come into play. Their corresponding heuristics are mentioned in Table 1.

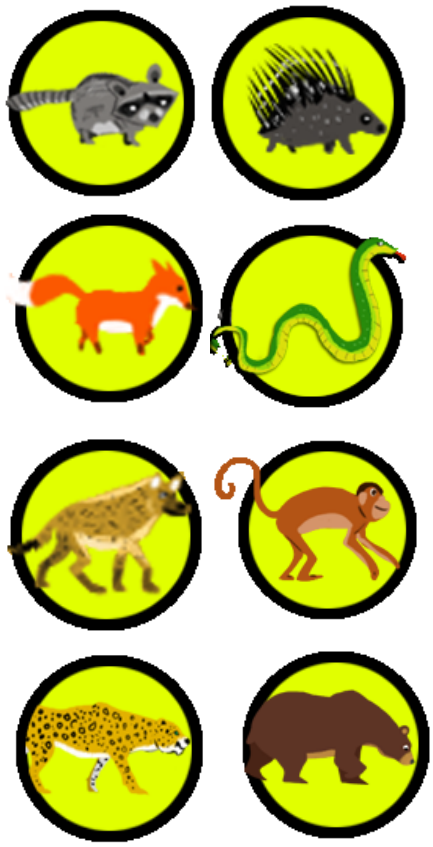


Figure A: The animals appearing in the game are (starting from top left) Raccoon, Porcupine, Fox, Snake, Hyena, Monkey, Leopard, and Bear.

"You still use the password after all" – Exploring FIDO2 Security Keys in a Small Company

Florian M. Farke
Ruhr University Bochum

Lennart Lorenz
tracekey solutions GmbH

Theodor Schnitzler
Ruhr University Bochum

Philipp Markert
Ruhr University Bochum

Markus Dürmuth
Ruhr University Bochum

Abstract

The goal of the FIDO2 project is to provide secure and usable alternatives to password-based authentication on the Web. It relies on public-key credentials, which a user can provide via security tokens, biometrics, knowledge-based factors, or combinations. In this work, we report the results of a qualitative study accompanying the deployment of FIDO2-enabled security tokens for primary authentication in a web application of a small software company operating in the life sciences industry. We assisted the company in implementing and setting up FIDO2-enabled authentication on its public test and evaluation server. Over four weeks, we observed the authentication routine of 8 employees out of 10 employees regularly using the web application, including sales representatives, software developers, project managers, and account managers. We gathered data through login diaries, server logs, and semi-structured interviews to assess themes regarding usability, perceived security, and deployability. We found that participants had several concerns, like losing the security token and longer authentication times, while the security benefits were largely intangible or perceived as unnecessary.

1 Introduction

User authentication by username and password is still the most dominant method on the Internet and remote authentication in general. However, password-based authentication has many usability and security flaws, and researchers and practitioners have been discouraging from using it for decades [5]. On the Internet, passwords are prone to phishing attacks. Phishing attacks become more and more sophisticated, leading to often transparent and nearly indistinguishable imitations of valid authentication requests [3, 27].

Many alternatives have been proposed, but their usage is minimal [5]. Biometric schemes such as fingerprint or face recognition are regularly used to unlock phones, but they are not used for remote authentication. Authentication with hardware tokens, typically in the form of two-factor authentication (2FA) combined with a knowledge-based scheme like passwords, provides high security, but distributing and managing the hardware keys can become a great hurdle. 2FA using possession-based factors such as smartphone apps or SMS tokens as a second factor is less secure, yet easier to set up and manage, but has only found limited adoption (e.g., less than 10 % of active Google accounts use 2FA [18]).

The FIDO2 project, including both the Fast IDentity Online (FIDO) Alliance [13] – an industry association – and the World Wide Web Consortium (W3C) [29], aims at offering an alternative to password-based authentication that is both usable and secure. It consists of two main components: the *Client to Authenticator Protocol 2 (CTAP2)*, governing the communication between the client and (external) authentication hardware, and the *Web Authentication (WebAuthn)* specification defining the server-facing API on the client. WebAuthn became an official web standard in March 2019, and several browsers (e.g., Google Chrome, Mozilla Firefox, and Microsoft Edge) support it already.

FIDO2 promises a largely improved authentication experience and is backed by several big companies, like the *Alibaba Group*, *Amazon*, *Apple*, *Facebook*, and *Google*. Thus, it is very interesting to understand the likely impact it can have in practice, including aspects of deployment and usability.

In this work, we present our experience with deploying FIDO2 in the context of a company. We report on a four-week evaluation phase in which we accompanied the deployment in a life sciences company. Eight voluntary participants used a FIDO2-based authentication scheme on a daily basis and kept login diaries, which we combined with server logs, a survey, and semi-structured interviews after the four weeks.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

To use the FIDO2-based authentication method, we gave the participants security keys (USB-based hardware tokens) and guided them through the setup. The security keys could be used as a full-fledged alternative to username and password in one of the company's software products. We used security keys because they are relatively inexpensive and were compatible with all computers the participants used at work. In particular, we were interested in:

- How do users behave when they have a security key as an authentication alternative to username and password and not only as a second factor?
- Do users use the security key in their daily routine?
- What differences do users identify between the authentication schemes, especially do they perceive the new method as more secure?
- What advances or hinders the adoption of the new security key-based mechanism?

We learned that even though the participants liked using the security key-based authentication scheme, they tended to fall back to username and password. On the one hand, this is because participants do not want to abandon a habit when there is no apparent necessity in doing so. On the other hand, they fear to lock themselves out when losing or breaking the security key. Participants who use a password manager with an auto-fill feature of a web browser also report that the authentication with the security key takes longer in comparison. In contrast, the participants assumed that the keys were providing a better security level, although they did not fully comprehend the authentication procedure's technical details.

Our qualitative study is a first attempt to explore FIDO2 in a business environment and sheds light on which problems arise when deploying FIDO2. In summary, we make the following key contributions:

- We explore a *passwordless* web authentication scheme rolled out as the first authentication factor in a real-world application.
- We provide insights into the daily usage of security keys in a company environment over four weeks by combining login diaries, server logs, and interviews. The data indicate that participants liked the passwordless authentication scheme because of its simple usage, yet from the users' perspective, there is no clear advantage over password-based authentication. The positive impression of the security keys is less pronounced if participants previously used password managers that already limit some negative aspects of password-based authentication.
- Analyzing the participants' feedback, we identify a set of adoption barriers, including the fear of getting locked out, a cumbersome integration of the security keys in the work environment, and the general routine in using passwords. Those barriers should be minimized before introducing new security measures in work environments.

2 Background

The FIDO2 project has a more general and thus flexible approach to user authentication than its predecessor, the Universal 2nd Factor (U2F) standard, and other authentication schemes. Compared to other 2FA/multi-factor authentication (MFA) approaches, the advantages of FIDO2 are (i) growing support by all major browser and operating system vendors, (ii) open and standardized protocols, (iii) making authentication via username and password not mandatory (although it is still possible), and (iv) building upon vetted asymmetric cryptographic principles and algorithms.

The FIDO2 project consists of the WebAuthn specification of the W3C [4] and the Client to Authenticator Protocol (CTAP) defined by the FIDO Alliance [6]. FIDO2 allows abstracting from the actual authenticator (e.g., a hardware token). Thus, the *Relying Party* (e.g., a web server) does not require knowledge about the implementation details of the authenticator. Figure 1 depicts the interplay of CTAP2 and WebAuthn as we used it in our study.

WebAuthn specifies a standardized, browser-independent JavaScript API that allows web services to interact with all sorts of facilities. Through this API, web services can implement user authentication in a way that is resilient to phishing, password theft, and replay attacks. Instead of relying on shared secrets like passwords, public-key cryptography is used to create unique credentials for every web service and only generated and stored on the client's device.

On the other hand, CTAP2 governs the communication between external authenticators and web browsers or other applications supporting WebAuthn. The proposed CTAP standard comprises two protocol versions – CTAP1, the protocol used for U2F, and CTAP2 a new protocol used for WebAuthn. At the time of the study, implementations for the operating systems *Android* and *Windows 10* were available.

The FIDO Alliance uses the term "passwordless" to describe single-factor authentication and multi-factor authentication with an authenticator or with an authenticator and a personal identification number (PIN) or biometric. While it is easy to agree with, for example, hardware tokens as a single factor being passwordless, this is not inevitably the case if additionally a PIN is used as a second factor. Although similar to a password, a FIDO2 PIN has some notable differences compared to passwords in the context of web authentication:

- No shared secret has to be entered, sent over network, or store on the server-side. Phishing attempts and data breaches do not affect the authenticity of the credentials.
- A single PIN unlocks the authenticator and all account credentials registered with the device. There is no need to set a unique password for every web service.
- Guessing or brute-forcing the PIN is limited to eight consecutive attempts. Reaching this limit resets the authenticator to factory settings, effectively invalidating all generated credentials [6].

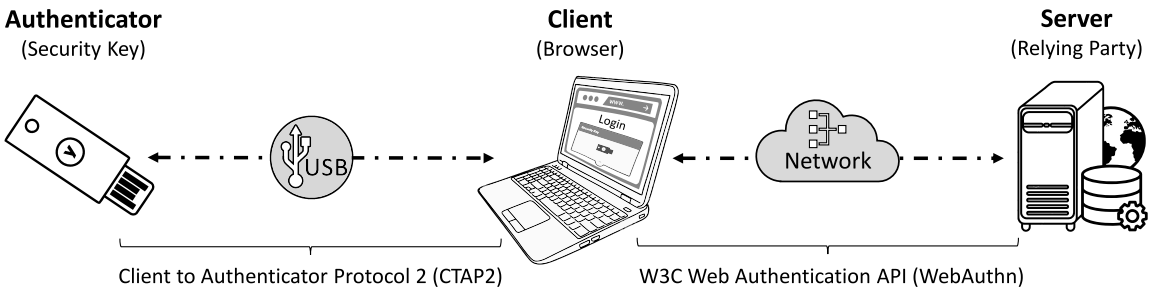


Figure 1: Communication via FIDO2. The cryptographic authenticator and the client communicate via USB using CTAP2. The client's browser implements the JavaScript-based Web Authentication API to communicate with the server over the network.

3 Methods

Our study's primary goal was to gain insights into the usability, user perception, and barriers or facilitators for the adoption of FIDO2. To increase ecological validity, we conducted the study on a web application in a small software company in the life sciences industry. We collected and analyzed qualitative data in the form of login diaries and semi-structured interviews, but also quantitative data in the form of server logs.

3.1 Study Environment

We conducted our study at *tracekey solutions GmbH*,[1] a small software company from Bochum (Germany). As a *software-as-a-service* provider, they develop and operate a product serialization service for small and medium-sized businesses in the pharmaceutical industry. The service offers a solution to fulfill the traceability requirements of this industry [12].

The service includes a web application that requires customers and employees to authenticate with a username and password. For our study, we extended the existing login form and authentication back end and added a new login option using WebAuthn with a roaming hardware token.

We decided only to allow PIN-protected security keys (in contrast to Lyastani et al. [17]) because some participants occasionally worked remotely, and losing the security key was a realistic scenario. However, this may have reduced the comfort of using the key (see Section 4.3). The company had an authentication policy for its software that required re-authentication after 30 minutes of inactivity. Due to this policy, we did not consider adopting a "remember me" option, which does not ask the user to authenticate for a particular time after a successful login on a device. We did not implement any fallback authentication procedure for the security key because the participants could still use their passwords, and manual account recovery was also a viable option.

[1] https://www.tracekey.com/, as of April 30, 2020

Following tracekey's three-week release cycles, the security key-based authentication method was developed on an internal test server and pushed to the public test and validation server before the study started. The test server was accessed regularly by tracekey employees and customers. Due to a delay in the release process, the new authentication method was available on the production server only in the last week of the study (see *Section 5.5*).

The participants used two variants of FIDO2-compliant hardware tokens, the *Security Key by Yubico* and *YubiKey 5 NFC*, both from *Yubico* [30]. Both variants had the same form factor and supported WebAuthn, CTAP, and U2F. YubiKey 5 NFC offered additional features (e.g., support for OTP algorithms, OpenPGP, etc.) and could also be used via NFC. We did not require any of the additional features so that both key variants could be used in the same way.

3.1.1 Recruitment

We asked all employees who were eligible for our study (i.e., having accounts on both the public test server and the production server, using the web application on a daily basis, and being available at the time of the study). We invited the 10 employees who fulfilled these criteria to an in-house workshop, in which 9 attended. During the training session, we briefed the attendees about the purpose of the study, the procedure, risks and benefits, and the option to withdraw from the study at any time without penalty. We asked them to read and sign a consent form containing the same information. Participation was voluntary and uncompensated since they took part in the study during working hours.

We informed the participants about the required operating system and browser version. The minimum requirement was Microsoft Windows 10 version 1903 because it implemented CTAP2 in Windows Hello. The participants used Mozilla Firefox, Microsoft Edge, and Google Chrome. Of those three, only the stable version of Chrome did not support Windows Hello at the time of the study. We asked participants who used Google Chrome to switch to one of the other browsers or the beta version of Google Chrome.

3.1.2 Implementation

We used a server-side WebAuthn library provided by Yubico to implement the new FIDO2-based authentication. This library was integrated because the Spring Security framework used by the web application did not support WebAuthn at the time of the study.

To implement the login, we used the *resident credential* feature of FIDO2. This feature allows storing credential information like the username and private key on any authenticator with built-in memory. The participants did not need to enter their username because of this feature.

In addition to this, we redesigned the login screen of the web application (see Step 1 of Figure 2) to present both login methods (i.e., WebAuthn with security key and username and password). We decided to display both options on the same website to allow participants to choose between them without additional clicking on the website.

3.1.3 Data Preparation

At that time of the study, not all browsers stable versions supported the WebAuthn options used by our FIDO2-based authentication like user verification (i.e., via PIN) and resident keys (i.e., the username is chosen from a list instead of manually entered). Thus, most WebAuthn related errors could be attributed to the use of an unsupported browser.

We removed entries of accounts without a registered security key from the logs before the analysis, leaving logs containing entries of study participants or failed logins, which we could not attribute. The information provided by the timestamps of each login attempt was used to calculate average authentication times for the different login types aggregated across all participants (see Figure 4).

3.2 Study Protocol

The study was conducted over six-week in June/July 2019 and was framed as a usability study on security keys. It consisted of three phases. The first phase was a workshop in which we briefed the participants on using the security key. It followed a four-week phase of day-to-day use of the security key during which we collected data via authentication diaries and server logs. Finally, we interviewed the participants to discuss their experience using the security key and debrief them.

3.2.1 Initial Workshop

We started the study with a one-hour workshop. During this workshop, we first introduced the study as well as its purpose. We gave all participants consent forms, which also contained information about our study and their participation, and let them read and sign the form. A fifteen-minute training session introduced the security key, demonstrated the setup and use of the key, and showcased the user interfaces of each phase. It is important to highlight that we are specifically targeting a corporate context, where such training sessions are relatively common. This situation is very different from introducing WebAuthn to consumers.

After the training session, participants were handed the security keys and asked to set up the key on their work laptop computer and one of their accounts in the web application. Since most of the participants had multiple accounts for the web application, we encouraged them to register it with additional frequently used accounts and assisted them if necessary.

At the end of the workshop, participants filled in a questionnaire in which we gathered demographic data and feedback on the user interface and workflow of the implementation. We were especially interested in participants' knowledge about web authentication and used a modified version of the *web-use skill index* of Hargittai and Hsieh [14] for this purpose, focusing on authentication-related terms. We also gathered free-text responses about the participants' experiences with security keys and 2FA, and asked them to explain how a security key may improve the security of their account.

3.2.2 Authentication Diary

Over the next four weeks, the participants' task was to use the security key in their work routine. Additionally, we encouraged them to keep an authentication dairy in which they noted all logins over the first two days, and later only their failed login attempts. We adapted the diary from the work of Steves et al. [26]. Each diary entry comprised authentication time and date, on which server the participant tried to log in (e.g., the public test server) and whether the participants used the security key or username and password. They could also rate their satisfaction with the login on a five-item emoji-based scale we adapted from previous work [2, 20]. Despite their potential drawbacks (e.g., varying representations of the same emotion [19]), emojis were found to be well suited for affective self-reports of participants [28]. Furthermore, each diary entry had a section for errors and comments.

To enrich the authentication diaries and to get insights into how long it took the participants to authenticate, we also collected the timings of each login from the server logs. Each log entry contained multiple timestamps, username, user agent, WebAuthn-related information (e.g., the WebAuthn credential ID or the WebAuthn error message), and information about the login's success or failure. Especially in the case of failed login attempts, the logs provided additional insights.

3.2.3 Interview

After the four-week usage phase, we invited the participants to interviews. The interviews took place in a conference room in the company, and each session lasted 15 to 20 minutes. All but one participant were German native speakers, and we conducted seven interviews in German and one in English.

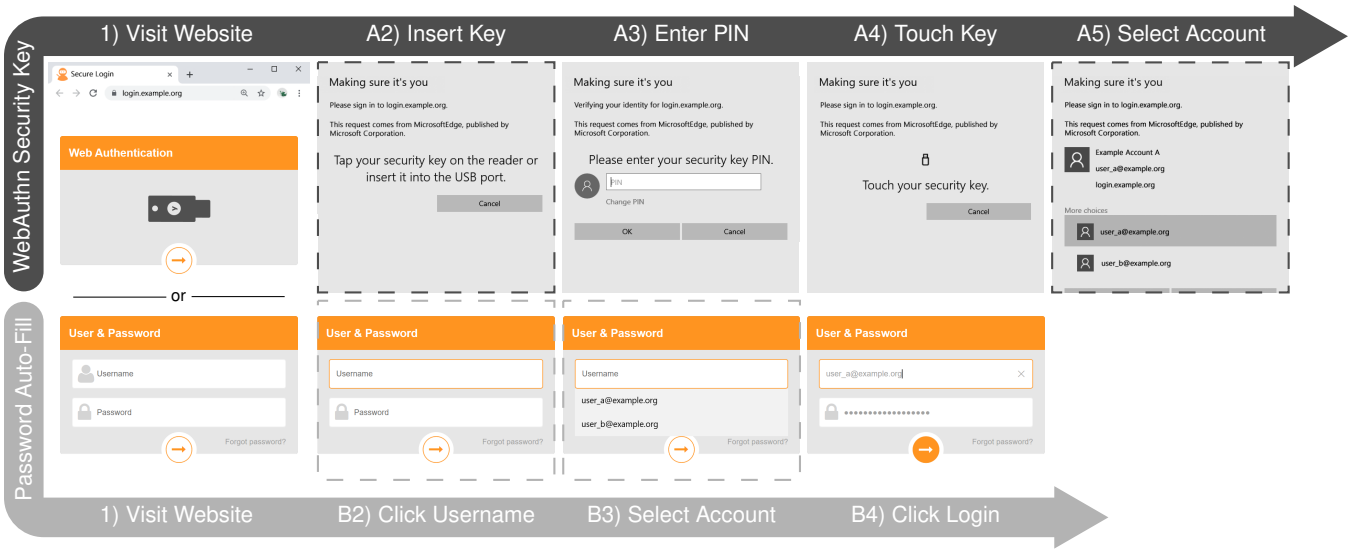


Figure 2: Comparison of the login procedures when using WebAuthn with a security key (top) vs. the browser's password auto-fill feature (bottom). The dashed borders are indicating omitted steps. If the security key is already plugged-in, the dialog in Step A2 does not show up. Having only one account stored on the key or in the browser, skips steps A5, B2, and B3.

The interviews were semi-structured and addressed the perceived usability of the security key, the differences between the key-based and password-based authentication, and the obstacles in using the key. Our goal was to examine why participants used the security key and what kept them from using it. The participants were also asked which of the two login methods they perceived as more secure.

We started each interview with a question about how they liked using the security key. The participants' answers gave us insights into their general impression of the security key and their experience of using it. We then asked them how frequently they used the web application and how often they used the key over the last four weeks. The participants were also asked to describe the difference between username password and security keys. If they mentioned 2FA, we let them explain what 2FA is and what the two factors are.

Additionally, we asked them to share their thoughts on which authentication method is more secure and their rationales. The full interview guide can be found in Appendix A.2.

In contrast to password-based authentication, our possession-based method required the user to have the security key with them. Thus, we wanted to know how hard it was to have the key at hand for the participants since this was the most obvious hurdle. However, we encouraged them to tell us about other obstacles they encountered as well. Finally, we asked them whether they would use the key in the future and for the reasons for their decision.

To analyze the interviews, we used a data-driven coding technique (as described in [10]). Two researchers independently coded all interviews through categorizing participant statements and identify recurring themes in each interview. They then compared the categories and themes across all interviews and created codes. A third researcher merged the themes and codes, derived a final codebook, and used it to code all interviews again. The full codebook is presented in Appendix B.

3.3 Demographics

Since we conducted our study in a small company, the participants had different backgrounds and positions. They were software developers, sales representatives, or project managers. One-third of the participants were female (3 out of 9) and the other two-thirds were male. The participants were 22–44 years old (mean: 30, SD: 6.3). Five (out of 9) participants had completed a master's degree, one holds a bachelor's degree, and two had studied at a university without completing a degree. One had completed a vocational training. All but one of the 9 people who attended the workshop completed the full study. The person who dropped out was on vacation during the second phase of the study and could not use the security key. Nevertheless, we included this person's feedback from the workshop in our evaluation.

Web Authentication Skills

To examine the participants' knowledge about web authentication, we adjusted the *Web-use Skill Measure* [14] to focus on web authentication, resulting in a skill survey containing ten authentication-related terms. We selected surveyed items from security awareness trainings, education materials, and infographics, including the NCSC glossary [22]. Table 1 shows the mean and standard deviation for all ten items.

Table 1: Web authentication skills are determined by rating the understanding of ten authentication-related items. The items are in the order of appearance in the questionnaire.

Item	Mean	SD
Malware	3.1	0.74
Phishing	3.3	0.94
Two-factor authentication	3.4	0.69
One-time password	2.5	1.26
Personal identification number	4.2	0.63
Auto-fill	3.2	1.55
Challenge-response	1.8	1.23
Password manager	3.9	0.99
Brute-force attack	2.4	1.43
Security question	4.0	0.82
Composite score	3.2	0.73
N	9	
Scale	5-point	
Cronbach's α	0.89	

The standard deviations were higher for items with a lower level of understanding (e.g., One-Time Password) than for items with a "high-level" understanding (e.g., personal identification number). These low-level items refer to more technical aspects of authentication. In contrast to the results obtained by Hargittai and Hsieh [14], our participants showed a higher understanding of the medium-level rated terms *Malware* and *Phishing*, indicating that the participants had an excellent understanding of authentication-related risks. A composite score of 3.2 indicates that the level of understanding lay between "some" and "good" understanding.

A Cronbach's α value of 0.89 as an estimate of the interrelatedness of items and internal consistency of the survey can be considered good, almost excellent. Due to the small sample size, statistical evidence is limited. However, the results met our expectations because all participants worked in a software company and reported to be tech-savvy.

3.4 Ethics

Our institution did not have a review board governing this type of study, so we discussed the study design with peers to validate our research's ethical perspective. We made sure to minimize any potential adverse effects from the study by following the ethical principles laid out in the Belmont report [21]. These principles included having an informed consent procedure at the beginning of the study and explaining to the participants that they could withdraw from the study without any negative consequences.

4 Results

Next, we present and discuss the results of our study. We evaluate the data we gathered through login diaries, server logs, and interviews.

4.1 Frequency of Authentication

Among the participants, the number of logins per day varied. Some participants logged into the web application multiple times a day, while others only used it once a week or less. Figure 3 shows how often the participants used the security key over the four weeks of the study broken down by participant and week of the study.

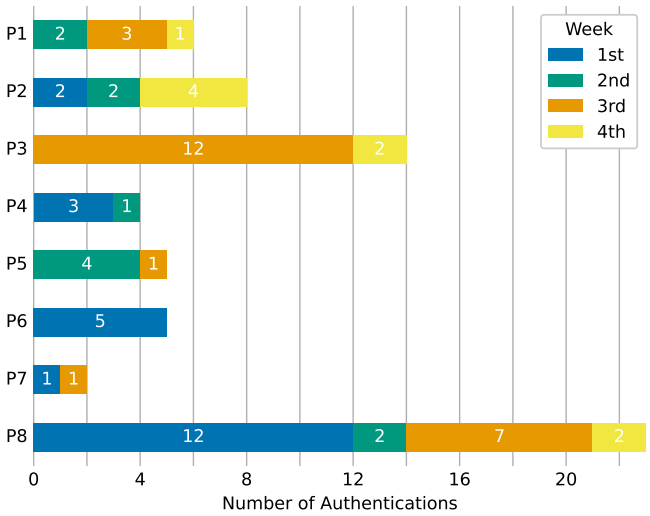


Figure 3: Breakdown of the number of authentications with the security key per participant and week. Except for participant P8, all participants used the security key only occasionally.

The reason for this discrepancy is that we conducted our study on the public test and evaluation server of the web application, which all the participants use but less often than the production server (see Section 5.5). We analyzed the server logs to gain insight into how often the participants used security keys. Table 2 presents a detailed breakdown of how many times the participants logged into the web application during the four weeks of the study.

For our analysis of the server logs, we filtered all log entries in which the WebAuthn method or the usernames of our participants were involved. After the filtering, we had 287 unique logins attempts over the four weeks. Surprisingly, only 67 (23.4 %) of these login attempts used a security key while 141 of the remaining 220 login attempts used browser auto-fill (i.e., browser password manager). We discuss this discrepancy in the number of logins in Section 4.4 in more detail.

Table 2: Breakdown of successful authentications per account, participant, and authentication method. Most of the participants registered their security keys for at least two accounts. Four participant pairs shared five accounts; for these accounts, we cannot distinguish which login belongs to which participant. *Manual logins* contain all login attempts for which the participants typed in the username and password manually or copied them from an external password manager/storage.

		Number of Authentication			
Account	**IDs**	**Security Key**	**Password Auto-fill**	**Manual Logins**	**Total**
A1	P1	4	7	10	21
A2	P1, P3	2	3	4	9
A3	P2	1	0	0	1
A4	P2	2	1	0	3
A5	P2	1	9	7	17
A6	P2, P4	2	33	3	38
A7	P2, P4	5	30	2	37
A8	P3	2	4	5	11
A9	P3	2	0	1	3
A10	P3	1	2	3	6
A11	P3	1	4	13	18
A12	P3, P4	9	6	12	27
A13	P5	0	20	2	22
A14	P5	5	6	8	19
A15	P6	4	9	3	16
A16	P6, P8	3	5	2	10
A17	P7	2	0	3	5
A18	P8	21	2	1	24
Total		67	141	79	287

To analyze login attempts with security key, we used the authentication diaries and the interviews. Participant P1 faced a problem where the touch sensor of the security key (see *A4) Touch key* in Figure 2) did not work at first, so he needed multiple tries until the login was successful. Participant P8 reported a similar problem. In this case, the software detected the security only after plugging it in again. Figure 2 shows the respective step labeled *A2) Search/insert key*. The other six participants reported no security key-related problems.

4.2 General Impressions

We started the interviews asking by participants about their general impressions using the security key over the four-week study period. The themes we found in our analysis of the interviews resulted in five categories of codes (cf. Appendix B): (i) Use of the security key, (ii) comparison of the security key with username and password, (iii) adoption barriers, (iv) general impression, and (v) perceived security.

Our warm-up question in the interviews was how they liked using the security key. Four participants appeared to be pleased about using the security key since they described the key as *easy to use* and its handling as *intuitive*. One participant was even enthusiastic when talking about impressions of the security key.

> P2: *"I don't need to remember anything. It's also faster, I am completely convinced. I think it's terrific."*

The four other participants started referring to minor issues which occurred when they used the key (e.g., being annoyed by touching the key because *"it is on the other side of the desk"* (P7)), two of which still rated its usage to be overall *"ok"* (P3, P5). The remaining two participants stated they did not use the key as often as intended and did not provide a clear judgment.

4.3 Authentication Timings and Convenience

During the interviews, the participants revealed that their convenience in authenticating varied in many aspects. In particular, their feedback on convenience using the security key highly depended on how they managed their password-based credentials since this was their ground truth against which they compared the new method.

We encountered three different ways of managing passwords: (i) The employees used a collaboration software where they stored their shared credentials and copy-and-pasted them into the respective login website, (ii) they saved their passwords using a third-party password manager software, or (iii) they used the browser's built-in password manager.

Five participants (P1, P2, P4, P5, P8) mentioned that the security key reduced the memory effort because they only needed to remember one PIN. Two participants who manually copy-and-pasted passwords (P2, P3) stated that using the security key was faster than entering username and password.

> P2: *"[...] but it is much more convenient if you can simply use this key, push it, enter your 4-digit PIN instead of your 12-character password [...]. It is also faster like this."*

In contrast to these two participants, five participants used browser built-in password managers with a password auto-fill feature and stated that the authentication with auto-fill required fewer steps than using the security key (P4, P6, P7) or was faster (P5, P6, P7, P8). The timings we extracted from the server logs support their statements. We measured the time starting when the login website was fully loaded and ending when the login form was submitted to the server for each authentication attempt. Figure 4 presents the timings of the different login variants. It shows that the security key was slower than the password auto-fill feature of a browser.

For authentication attempts using the security key, we could not determine whether the measured times include the time to reach for and plug in the security key or if it were already plugged in before plugged in before. We expect our result set to comprise both scenarios. However, since we consider physical interactions with the security key part of the authentication ceremony, our results provide a best-case estimate. If the measured times had not included preparing the key, the gap compared to the password auto-fill timings would have become even larger.

Figure 2 illustrates the steps required to log in with the key compared to the browser's auto-fill feature. Using the security key requires three steps in a best-case scenario. This scenario requires that the key is already plugged into the computer (Step A2), only the security key is configured for Windows Hello (Step A2), and only one account is registered for the key (Step A5). The three steps are (i) clicking the login button beneath the security key symbol (see Step 1), (ii) entering the PIN of the security key as shown in Step A3), and (iii) touching the key (see Step A4).

In contrast, the credential auto-fill via browser requires at best (with only one username-password pair stored in the browser for the website) one click on the login button, as shown in Step B4. Otherwise, the user needs to click the text field for the username (Step B2) and select the desired account (Step B3). In both cases, the auto-fill procedure requires fewer steps and no physical interaction with additional hardware, which explains why it is the faster authentication procedure.

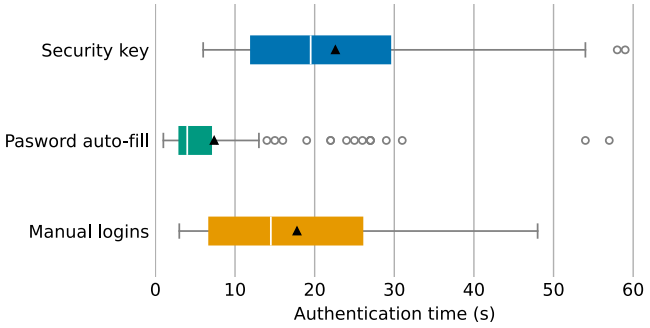


Figure 4: Authentication timings for different login variants. The time spent by participants to authenticate varies depending on the login type used. (▲ denotes the mean.)

Participant P4 provided the most differentiated feedback, taking into account both manually copying and pasting passwords as well as using the browser's password manager when assessing the convenience using the security key.

> P4: *"I already have the passwords saved in my Edge account for all different accounts. So that was more convenient for me because it's hardly one click [...] Even if I save something in the browser, it will work that way as well but if you use the security key it will definitely [be] time saving as well [...]."*

Another participant referred to the use of the 2FA-protected third-party password manager and preferred using the security key because it was less cumbersome.

> P5: *"In comparison I think that the key is more user-friendly, it requires less effort than invoking both WinAuth and KeePass."*

4.4 Weighing Security and Purposes

During the interviews, participants indicated awareness of different security requirements for different services. We unveiled such tendencies when asking participants about the purposes they had used the security key for. Getting a full view is a two-step process since we also needed to capture how the participants estimated the security of the available authentication options.

4.4.1 Security of Authentication Schemes

While six participants (P3–P8) rated the security key as more secure than the password-based authentication schemes, participant P1 guessed that the password copy-and-paste mechanism is presumably more secure than the security key. However, he mentioned the risk of losing the key and, consequently, becoming unavailable to log in, as a reason for the key's reduced level of security.

> P1: *"I guess you can use such a security key, and how do you log in if you don't have it? [...] So, I believe the standard way [i.e., using passwords] is maybe more secure? Well, I'm not sure if it's 'more secure' but I can log in in any case [...]"*

Another participant stated that the security offered a good security level but refrained from deeming one scheme more secure than another. Participants P2, P5, and P8 explained that the security of the key-based authentication is more secure because it relies on "two factors", i.e., possession of the key and knowing the correct PIN. Furthermore, participant P5 elaborated his understanding of how the security key works, alleging that the passwords are stored on the key.

> P5: *"It is secure, if I understand correctly, because the passwords are stored on the key and, therefore, are not affected if I have a compromised machine [...]"*

Even though this explanation is not correct from a technical perspective, the idea of P5 can give non-expert users some useful intuition why the key is more secure than passwords.

4.4.2 Use of the Security Key

In total, three participants talked about different security requirements depending on a service's purpose. Two participants named online banking (P6, P7), another authentication at work (P5) as use cases that require a higher level of security.

When asked whether they want to keep using the key after the study, four participants signaled willingness (P2, P3, P4, P8), two were not sure (P5, P6), and two did not want to use the key in the future (P1, P7). Opinions of those who could imagine continuing using the key ranged from plain approvals (P2, P8: *"yes"*) to readiness to use the key exclusively if the working environment supported this:

> P3: *"Yes, and if it would work with the production server, I would work exclusively using the key."*

However, participants revealed different opinions on extending the use to personal accounts. While one participant claimed willingness to use the key for personal purposes (P2), two participants stated that they would not use a security key off the job.

While P4 stated to use a password manager for personal accounts, P7 unambiguously explained that the additional time required to use the key compared to a password manager is an unacceptable trade-off in everyday use.

> P7: *"In the time it takes to dig it up, plug it in, enter the PIN, and push it – I could have already bought two pairs of shoes."*

Another participant remained rather indecisive when asked about using the key for personal use, due to the additional overhead compared to the use of a password manager.

> P8: *"When you have stored your passwords in your browser, it is still faster than picking the key, plugging it in and entering the PIN."*

4.5 Adoption Barriers

The authentication diaries and responses in the interviews show that usage rates of the security key were rather low (see Section 4.1), so we asked the participants what prevented them from using the security key in particular situations. They reported several hurdles when using the key, e.g., the fear of losing access to their accounts, the additional effort/time required to plug the key in or unlock it, and the habitual use of passwords.

Participants P5 and P7 mentioned that the additional effort and time to interact with the key makes it less convenient than authenticating with username and password.

> P5: *"Well, it would start to make a real difference if I didn't have to enter anything at all but only had to touch the key."*

The time factor was not only important for authentication, two participants (P5, P6) also mentioned that they needed to invest time ("*5–10 minutes*") to set up the key for an account, which implies that even comparably small amounts of time can be an adoption hurdle.

> P6: *"There is this small initial effort you need to find five to ten minutes for."*

Besides the time aspects, participants' feedback from the interviews unveiled further details about additional obstacles towards the adoption of the security key.

4.5.1 Fallback Authentication

Several participants expressed concerns about not being able to log in if they do not carry their keys (P2, P3) or even lose them (P1). Participant P3 further mentioned a potential risk of technical flaws, hampering the ability to log in. The majority of participants had no problems with carrying the key, e.g., by attaching it to their key rings.

> P6: *"I have it on my [...] key ring. Thus, I have it with me all the time."*

However, this was not a proper solution for every participant. P7 explicitly deemed this a disadvantage and suggested providing a solution to attach the key to a smartphone.

> P7: *"That's cumbersome, and I also have quite few keys, and I don't want to plug them all in."*

Potential problems about handling the key were also mentioned by another participant (P2), who feared to destroy the key due to its tiny size and shakiness when plugged in and suggested a more sturdy design.

These answers show that participants' views were not limited to the scope of the study but also widened for the security key's general use as a single first authentication factor. Not being able to log in due to losing the key was not a real risk in our scenario since participants could always choose between password-based and key-based authentication during the study.

4.5.2 Workflow and Environment

The workflow of the authentication or the work environment can also be an adoption barrier. As discussed in Section 4.3, the security key requires more interactions for each authentication. Participant P6 explicitly mentioned the higher click-count when using the security key (i.e., the Windows Hello user interface) compared to entering the username and password or using the browser's password manager.

> P6: *"The workflow needs to be simple. Like even faster. Fewer clicks."*

Another participant saw the account selection of Windows Hello as a hurdle for adoption.

> P1: *"I have 40 accounts or so. When I register the key for all of them and I want to log in then, I need to scroll through all of them [...] That's a little bit time-consuming."*

Two participants indicated that the characteristics of their work environment made it more challenging to integrate the key into their routines. For instance, picking an account to log in when multiple accounts were used on the same website, since the names of all accounts followed a naming convention made them very similar.

> P3: *"I cannot see at first sight to which account [the username] actually belongs."*

Similar, another participant denoted that the login is slightly different depending on the operating system and browser (here Windows 10 version 1903 and Google Chrome 76).

> P6: *"[...] on my computer the Windows update has not been rolled out yet, so, I could not use Firefox [...] and [in the other browser] I need to click three times and touch [the key] twice to log in."*

4.5.3 Routine in using Passwords

Considering aspects of introducing and integrating new security measures into a well-functioning working environment raises questions about how to overcome long-established security behaviors. Participant P7, who showed a generally rather reserved attitude towards using the security key, remarked that routine in handling passwords was a reason not to use the security key for authentication, especially in cases when something needed to be checked quickly.

> P7: *"It was not an active decision [to use the password], but rather a situation when I just had to get things done. [...] I'm just used to it, because I know the password for this application. [...] It's like an addiction. You still use the password after all."*

This statement suggests that deeply ingrained security habits cannot be challenged, let alone be replaced easily, not within the four weeks of our study.

5 Discussion

In contrast to prior work investigating security keys in the context of 2FA [8, 9, 15, 16, 24], we focused on a secure authentication scheme without username and password. From a user perspective, the difference is to have a new authentication scheme instead of adding steps to a well-known one. The resulting process still suffers from similar issues, as found by studies on security keys in a 2FA context [8, 16].

However, our analysis of the interviews and authentication dairies identified three problem areas unique to or more important for the security key as a primary authentication factor: (i) concerns about account recovery in case of key loss or defect, (ii) having a more complex and possibly slower authentication process, and (iii) security benefits are intangible.

Losing access to the account through defect or loss of the security key is the primary concern for users, especially if it is the only way to authenticate. A possible solution to this problem is registering multiple security keys for one account, but this seems to be an additional burden to the users. Another option is using username and password as fallback authentication, but this nullifies the security benefits of FIDO2. The question of how to realize secure fallback authentication for FIDO2 is still open for future research.

The differences in the timings between the security key and other login methods, as indicated in Figure 4 and also mentioned by multiple participants, needs to be tested with a larger sample. Other authenticators could make "passwordless" FIDO2-based authentication faster and less complex.

Even though most participants found the security key to be more secure than username and password, the reasons why the key was, in fact, more secure were hard to grasp for them. More research on how to explain the security benefits of FIDO2-based authentication schemes is needed. In the following, we discuss our results using the three categories to assess authentication schemes proposed by Bonneau et al. [5].

5.1 Usability

Our study indicates that the security key is usable in the sense that all the participants understood how to use the key correctly and comprehended the on-screen instructions. The problems the participants encountered were rare, and they could solve all of them.

Authentication times appeared to be more of a limiting factor. Using the browser's auto-fill feature was the fastest authentication method in our scenario. Even "manual logins" (e.g., copy and paste the password) may be faster than using the security key. This speed difference may be one of the primary adoption barriers.

The level of routine and habituation users developed with passwords is high. Some interview statements implied that password-based authentication does not necessarily induce friction but works as an unconscious background process that makes it more challenging to get used to a new scheme like the security key. Protecting the key with a PIN also undermines its benefit of allowing "passwordless" authentication.

Physical aspects and related handling of the key could also be an obstacle. Actions like carrying the security key around, e.g., by attaching it to a key ring, inserting it, or touching its button were a hurdle for some participants. However, the degree of inconvenience appears to depend highly on the user's perception and predispositions.

5.2 Deployability

WebAuthn support, while increasing, is the biggest deployment issue. This lack of support comprises operating systems, browsers, but also software frameworks to ease integration. Only recent versions of operating systems and browsers work with all WebAuthn features, thus requiring web service owners to offer alternative authentication schemes and show appropriate error messages in case of unsupported operations.

Missing best practices on login form design for WebAuthn-based authentication hinders a consistent user experience across different web services. While a typical design has emerged for password-based login forms over time, there is no such design for WebAuthn. Research on the impact of such forms is scarce.

On the other hand, direct costs for hardware or support and indirect costs through lost productivity are negligible. In a company with tech-savvy personnel, the expense for the adoption of a security key-based authentication (not including implementation) should not be too high. As Lang et al. [16] showed with a predecessor of U2F and thus FIDO2, the same can be true for large companies.

5.3 (Perceived) Security

Security is only secondary to usability when choosing an authentication scheme in daily work life. If security is not promoted as an essential part of work instead of being just an obstacle to other tasks, that fact remains [11].

The benefits of the security key, especially when it requires a PIN, need to be conveyed clearly. While creating risk awareness helps users to make informed decisions, reminding users of the benefits provided by a scheme seems to be even more promising [9].

5.4 Overcome Adoption Barriers

We think that the FIDO2 project can replace password-based authentication on the Web in the long run. However, at the moment, only a few applications or Internet services support FIDO2-based authentication, which impedes its adoption, and the lack of reference implementations of the WebAuthn server-side hinders its integration. These obstacles will probably resolve over time. To overcome some of the other adaption barriers we found, we have the following suggestions:

- Support multiple different authenticators (platform and roaming authenticators if possible);
- Require adaption of the security key in organizations (as suggested by Colnago et al. [8] for 2FA);
- Make FIDO2-based authentication available for as many systems as possible
- For PIN-protected security keys, allow to "remember" the PIN until the key is unplugged.

5.5 Limitations

Due to the qualitative nature of the study and the small sample size, it can only provide a first insight into "passwordless" authentication with security keys. Our results may not apply to a broader population but indicate potential interesting topics and raise new research questions.

We deployed the new authentication method on a public test server for our study. Although all participants had access to the server and user accounts on it, half of them (4 participants) mentioned in the interviews that they used the test server less often than the production system. The low use of the key affected how the participants used the security key and might have had an impact on the perceived usefulness of it.

6 Related Work

The FIDO2 project has not been around for a long time, which is why research in this specific area is limited. Most related to our study is the work by Lyastani et al. [17], as it does investigate the use of tokens for primary authentication in the context of FIDO2. Lyastani et al. [17] conducted a lab study with end-users to get insights into the perception, acceptance, and concerns when security keys are used for passwordless authentication. Their results conform to ours in terms of a mixed impression with an overall positive impression of token-based authentication but also the fear of the participants to lose the token locking themselves out.

In contrast to Lyastani et al. [17], we implemented the token-based authentication in combination with a PIN because it counteracts one of the disadvantages the participants in the study by Lyastani et al. [17] mentioned, namely, the risk of illegal access when the authenticator is lost. However, this combination made the workflow more complex, which again can impede adopting the key. Additionally, we concentrate on the long-term and real-world impact of using token-based authentication in a corporate context. Through this different focus in the study, we saw that overcoming the routine users gained in using passwords as an additional adoption barrier.

Besides FIDO2, Pico, proposed by Stajano [25], is another example of a token-based login method. In a study by Aebischer et al. [1], users appreciated the ability to avoid passwords because of the known drawbacks, but adoption was still identified as a problem as users prefer to stick to the familiar password-based authentication. We observed a similar phenomenon, among the participants who used a password manager. Although they were convinced that the security key-based solution was more secure, but they preferred the password managers because they were fast.

While we are interested in hardware tokens as a first factor, research into 2FA is also related to our work as it gives important insights into the use of tokens for authentication. Despite what factor is used, the initial setup of an additional second factor is one of the major issues for most users [7, 15, 23, 24].

A corporate context – in which we were interested in this study – allows counteracting this problem by offering customized guidance for the setup phase. Because of this, we walked the participants through the initial steps.

Studies by Das et al. [9] and Ciolino et al. [7] further analyzed why users decide not to use 2FA security keys. They found two main reasons: (i) Users are afraid of losing their key locking themselves out and (ii) some users also do not fear an account takeover, which is why they see no necessity for the additional effort associated with the use of a security key. These findings are supported by adoption rates reported by Google with less than 10 % of active Google accounts having 2FA enabled [18]. While these findings are relevant for the end-user, the situation in a corporate context is different. Here, the motivation to use security keys is driven by the company, and using them can be made mandatory for the employees.

Furthermore, it was found that users have an overall positive attitude towards security keys once they are in place as a second factor. [7, 9, 24]. They are seen as easy to use and increase the perceived security [8, 16]. We come to a similar conclusion for the case when security keys are used as a primary factor. Regarding the timing of security key-based logins, Reese et al. [23] found that login times decrease the longer and the more often security keys are used. Some participants also mentioned the fast authentication time, yet some abandoned the security keys in favor of a password manager because it allows an even more efficient authentication.

7 Conclusion

We conducted a qualitative study on the usability of FIDO2, using USB-based hardware tokens in the form of security keys in the context of a small company. The core components of FIDO2 – WebAuthn and CTAP – offer promising alternatives to the dominant username-password scheme used for web authentication. FIDO2 security keys present a phishing-resistant form of hardware tokens suited as the primary authentication factor for web applications.

In contrast to previous work on authentication via security keys, we focused on using the key as a primary authentication factor instead of having it as an additional factor in a 2FA setting. Although most participants considered the security key-based login as usable, several of them stopped using the key as it was slower than using the password manager built in their browsers. Furthermore, the security benefits were largely intangible or perceived as unnecessary by the participants. Another issue was the missing support of some browser and operating systems at the time of the study. All these adoption barriers should be minimized before introducing FIDO2 (with security keys) to replace username and password-based authentication in a company.

Acknowledgments

The authors would like to thank *tracekey solutions GmbH* for allowing them to conduct a study in their company and providing the necessary software and hardware. This research was partially funded by the MKW-NRW research training groups SecHuman, “Human Centered Systems Security” sponsored by the state of North Rhine-Westphalia, Germany, and the German Research Foundation (DFG) within the framework of the Excellence Strategy of the Federal Government and the States – EXC2092 *CASA* – 390781972.

References

[1] Seb Aebischer, Claudio Dettoni, Graeme Craig Jenkinson, Katarzyna Kinga Krol, David Llewellyn-Jones, Toshiyuki Masui, and Francesco Mario Stajano. Pico in the Wild: Replacing Passwords, One Site at a Time. In *European Workshop on Usable Security*, EuroUSEC ’16, Paris, France, April 2017. ISOC.

[2] Sarah Alismail and Hengwei Zhang. The Use of Emoji in Electronic User Experience Questionnaire: An Exploratory Case Study. In *Hawaii International Conference on System Sciences*, HICSS ’18, Waikoloa Village, Hawaii, USA, January 2018. ACM.

[3] Anti-Phishing Working Group. Phishing Attack Trends Report – 2Q 2018, October 2018. https://docs.apwg.org/reports/apwg_trends_report_q2_2018.pdf, as of April 30, 2020.

[4] Dirk Balfanz, James C. Jones, Akshay Kumar, Huakai Liao, Jeff Hodges, Emil Lundberg, Czeskis Alexei, Rolf Lindemann, and Michael Jones. Web Authentication: An API for accessing Public Key Credentials Level 1. Recommendation REC-webauthn-1-20190304, World Wide Web Consortium, March 2019.

[5] Joseph Bonneau, Cormac Herley, Paul C. Van Oorschot, and Frank Stajano. The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes. In *IEEE Symposium on Security and Privacy*, SP ’12, pages 553–567, San Jose, California, USA, May 2012. IEEE.

[6] Christiaan Brand, Alexei Czeskis, Jakob Ehrensvärd, Michael B. Jones, Akshay Kumar, Rolf Lindemann, Adam Powers, and Johan Verrept. Client to Authenticator Protocol (CTAP). Proposed Standard fido-v2.0-ps-20190130, FIDO Alliance, January 2019.

[7] Stéphane Ciolino, Simon Parkin, and Paul Dunphy. Of Two Minds about Two-Factor: Understanding Everyday FIDO U2F Usability through Device Comparison and Experience Sampling. In *Symposium on Usable Privacy*

and Security, SOUPS '19, pages 339–356, Santa Clara, California, USA, August 2019. USENIX.

[8] Jessica Colnago, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Lorrie Cranor, and Nicolas Christin. "It's Not Actually That Horrible": Exploring Adoption of Two-Factor Authentication at a University. In *ACM Conference on Human Factors in Computing Systems*, CHI '18, pages 456:1–456:11, Montreal, Quebec, Canada, April 2018. ACM.

[9] Sanchari Das, Andrew Dingman, and L. Jean Camp. Why Johnny Doesn't Use Two Factor A Two-Phase Usability Study of the FIDO U2F Security Key. In *Financial Cryptography and Data Security*, FC '18, pages 160–179, Santa Barbara Beach, Curação, February 2018. Springer.

[10] Jessica T. DeCuir-Gunby, Patricia L. Marshall, and Allison W. McCulloch. Developing and Using a Codebook for the Analysis of Interview Data: An Example from a Professional Development Research Project. *Field methods*, 23(2):136–155, December 2011.

[11] Geoffrey B. Duggan, Hilary Johnson, and Beate Grawemeyer. Rational Security: Modelling Everyday Password Use. *International Journal of Human-Computer Studies*, 70(6):415–431, June 2012.

[12] European Parliament and Council of the European Union. Directive 2011/62/EU of the European Parliament and of the Council of 8 June 2011 amending Directive 2001/83/EC on the Community code relating to medicinal products for human use, as regards the prevention of the entry into the legal supply chain of falsified medicinal products. *Official Journal of the European Union*, 174(1):74–87, July 2011.

[13] FIDO Alliance. The FIDO ("Fast IDentity Online") Alliance – Industry Association to Promote Authentication Standards, February 2013. https://fidoalliance.org, as of April 30, 2020.

[14] Eszter Hargittai and Yuli Patrick Hsieh. Succinct Survey Measures of Web-Use Skills. *Social Science Computer Review*, 30(1):95–107, February 2012.

[15] Kat Krol, Eleni Philippou, Emiliano De Cristofaro, and M. Angela Sasse. "They brought in the horrible key ring thing!" Analysing the Usability of Two-Factor Authentication in UK Online Banking. In *Workshop on Usable Security*, USEC '15, San Diego, California, USA, February 2015. ISOC.

[16] Juan Lang, Alexei Czeskis, Dirk Balfanz, Marius Schilder, and Sampath Srinivas. Security Keys: Practical Cryptographic Second Factors for the Modern Web. In *Financial Cryptography and Data Security*, FC '16, pages 422–440, Accra Beach, Barbados, February 2016. Springer.

[17] Sanam Ghorbani Lyastani, Michael Schilling, Michaela Neumayr, Michael Backes, and Sven Bugiel. Is FIDO2 the Kingslayer of User Authentication? A Comparative Usability Study of FIDO2 Passwordless Authentication. In *IEEE Symposium on Security and Privacy*, SP '20, San Francisco, California, USA, May 2020. IEEE.

[18] Grzergor Milka. Anatomy of Account Takeover. In *USENIX Enigma Conference*, Enigma '18, Santa Clara, California, USA, January 2018. USENIX.

[19] Hannah Jean Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. "Blissfully Happy" or "Ready to Fight": Varying Interpretations of Emoji. In *AAAI Conference on Web and Social Media*, ICWSM '16, pages 259–268, Cologne, Germany, May 2016. AAAI.

[20] Adam Moore, Christina M. Steiner, and Owen Conlan. Design and Development of an Empirical Smiley-Based Affective Instrument. In *Workshop on Emotions and Personality in Personalized Services*, EMPIRE '13, pages 41–52, Rome, Italy, January 2013. CEUR.

[21] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research, September 1978. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html, as of April 30, 2020.

[22] National Cyber Security Centre. Common Words and Phrases Relating to Cyber Security, January 2018. https://www.ncsc.gov.uk/information/ncsc-glossary, as of April 30, 2020.

[23] Ken Reese, Trevor Smith, Jonathan Dutson, Jonathan Armknecht, Jacob Cameron, and Kent Seamons. A Usability Study of Five Two-Factor Authentication Methods. In *Symposium on Usable Privacy and Security*, SOUPS '19, pages 357–370, Santa Clara, California, USA, August 2019. USENIX.

[24] Joshua Reynolds, Trevor Smith, Ken Reese, Luke Dickinson, Scott Ruoti, and Kent Seamons. A Tale of Two Studies: The Best and Worst of YubiKey Usability. In *IEEE Symposium on Security and Privacy*, SP '18, pages 872–888, Oakland, California, USA, May 2018. IEEE.

[25] Frank Stajano. Pico: No More Passwords. In *Workshop on Security Protocols*, SPW '11, pages 49–81, Cambridge, United Kingdom, March 2011. Springer.

[26] Michelle Steves, Dana Chisnell, M. Angela Sasse, Kat Krol, Mary Theofanos, and Hannah Wald. Report: Authentication Diary Study. Technical Report 7983, NIST, February 2014.

[27] Symantec Corporation. Internet Security Threat Report Volume 23, 2018. https://www.symantec.com/content/dam/symantec/docs/reports/istr-23-2018-en.pdf, as of April 30, 2020.

[28] Alexander Toet, Daisuke Kaneko, Shota Ushiama, Sofie Hoving, Inge Kruijf, Anne-Marie Brouwer, Victor Kallen, and Jan Erp. EmojiGrid: A 2D Pictorial Scale for the Assessment of Food Elicited Emotions. *Food Research International*, 115:541–551, January 2019.

[29] W3C. The World Wide Web Consortium (W3C) – Standards Organization for the World Wide Web, October 1994. https://www.w3.org, as of April 30, 2020.

[30] Yubico AB. Yubico | YubiKey strong two factor authentication, February 2020. https://www.yubico.com/, as of April 30, 2020.

A Study Materials

This appendix contains the materials we used to conduct the study. All information about the authors or participants have been removed.

A.1 Initial Workshop Questionnaire

How familiar are you with the following authentication and security-related items? Please choose a number between 1 and 5 where 1 represents "no understanding" and 5 represents "full understanding" of the item.

	None (1)	Little (2)	Some (3)	Good (4)	Full (5)
Malware	◯	◯	◯	◯	◯
Phishing	◯	◯	◯	◯	◯
Two-factor authentication	◯	◯	◯	◯	◯
One-time password	◯	◯	◯	◯	◯
Personal identification number	◯	◯	◯	◯	◯
Auto-fill	◯	◯	◯	◯	◯
Challenge-response	◯	◯	◯	◯	◯
Password manager	◯	◯	◯	◯	◯
Brute-force attack	◯	◯	◯	◯	◯
Security question	◯	◯	◯	◯	◯

Please answer all of the following questions.

1. Have you ever used a security key before?
   - ◯ Yes, and I still do
   - ◯ Yes, but I stopped using it
   - ◯ No
   - ◯ I do not know
2. Why? Why not?

   Answer: _____
3. Have you ever used two-factor authentication for any of your online accounts?
   - ◯ Yes, and I still do
   - ◯ Yes, but I stopped using it
   - ◯ No
   - ◯ I do not know
4. Why? Why not?

   Answer: _____
5. What did you like about the setup procedure of the security key?

   Answer: _____
6. How would you improve the setup procedure of the security key?

   Answer: _____
7. How does a security key make your account more secure?

   Answer: _____
8. Do you have any comments, ideas, or suggestions for improvement?

   Answer: _____

A.2 Interview Guideline

(a) Introduction

- Thanks again for taking part in the security key evaluation over the past four weeks and also thank you for agreeing to this interview.
- The interview will take 10-15 minutes.
- Are you OK with me recording our interview?
- <*Start recording.*>
- There are obviously no right or wrong answers here, I am just interested in your personal perceptions and your honest opinions.
- Any questions? Can we start?

(b) Interview

- You were able to test a security key over the last 4 weeks, how do you like it?

  . . . . . .
- How many times a day do you log in on average?
- What do you think, how often have you used the security key during the last 4 weeks?
- Have you registered additional accounts with the security key after the training?

  . . . . . .
- <*Ask questions based on the authentication diary.*>

  . . . . . .
- What are the differences between security keys and passwords?
- Do you think the security key offers any benefits compared to username and password?
- What do you think is more secure?
- How easy or difficult was it to have your security key with you whenever you needed it?
- What kept you from using the security keys?

  . . . . . .
- What was the best part of using the security key?
- What would you improve about the user-friendliness of using the security key?
- Would you like to continue to use the security key?

(c) Debriefing

- <*Briefly summarize interview.*>
- Study goal: We investigate security keys as a password "replacement".
- We are interested in usability issues of these keys.
- Do you have any questions about the interview or the study?
- <*Stop recording.*>

B Codebook

Table 1: **Category:** Use of the security key – The participant indicates if or for what purposes they want to use the security key.

Code	Freq.	Description	Example
Continue using key	6	The participant states to continue using the security key after the study.	*"For all new projects which I'll get, I'll use the security key."* (P4)
Personal use	5	The participant considers using the security key also for their personal accounts.	*"[...] and I'd also use it in my private life."* (P2)
Only for work	5	The participant states to only use the security key for work.	*"I feel like I will just use the security key for work and KeePass and all that for personal stuff."* (P4)
Key is impractical for multiple accounts	4	The participant states that using the security key is impractical for users with more than 5 accounts on one website (Windows Hello selection dialog).	*"As I said, I have 40 different accounts for work. In this case, it's not really practical."* (P1)
Only for sensitive accounts	4	The participant indicates to use the security key only for sensitive accounts (e.g., online banking).	*"And I think in cases where one needs more security, it's good and I could understand it [to use a security key]."* (P7)
No need for the key	2	The participant mentions that using the security key for only one account is not worth the effort.	*"If I'd say: 'Okay, I just got 20 new accounts', then maybe, but with one account, no."* (P7)
Stop using key	2	The participant states to not continue using the security key after the study.	*"I don't think I would continue using it."* (P1)

Table 2: **Category:** Comparison of the security key – The participant compares a certain aspect of the security key with username and password.

Code	Freq.	Description	Example
Key requires more steps than browser PW	7	The participant states using the security key requires more steps/clicks compared to using the built-in password manger of the browser.	*"For already existing accounts, I already had the passwords saved. So, that was more convenient."* (P4)
Key slower than browser PW	6	The participant indicates that authentication with the security key takes longer than using the built-in password manger of the browser.	*"When you have stored your passwords in your browser, it is still faster than picking the key, plugging it in, and entering the PIN."* (P8)
Key memory-wise less effort	5	The participant states using the security key is compared to username and password more convenient because they only need to remember one PIN.	*"I'd prefer to use the key. I think it's easier to only remember the PIN, like just this one PIN and nothing else."* (P5)
Key faster	5	The participant indicates that authentication with the security key is faster than with username and password.	*"It's time saving. Absolutely."* (P4)
No difference	1	The participant finds both security key and username and password equally convenient/inconvenient.	*"But for me it doesn't make a huge difference whether I manually type in a password, or if I type in the PIN for the key. Unfortunately, it doesn't make a huge difference."* (P5)
Key more cognitive effort	1	The participant states that using the security key requires more thinking than entering username and password.	*"Touching the key is something different. [...] sometimes I don't think about what I'm doing, I just do it. And then I find myself using the password again."* (P7)

Table 3: **Category:** Adoption barriers – The participant refers to an obstacle or possible obstacle when using the security key.

Code	Freq.	Description	Example
Higher effort	5	The participant finds carrying/plugging in/unlocking the security key cumbersome or time-consuming.	*"Well, it would start to make a real difference if I didn't have to enter anything at all but only had to touch the key."* (P5)
Fear to lock out	4	The participant fears to lose access to web application through loss or defect of the key.	*"Well, if I forget or loose it, I couldn't get into my account"* (P3)
Routine with passwords	3	The participant mentions to use username and password instead of the security key out of habit.	*"It's like an addiction. You still use the password after all."* (P7)
Setup per account	3	The participant thinks the effort to set up the key for multiple accounts can be an obstacle to adoption.	*"There is this small initial effort you need to find five to ten minutes for."* (P6)
More complex workflow	2	The participant reports usability issues, like higher click count or confusing account selection in Windows Hello.	*"The workflow needs to be simple. Like even faster. Fewer clicks."* (P6)
Forgot to take key	1	The participant states to have forgotten the key and therefore not being able to use it to log in.	*"Well, I assume I always forgot to take it with me."* (P2)
Key perceived as fragile	1	The participant reports difficulties with the form factor of the security key (e.g., the fear to break it accidentally).	*"It always bends so easily, and I thought: 'Oh my god, now I'm breaking the poor thing.'"* (P2)

Table 4: **Category:** General impression – The participant mentions their general impression of the security key.

Code	Freq.	Description	Example
Key is usable/convenient	9	The participant finds the security key generally usable or convenient.	*"You click it [the account name], quickly enter your PIN, touch the key with your finger, and you're done. It's smooth."* (P5)
Key is easy/intuitive	6	The participant finds the security key generally easy or intuitive to use.	*"It's easy, like really easy. I'm a huge fan I have to say."* (P2)
Key is cool/novel	2	The participant shows enthusiasm because the security key is "new" or "cool" technology.	*"Well, at the beginning I started very enthusiastically. I really thought it's cool thing. "* (P6)

Table 5: **Category:** Perceived security – The participant comments on the security of the security key.

Code	Freq.	Description	Example
Key more secure	11	The participant states that the security key is more secure than username and password	*"Yes, I think one thing I like is that no password is sent, if I understand that correctly, over the Internet but it [the password] just decrypts the key locally."* (P6)
Unsure	3	The participant is unsure whether the security key is more secure or not.	*"I guess you can use such a security key, and how do you log in if you don't have it? [...] So, I believe the standard way [i.e., using passwords] is maybe more secure? Well, I'm not sure if it's 'more secure' but I can log in in any case [...]"* (P1)

Knock, Knock. Who's There? On the Security of LG's Knock Codes

Raina Samuel
New Jersey Institute of Technology
res9@njit.edu

Philipp Markert
Ruhr University Bochum
philipp.markert@rub.de

Adam J. Aviv
The George Washington University
aaviv@gwu.edu

Iulian Neamtiu
New Jersey Institute of Technology
ineamtiu@njit.edu

Abstract

Knock Codes are a knowledge-based unlock authentication scheme used on LG smartphones where a user enters a code by tapping or "knocking" a sequence on a 2x2 grid. While a lesser-used authentication method, as compared to PINs or Android patterns, there is likely a large number of Knock Code users; we estimate, 700,000–2,500,000 in the US alone. In this paper, we studied Knock Codes security asking participants in an online study to select codes on mobile devices in three settings: a control treatment, a blocklist treatment, and a treatment with a larger, 2x3 grid. We find that Knock Codes are significantly weaker than other deployed authentication, e.g., PINs or Android patterns. In a simulated attacker setting, 2x3 grids offered no additional security. Blocklisting, on the other hand, was more beneficial, making Knock Codes' security similar to Android patterns. Participants expressed positive perceptions of Knock Codes, yet usability was challenged. SUS values were "marginal" or "ok" across treatments. Based on these findings, we recommend deploying blocklists for selecting a Knock Code because they improve security but have a limited impact on usability perceptions.

1 Introduction

Mobile device unlock authentication has many variations and there have been extensive user-based studies on the security of knowledge-based mobile authentication, including Android graphical unlock patterns [4, 47], PINs [10, 38, 50], as well as using passwords on mobile devices [40]. The conclusion of most of this work is that mobile device users, much like with traditional password selection [18, 28, 39], opt for predictable and easily guessed authenticators. Additionally several physical attacks have been proposed on knowledge-based mobile authentication, such as smudge attacks [6], sensor attacks [7, 12], vision attacks [51], acoustic signals [52], and shoulder surfing [5, 19, 22].

Into this space, LG developed *Knock Codes* as a new mobile authentication system that is designed to combat some of these attacks[1] and provide, per LG's advertising,[2] "perfect security." Knock Codes require a user to recall a pre-selected series of at least 6 and at most 10 knocks[3] (or taps) on a 2×2 quadrant which is displayed upon setup and can be entered with the phone screen on or off. Knock Codes are used less frequently than PINs or Android patterns, but we estimate that there is a large number of Knock Code users, 700,000–2,500,000 in the US alone.

To evaluate the security and usability of Knock Codes, we conducted two online user studies on Amazon Mechanical Turk: a preliminary study ($n = 218$) and a main study ($n = 351$), analyzing a total of 1,138 Knock Codes (436 in the preliminary study and 702 in the main study). In the main study, we evaluated three between-group treatments: a control treatment, where participants used the current 2x2 Knock Code interface; a blocklist treatment, where participants selected 2x2 Knock Codes with some popular codes, as measured in the preliminary study, being disallowed; and finally, a big grid treatment, where participants selected Knock Codes on a larger, 2x3 grid.

We analyzed the selected Knock Codes across treatments and scenarios for security using standard guessing metrics, considering both an offline attacker with unlimited guesses and an online attacker with a limited number of guesses. We find that Knock Codes, as currently deployed, offer *worse* security (51.3 % guessed after 30 attempts) as compared to

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

[1] https://youtu.be/0Imk5JILUc0 (as accessed on June 11, 2020)
[2] https://youtu.be/NRInfu-Lhnc (as accessed on June 11, 2020)
[3] In earlier models, like the 2014 LG G2 [46], where this method first appeared, codes required at least 3 and at most 8. Newer models require 6 to 10 knocks occurring in at least 3 quadrants.

other widely available unlock authentication schemes, e.g., 4-digit PINs (28.0 %), 6-digit PINs (25.4 %) and Android unlock patterns (36.6 %).

While it seems like a straightforward attempt to increase security, an expanded Knock Code grid to 2x3 does not increase, and sometimes worsens, security as compared to 2x2 Knock Codes. After 30 attempts, a simulated attacker correctly guesses *more* 2x3 Knock Codes compared to 2x2 (41 % vs. 37 %). However, blocklisting common Knock Codes (as collected in the preliminary study) is more effective at improving guessing security: only 19 % of these codes were guessed within 30 attempts in simulation.

Overall, participants perceived Knock Codes (across treatments) as secure; however, among all treatments, participants were more hesitant to rate Knock Codes as *more secure* than PINs, Android Unlock Patterns, or alphanumeric passwords. Despite the fact that participants reported Knock Codes as "simple" and "memorable", responses to the SUS [11] questions averaged to "marginal" or "ok" usability (69.8, 68.1, and 64.3, for the control 2x2 treatment, the larger 2x3 treatment, and the blocklist informed 2x2 treatment, respectively). Entry and recall times for Knock Codes were also much slower than what was reported for PINs and Android patterns [27, 38], suggesting lower usability.

Based on the survey and analysis, we make the following contributions and findings:

- We conducted a user study of Knock Codes that considers usability and security analysis.
- We find that Knock Codes, as currently deployed, offer worse security compared to other available methods, both in terms of an online and offline guessing analysis.
- We evaluated different designs for Knock Codes, finding that larger grid sizes offer no benefits (and might actually be less secure) while blocklisting offers promise for improving security.
- We analyzed both qualitative and quantitative feedback of the perceptions of security and usability of Knock Codes, finding that while there are some features of Knock Codes that users like the overall usability was "ok" or "marginal" and the security perceptions were weak compared to other available schemes.

These results indicate that users are interested in new forms of mobile authentication, in particular ones that have options for unlocking with the display off. However, given the usability and security challenges of Knock Codes, *we would not recommend further deployment as currently configured.* For users and developers who wish to continue to use Knock Codes, we would recommend using a blocklist to inform selection as it provides increased security with small effects on usability.

2 Related Work and Background

While Knock Codes have not been broadly studied in the community, other mobile authentication methods have been investigated widely, namely PINs [16, 20], patterns [4, 44, 47], passwords [29, 35], and biometrics [42], as well as adoption rates [27] and authentication times [26].

Research on user-chosen authentication has shown that users tend towards predictable and popular choices, regardless of the authentication method. For instance, Bonneau et al. [10] studied 4-digit PINs and concluded that while 4-digit PINs fare better in user management and choices, guessing the birthday is an effective strategy to access a user's account. Wang et al. showed that 6-digit PINs have marginally better security than 4-digit PINs, yet both English and Chinese users fall into certain patterns when choosing PINs [50].

Markert et al. collected PINs specifically primed for mobile authentication and demonstrated that 6-digit PINs offer little (and perhaps worse) benefit than 4-digit PINs against a throttled attacker. Moreover, non-enforcing blocklists (as deployed by iOS) do not increase security [38]. We use an enforcing blocklist in our data collection, as recommend by Markert et al., and compare Knock Codes to the same RockYou [18] and Amitay [1] datasets used by Wang et al. and Markert et al.

Patterns, or graphical passwords, have been studied in multiple contexts, including smudge attacks [6], shoulder-surfing [5, 19, 23, 37], and user strength perceptions [2, 3]. The selection process has also been studied [4, 44, 47], and in all cases, users choices are predictable. We compare our results to those from Uellenbeck et al. [47] and Aviv et al. [4].

There have also been proposals for incorporating more tactile interaction into mobile authentication. For example, Deyle and Roth suggested using "tactile pins" [21]. Kuber et al. [32–34] studied tactile stimuli: a special mouse with a 4x4 matrix of PINs for selecting a "tactile password." Krombholz et al. considered extra touch interactions through pressure-sensitive touches on iPhones to enhance PINs [31]. However, these user interaction modalities are very different from Knock Codes. Similar to Knock Codes, "personal identifiable chords" (PIC) for smartwatches (a multi-touch PIN entered on a 2x2 grid) have been proposed [41]; these differ in setting (smartwatches) and input type (multi-touch), but the approach could be used to improve Knock Codes by adding multi-touch.

Along with security, usability is an important facet regarding the adoption of authentication methods, thus, quantifying user feedback of such methods is pertinent [43]. Regarding biometric adoption and perceptions, users considered biometrics to be more secure than PINs according to Bhagavatula et al. [8]. In addition, usability factors (such as poor lighting for facial recognition) contributed to users' negative feedback and reluctance to adopt this method versus a more convenient method such as fingerprint recognition. Even with biometrics, this can lead to users choosing weaker forms of knowledge-based authenticators [14].



Figure 1: Screenshot of a video exploring Knock Codes (https://youtu.be/tPYypLe8LEU) where a user enters a Knock Code with the screen off to unlock the phone. This was used to provide instructions and background information to users on Knock Codes.

3 Methodology

We collected data via Amazon Mechanical Turk (MTurk) using an online survey whereby participants were directed to use their mobile devices (checked via the user-agent) to select *two* Knock Codes as well as answer general questions about Knock Codes and their demographics. The two Knock Codes were primed based on different security scenarios, as informed by prior work of Loge et al. [36]. We found some, but minor, differences between Knock Codes in each scenario, similar to Loge et al.'s findings for Android patterns.

We conducted two studies: a preliminary study and a main study which is based on the preliminary study and presented here. The main difference between the two studies is that the main study was focused on participants using mobile devices while the preliminary allowed participants to use traditional computers. From the preliminary study, we were able to refine the main study as well as develop a blocklist of the 30 most common Knock Codes selected in the preliminary study (see Table 3). We provide all study material in the Appendices. Both studies were approved by our institutional review board (IRB).

We found that usage and awareness of Knock Codes are relatively uncommon. Only 3% of our participants in the main study responded that they use Knock Codes, see Table 2 and only 1% reported so in our preliminary study. Despite the low percentages, this suggests that 700K-2.5M users may deploy Knock Codes in the US alone, and we would ideally focus our study just on these users. This is unfortunately not feasible due to the low concentration on MTurk, and as such, we consider a broader set of study participants who may (or may not) be aware of Knock Codes. For those unaware of Knock Codes, our survey would simulate their first experience, as would be the case if they were selecting Knock Codes for the first time on a new device.

Detailed description of the survey. The survey consisted of 12 parts as described below. Please see Appendix A for the exact questions and wording on the pages. We refer to specific questions within a survey part using the page name and question number.

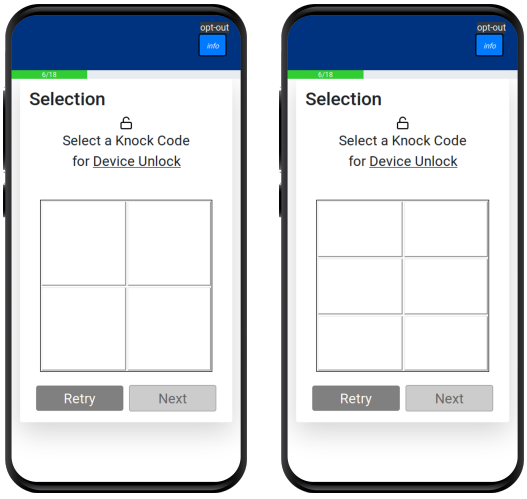
1. *Overview and Informed Consent:* Upon starting the survey, participants were informed about the nature of the research (per the requirements of our IRB), and provided general instructions for proceedings.

2. *Device Usage Questions:* Participants reported on the number of mobile devices (as defined by a smartphone but excluding tablet computers and laptops) they own, the brands they use, and which types of mobile authentication they use on those devices. We use this data, normalized to US census data, to estimate Knock Code usage.

3. *Instructions:* As we could not expect participants to be familiar with Knock Codes, we provided detailed instructions of Knock Codes. This included a GIF animation of a user entering a Knock Code (see Figure 1), a display of the entry screen used later in the survey (see Figure 2), and requirements of Knock Codes (use at least 3 different regions and at least 6 total knocks). We also introduced the size of the grid, 2x2 for participants who were assigned to the control or blocklist treatment, and 2x3 for the group that tested a larger grid. Those in the blocklist treatment were *not* informed of the existence of the blocklist. A detailed description of the treatments is given later in this section.

4. *Practice:* After the instructions, participants could practice selecting a sample Knock Code and familiarize themselves with the interface, before proceeding to the actual Knock Code selection. It was clearly stated that this stage was for practice purposes only. Participants practiced on the appropriate grid size for their treatment and for those in the blocklist treatment, there was no blocklist in place yet, i.e., no indication that a code would or would not be allowed.

5. *Scenario Overview:* In addition to a treatment, each participant was assigned to two scenarios under which they would select Knock Codes for protection. The first of the scenarios was always *Device Unlock*; the other was either *Banking App* or *Shopping Cart*. These scenarios were adapted from prior work of Loge et al. [36] for collecting Android patterns. Participants were made aware of *both* scenarios before proceeding and the order in which they would be asked to select Knock Codes. On this page, we also highlighted that the selected Knock Code will have to be recalled later, hence, participants were asked to "choose something that is secure and memorable."



(a) con-2x2 & bl-2x2 (b) big-2x3

Figure 2: (a) Interface for selecting 2x2 Knock Codes and (b) interface for selecting 2x3 Knock Codes. Both designs mimic the look and feel of LG's Knock Code implementation.



Figure 3: Blocklist warning display, which mimics blocklist warnings as used by iOS for PINs.

6. *Select and Confirm (2x):* Participants were prompted to select a Knock Code for the scenario, and confirm it before proceeding. The respective pages are shown in Figure 2. Participants of the blocklist treatment saw the warning message shown in Figure 3 if any selection was disallowed. Table 3 contains the list of blocklisted codes as collected in the preliminary study.
7. *Selection Feedback (2x):* After selecting and confirming a Knock Code, participants were asked for feedback about their views on the security of their code and any difficulties in selecting a secure and usable code. Data was collected in both Likert agreement and through open answer forms.
8. *Security Prompts:* Now with more familiarity with Knock Codes, participants answered questions about the perceived security of Knock Codes, and also compared it to PINs and Android Unlock Patterns. Participants also provided qualitative feedback on their security likes and dislikes related to Knock Codes in general.
9. *Usability Prompts:* We asked the 10 System Usability Scale questions [11] related to Knock Codes (plus an attention test).
10. *Recall (2x):* Participants were asked to recall their selected Knock Codes. We allowed up to three guesses for each of the scenarios and forwarded participants if they were not able to recall their Knock Code within this limit.
11. *Demographic Questions:* Participants answered basic demographic questions about their age, gender, dominant hand, educational background, and technology background. We also included another attention check question on this page.
12. *Submission:* The survey ended with participants answering an honesty question (i.e., indicated yes/no to "I honestly participated in this survey and followed instructions completely."). Negative responses were removed from the results, however, all participants were compensated for their work.

Treatments. As part of the study, we assigned participants to one of three treatments. In addition to the standard implementation of LG's Knock Code, which we refer to as **control 2x2** or **con-2x2** throughout this paper, we tested two additional ones.

We first include a blocklist treatment (**blocklist informed 2x2** or **bl-2x2**) which differs from the control 2x2 treatment by the fact that we blocklisted 30 Knock Codes. These codes were the most frequently used as measured in the preliminary study (see Table 3). The blocklist warning, shown in cases of a blocklist hit, is depicted in Figure 3 and is a copy of a warning used by Apple on iOS devices to warn users about an insecure PIN choice.

We conjecture that by disallowing participants from selecting these common codes, the Knock Codes they eventually select would be stronger (harder to guess). There is a risk with blocklists as they may increase frustration during the selection process by having to perform selection multiple times. But as setting up an authentication method is a one-time event, we wished to understand if blocklists can improve the security of Knock Codes.

As another method for increasing security, we considered a modification to the Knock Code interface. The **larger 2x3** treatment (**big-2x3**) uses a 2x3 instead of 2x2 grid and provides participants with more options for creating a Knock Code. Theoretically, this increase makes a substantial difference with 72,520,440 possible 2x3 Knock Codes of length 6-to-10, as compared to 1,384,872 2x2 Knock Codes of similar length. The layout is shown in Figure 2b.

We decided to use a 2x3 grid rather than a horizontal extension (3x2) or making a square (3x3) because of the form factor of the phone's screen, which is taller than it is wide. The 2x3 grid offers a natural extension that fits within the form factor of the screen and mirrors the same interface.

Table 1: Overall demographics of the participants from the main study. Note, zero responses are not shown.

		Male		Female		Other		Total	
		No.	%	No.	%	No.	%	No.	%
Age	18 – 24	25	7 %	10	3 %	1	0 %	36	10 %
	25 – 34	131	37 %	64	18 %	2	1 %	197	56 %
	35 – 44	46	13 %	31	9 %	0	0 %	77	22 %
	45 – 54	19	6 %	13	3 %	0	0 %	32	9 %
	55 – 64	2	1 %	6	2 %	0	0 %	8	3 %
	Prefer not to say	0	0 %	0	0 %	1	0 %	1	0 %
Dexterity	Left-handed	31	9 %	14	4 %	0	0 %	45	13 %
	Right-handed	182	52 %	103	29 %	3	1 %	288	82 %
	Ambidextrous	9	3 %	7	2 %	0	0 %	16	5 %
	Prefer not to say	1	0 %	0	0 %	1	0 %	2	0 %
Location	Urban	91	26 %	44	12 %	0	0 %	135	38 %
	Suburban	99	29 %	57	16 %	1	0 %	157	45 %
	Rural	33	9 %	23	7 %	2	0 %	58	17 %
	Prefer not to say	0	0 %	0	0 %	1	0 %	1	0 %
Education	High School	36	10 %	6	2 %	1	0 %	43	12 %
	Some College	45	13 %	25	7 %	0	0 %	70	20 %
	Training	8	3 %	9	3 %	0	0 %	17	6 %
	Associates	22	7 %	17	5 %	1	0 %	40	12 %
	Bachelor's	91	26 %	55	16 %	1	0 %	147	42 %
	Master's	19	6 %	10	2 %	0	0 %	29	8 %
	Professional	1	0 %	1	0 %	0	0 %	2	0 %
	Doctorate	1	0 %	1	0 %	0	0 %	2	0 %
	Prefer not to say	0	0 %	0	0 %	1	0 %	1	0 %
Backgrnd.	Technical	102	30 %	28	8 %	2	0 %	132	38 %
	Non Technical	110	31 %	94	27 %	1	0 %	205	58 %
	Prefer not to say	11	3 %	2	0 %	1	0 %	14	4 %
	Total	223	64 %	124	35 %	4	1 %	351	100 %

Recruitment. The survey was distributed as an Amazon Mechanical Turk task, paying $1.25. On average, it took our participants 8.5 minutes to complete the survey. We ran the survey over the course of two days in June 2019. We recruited 351 participants, each creating two Knock Codes, for a total of 702 selected and confirmed Knock Codes, but also additional Knock Codes that were not confirmed, either due to memorability or the blocklists. We do not consider the practice Knock Codes in our analysis.

The demographics and backgrounds of the participants are listed in Table 1 and 2. As usual for Amazon Mechanical Turk, the participants tended to be younger and predominantly male, but there was diversity in other categories. A number of our participants reported using Knock Codes on their devices as part of their authentication choice. As Knock Codes were a new interface to many participants, our design models the scenario where a user acquires and first uses an LG phone to perform the initial Knock Code set-up.

Estimating US Knock Code Usage. We generalized our participants' device usage and authentication methods based on age and normalized it to the US population using census data [48, 49]. We saw that LG's market share in the US had a range between 8% to 12% among the estimated 285,300,000 smartphone users [17, 45]. Using that, as well as a 95% confidence interval, as our lower and upper bounds, we conclude

Table 2: Answers of the participants from the main study regarding their device usage.

		Male		Female		Other		Total	
		No.	%	No.	%	No.	%	No.	%
No. Devices	One device	144	41 %	86	24 %	2	0 %	232	66 %
	Two devices	61	18 %	34	10 %	1	0 %	96	28 %
	Three devices	14	4 %	4	1 %	0	0 %	18	5 %
	Four or more devices	4	1 %	0	0 %	1	0 %	5	1 %
Device Usage	Apple	23	5 %	13	3 %	0	0 %	36	8 %
	Google	26	6 %	11	2 %	0	0 %	37	8 %
	Huawei	9	2 %	4	1 %	1	0 %	13	3 %
	LG	51	11 %	26	6 %	0	0 %	77	22 %
	Motorola	40	9 %	16	4 %	0	0 %	56	13 %
	Samsung	115	25 %	77	17 %	2	0 %	194	43 %
	ZTE	7	1 %	4	1 %	1	0 %	12	2 %
	Miscellaneous	23	5 %	6	1 %	0	0 %	29	6 %
Authentication Usage	4 digit PIN	121	21 %	67	13 %	2	0 %	190	34 %
	6 digit PIN	19	3 %	10	2 %	1	0 %	30	5 %
	6+ digit PIN	12	2 %	5	1 %	0	0 %	17	3 %
	Android pattern	69	12 %	22	4 %	0	0 %	91	16 %
	Knock Code	9	2 %	4	1 %	0	0 %	13	3 %
	Fingerprint	96	17 %	41	7 %	2	0 %	139	24 %
	Facial Recognition	33	6 %	14	3 %	0	0 %	47	9 %
	Other	0	0 %	1	0 %	0	0 %	1	0 %
	No Authentication	17	2 %	20	4 %	1	0 %	38	6 %

that there are potentially many Knock Code users: 728,693 to 2,567,207 in the US alone. We believe, though, that the actual adoption rate is most likely on the lower end. While this may be an optimistic estimate, it still suggests that there is a substantial number of Knock Code users in the general public, particularly worldwide.

Even though Knock Codes are not as widely adopted as other traditional methods of mobile authentication, it is still important to study user behavior with real-world, deployed authentication systems. In addition, on Google Play many Knock Code apps can be installed on any Android device, thus not limiting Knock Codes to solely LG devices. For instance, the most highly rated Knock Code app on Android, "Knock Lock," boasts more than 1 million installations and claims that it is an innovative lock screen that "will leave intruders baffled" [30]. This app is just one among the plethora of Knock Code knock-off apps that can be found on Google Play, indicating that this authentication method may have a higher adoption rate and influence on mobile authentication systems than appears initially.

4 Limitations

There are a number of limitations associated with our methodology and survey design. One such limitation is that the survey's recall component occurred within a short time frame with minimal distraction tasks. While we can report on short-term memorability of Knock Codes, we cannot report on the memorability over extended time periods, e.g., days.

However, as a mobile unlock authentication method, users must recall their codes frequently, hence short-term recall

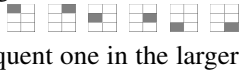
is still relevant. The increased use of biometrics, which reduces the number of knowledge-based recalls, confounds the issue though, and more research would be needed to better understand long-term memorability of Knock Codes.

There are also some limitations on how likely the selected Knock Codes would be real Knock Codes of real users. We believe that the simple interface and the nature of the initial device setup suggest that these Knock Codes would be akin to those used on real devices. Most of our participants were unfamiliar with Knock Codes when taking the survey and so would be new users of LG devices setting up their Knock Code for the first time. It should also be noted that a few participants who do use Knock Codes (both in the preliminary study and main study) reported that they reused their Knock Code in the survey.

Nevertheless, we attempted to address this limitation and thus decided to provide different security scenarios for which participants should create Knock Codes. This technique was used by Loge et al. [36] when collecting Android Unlock Patterns. The motivation is that different scenarios, one always being device unlock, will help users to be more careful about their choices, similar to how they may be during device setup. In analyzing the data (Section 6), we did not find significant differences between the Knock Codes selected under each scenario for the bl-2x2 treatment but did see some differences for the con-2x2 and larger 2x3 treatment.

5 Statistics of Knock Codes

The first step in analyzing Knock Codes is to determine the frequency statistics. Table 4 displays the 30 most frequent patterns, combined, across the scenarios for three treatments of the main study. The frequencies which we observed in the preliminary study are shown in Table 3. The preliminary study codes and the con-2x2 codes have a lot of overlap, with 42.0% of the Knock Codes from the preliminary study appearing in the top-30 most frequent codes in the Control 2x2 treatment. This helps justify using the most frequent preliminary study codes as the basis of the blocklist for the bl-2x2 treatment.

Code frequency. The most common Knock Code in our control dataset is  ($freq = 6.9\,\%$). It starts in the upper left corner, follows a left-to-right sequence, and is repeated until the minimum length of 6 is reached. We observe a similar strategy for the code  ($freq = 4.6\,\%$) which is the most frequent one in the larger 2x3 treatment. However, participants were able to reach the minimum length without repeating the pattern because of the larger grid.

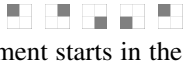
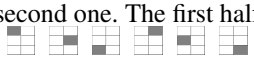
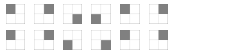
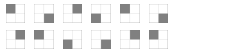
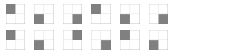
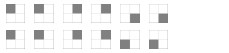
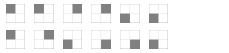
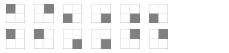
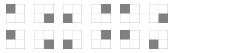
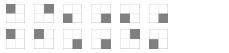
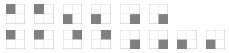
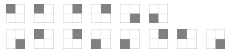
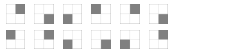
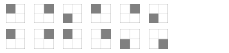
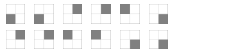
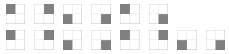
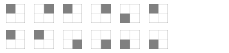
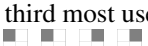
The second most common Knock Code  ($freq = 3.9\,\%$) in the control 2x2 treatment starts in the upper left quadrant, moving clockwise. In contrast to this,  ($freq = 4.2\,\%$), the second most used code in the larger 2x3 treatment, has different attributes: participants proceed diagonally over the grid, going down in a right-left movement for the first diagonal and up in a left-right movement for the second one. The first half of the third most used Knock Code  ($freq = 3.8\,\%$) is identical, yet, it differs at the second diagonal which follows a top-down movement instead of bottom-up.

Table 3: Top 30 most frequent Knock Codes from the preliminary study, which were used as the blocklist in the bl-2x2 treatment of the main study.

Rank	Knock Code	No.	%
1		28	6.4 %
2		25	5.7 %
3		19	4.4 %
4		7	1.6 %
		7	1.6 %
		7	1.6 %
		7	1.6 %
		7	1.6 %
9		6	1.4 %
		6	1.4 %
11		5	1.1 %
		5	1.1 %
		5	1.1 %
		5	1.1 %
		5	1.1 %
		5	1.1 %
		5	1.1 %
18		4	0.9 %
		4	0.9 %
		4	0.9 %
		4	0.9 %
		4	0.9 %
23		3	0.7 %
		3	0.7 %
		3	0.7 %
		3	0.7 %
		3	0.7 %
		3	0.7 %
		3	0.7 %
		3	0.7 %

The third most used Knock Code in the control 2x2 treatment ( , $freq = 3.5\,\%$) pursues a left-to-right sequence again, however, participants used double taps to comply with the required minimum length of 6 knocks.

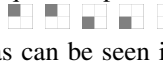
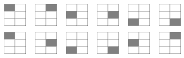
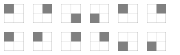
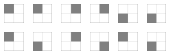
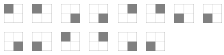
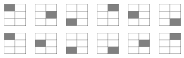
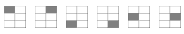
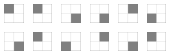
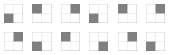
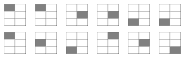
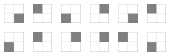
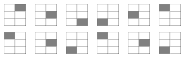
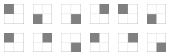
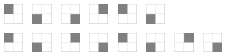
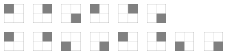
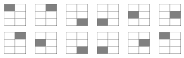
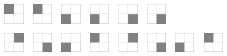
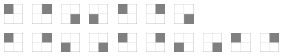
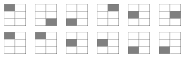
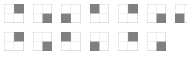
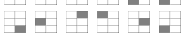
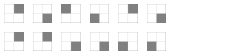
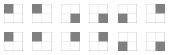
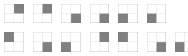
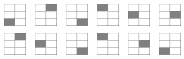
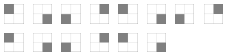
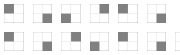
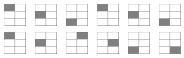
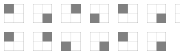
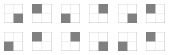
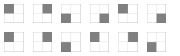
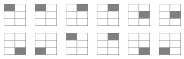
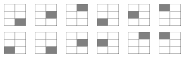
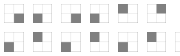
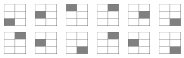
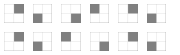
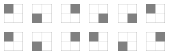
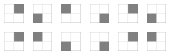
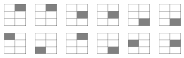
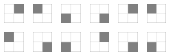
Participants of the blocklist informed 2x2 treatment used this strategy to an even greater extent: the three most used Knock Codes all contain multiple double taps and 51.0 % of all codes created for this treatment include one or more repeated taps. In contrast to this, only 41.0 % of the codes in the control 2x2 treatment and 29.0 % of the codes in the larger 2x3 treatment contain at least one repeated tap. Moreover, the distribution of Knock Codes in the blocklist informed 2x2 treatment is more equal compared to the other two. The most used Knock Code,  , occurs in only 2.6 % of the cases and as can be seen in Table 4 the distribution flattens the fastest.

Table 4: Top 30 most frequent Knock Codes in all three treatments.

Rank	All Control 2x2 Knock Code	No.	%	Rank	All Blocklist 2x2 Knock Code	No.	%	Rank	All Large 2x3 Knock Code	No.	%
1		16	6.9 %	1		6	2.6 %	1		11	4.6 %
2		9	3.9 %	2		5	2.2 %	2		10	4.2 %
3		8	3.5 %			5	2.2 %	3		9	3.8 %
4		6	2.6 %	4		3	1.3 %			9	3.8 %
5		5	2.2 %			3	1.3 %	5		8	3.4 %
		5	2.2 %			3	1.3 %	6		7	2.9 %
6		4	1.7 %			3	1.3 %			6	2.5 %
		4	1.7 %			3	1.3 %	8		5	2.1 %
		4	1.7 %			3	1.3 %			5	2.1 %
		4	1.7 %			3	1.3 %			5	2.1 %
		4	1.7 %			3	1.3 %	11		4	1.7 %
		4	1.7 %			3	1.3 %			4	1.7 %
13		3	1.3 %			3	1.3 %			4	1.7 %
		3	1.3 %			3	1.3 %	14		3	1.3 %
		3	1.3 %			3	1.3 %			3	1.3 %
		3	1.3 %	16		2	0.9 %			3	1.3 %
		3	1.3 %			2	0.9 %			3	1.3 %
		3	1.3 %			2	0.9 %			3	1.3 %
		3	1.3 %			2	0.9 %			3	1.3 %
		3	1.3 %			2	0.9 %			3	1.3 %
		3	1.3 %			2	0.9 %	21		2	0.8 %
		3	1.3 %			2	0.9 %			2	0.8 %
23		2	0.9 %			2	0.9 %			2	0.8 %
		2	0.9 %			2	0.9 %			2	0.8 %
		2	0.9 %			2	0.9 %			2	0.8 %
		2	0.9 %			2	0.9 %			2	0.8 %
		2	0.9 %			2	0.9 %			2	0.8 %
		2	0.9 %			2	0.9 %			2	0.8 %
		2	0.9 %			2	0.9 %			2	0.8 %
		2	0.9 %			2	0.9 %			2	0.8 %

To summarize, the frequencies of the Knock Codes show different characteristics depending on the assigned treatment, suggesting natural, human tendencies in the selection that can be leveraged in predicting and guessing Knock Codes. We take advantage of this observation when guessing codes. Participants in the blocklist informed 2x2 group use more repeated taps whereas codes created for the 2x3 treatment make use of the larger grid and follow directional patterns. Knock Codes created for the control 2x2 depict a mix and follow both strategies equally.

Start/end quadrant frequency. Figure 4 and 5 present the frequency of start and end taps in the Knock Codes. Clearly, there is a strong tendency to begin codes in the upper-left. Similar observations were made for Android Graphical Patterns [47] and is likely due to the left-to-right nature of the English language which is dominant among our participants. The least common starting points in the preliminary study as well as the control and blocklist treatment were in the lower row. In the larger 2x3 treatment, on the other hand, the middle row is used the least often.

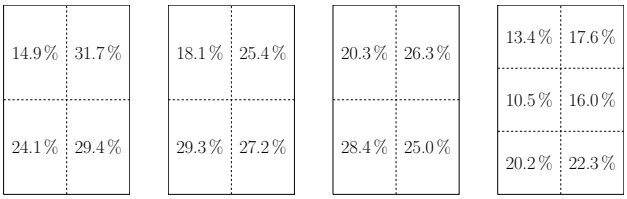
To understand the left/right and up/downshifting of the Knock Codes' start locations we mapped the Cartesian coordinate to each quadrant in the grid, where (-1,1) is the upper left quadrant  , (1,1) is the upper right quadrant  , (-1,-1) is the lower left quadrant  , and (1,-1) is the lower right quadrant . Similarly, in the larger 2x3 treatment, we mapped the coordinates (-1,1), (1,1), (-1,0), (1,0), (-1,-1), and (-1,1) to the grid spaces, scanning left to right, top to bottom. We then computed the average x and y coordinate for the start and end taps, across treatments.

A Shapiro Wilk's test ($p < 0.001$) indicated that the generated frequencies are not normally distributed, so a Mann-Whitney U test was used to identify any initial significance, followed by a posthoc test with Bonferroni correction. We found significant differences between both the control 2x2 and larger 2x3 treatment ($p < 0.001$) as well as blocklist informed 2x2 and larger 2x3 ($p < 0.001$), suggesting that the larger grid size affected how participants chose to start and end their codes.



(a) Preliminary Study (b) Control 2x2 (c) Blocklist 2x2 (d) Large 2x3

Figure 4: Frequency of start quadrants per treatment combined, across all scenarios. More detailed figures with frequencies for every single scenario can be found in the Appendices.



(a) Preliminary Study (b) Control 2x2 (c) Blocklist 2x2 (d) Large 2x3

Figure 5: Frequency of end quadrants per treatment across all scenarios. More detailed figures with frequencies for every single scenario can be found in the Appendices.

Code length. We also analyzed the Knock Codes with respect to length. The average code length was 6.4, 6.5, and 6.2 in each treatment, con-2x2, bl-2x2, and big-2x3, respectively. We observed statistical differences using ANOVA ($f = 11.57, p < 0.001$) between the treatments. In post hoc analysis, using pairwise *t*-test comparison, the difference lies primarily in the longer big-2x3 Knock Codes, which was statistically different from both bl-2x2 ($p < 0.001$) and the con-2x2 ($p < 0.001$). Surprisingly, the larger grid size encouraged slightly shorter Knock Codes. Regardless, the vast majority of Knock Codes were of length 6, which was the median value, or 8, with a few codes of length 10.

6 Security Analysis

We now analyze the security of Knock Codes. We start by describing the threat model which we are considering for the attack. Afterwards, we analyze the security of Knock Codes by using a perfect knowledge metric in Section 6.1 to define an upper bound on generic attack performance. In Section 6.2, we assess the success rate of a simulated attacker to provide a more realistic security estimation.

Threat Model. We consider a generic, non-targeted attacker that attempts to access an arbitrary victim's device by guessing the Knock Code without additional knowledge or previous observations of the victim. A targeted attacker who may know the victim's tendencies or previously observed an entry (e.g., via a shoulder surfing attack) would likely perform better than the generic attacker. A generic attacker, though, provides a lower bound on the scope of attacker performance, and it also provides a clear comparison point to other reported results [4, 10, 38, 47, 50] which use the same threat model.

For the security analysis, we employ two different attacker variations. First is a *perfect knowledge attacker*, which assumes that the attacker has complete knowledge of the frequency order Knock Codes, from most to least frequent. This attack is still generic as the same strategy is assumed for every victim, and it allows one to estimate the security of the Knock Codes as selected by users. See Section 6.1 for more details.

Second, a *simulated attacker* who knows a subset of the Knock Codes and constructs a model based on that observed distribution. The attacker then attempts to guess a set of arbitrary victims' (unknown) Knock Codes. We use a cross-fold validation to mimic the attacker, whereby the attacker trains on a subset of the data and guesses on an unknown test set.

First-Entry 2x2 Codes. Throughout this section, we refer to a *First-Entry 2x2* dataset which contains participants' first entered codes in the control and blocklist treatment. These codes may or may not have been confirmed (i.e., on the confirm entry screen) either due to lack of recall or because of the blocklist. We include this dataset, as it offers the perspective of an ideal user choice for how the authenticator may have been selected in the absence of external influences. As we expected, this dataset is slighter more secure than that of the confirmed control 2x2 codes and offers insights into how users compromise on security to gain more memorable codes.

6.1 Perfect Knowledge Strength Estimations

We consider the guessing strength of Knock Codes against a perfect knowledge attacker as described by Bonneau et al. [9]. A perfect knowledge attack depicts the upper bound for an attack as it assumes that the attacker knows the attacked dataset and always guesses in the ideal order, that is, the Knock Code with the next highest frequency. This approach has been regularly applied to analyzing mobile authentication, such as Android Patterns [4, 44, 47] or PINs [38, 50].

We use two different perfect-knowledge guessability metrics to evaluate Knock Codes, one based on an offline attack model and one based on an online (or throttled) attack model. An offline attack model assumes that the attacker can guess as many times as possible, while an online attack model assumes an attacker with a limited number of attempts. The online attack model better matches the realities of mobile authentication, where users typically have a maximal number of attempts before the device is locked out. The offline attack model, on the other hand, provides a more holistic approach to measuring the security of a set of user-chosen passwords.

Table 5: Comparison of the guessing metrics for a perfect-knowledge attacker between the treatments and other schemes. A comparison between the scenarios is shown in Appendix B.

Dataset	Online Guessing (Success %)			Offline Guessing (bits)			
	λ_3	λ_{10}	λ_{30}	H_∞	$\widetilde{G}_{0.1}$	$\widetilde{G}_{0.2}$	$\widetilde{G}_{0.5}$
All Control 2x2	14.2 %	28.0 %	51.3 %	3.86	4.20	4.79	5.69
All Blocklist 2x2	6.9 %	16.0 %	35.4 %	5.27	5.79	6.03	6.72
All Large 2x3†	12.9 %	31.5 %	53.4 %	4.40	4.53	4.70	5.54
All First-Entry 2x2†	10.8 %	22.8 %	43.1 %	4.40	4.79	5.35	6.19
3x3 Pattern [4]†	8.6 %	19.4 %	36.6 %	4.69	5.21	5.72	6.76
4x4 Pattern [4]†	7.8 %	18.1 %	32.3 %	5.05	5.47	5.92	7.00
4-digit PINs [1]†	9.5 %	17.2 %	28.0 %	4.40	5.14	6.05	7.21
6-digit PINs [50]†	13.4 %	16.8 %	25.4 %	3.10	3.10	6.38	7.32

†: For a fair comparison we downsampled all marked datasets to the size of Control and Blocklist (232 Knock Codes).

For an offline attack metric, we use *partial guessing entropy* or *α-guesswork* ($\widetilde{G}_\alpha$). Partial guessing entropy estimates the amount of guesswork that is needed to guess a fraction α of all codes. The Min-entropy H_∞ depicts a special case as it is only based on the most frequent Knock Code. As an online (or throttled) attack metric, we use *β-success rate*. It essentially measures what fraction of codes would be guessed if the attacker only had β guesses, e.g., λ_3 considers an attack which is limited to 3 guesses.

Table 5 shows the guessing results for our three treatments as well as the combined dataset First-Entry 2x2. As an additional comparison we included datasets from previous studies for Android patterns [4] as well as 4- and 6- digit PINs [1,50]. Because the datasets all differ in size which would influence the results, we downsampled all marked datasets to the size of control 2x2 and blocklist 2x2 (232 entries) and calculated the statistics for the samples. To rule out any sampling bias, we repeated this process 500 times, removed outliers using Tukey fences with $k = 1.5$, and report the median value of the remaining set in Table 5. With a 95 % confidence level the margin of error is lower than 0.3 % for the online guessing and lower than 0.1 bits for the offline case.

Across all comparisons, we find that Knock Codes in the control 2x2 are significantly weaker in terms of their guessability. This means, Knock Codes as they are currently deployed are more guessable than both 4- and 6-digit PINs as well as Android Patterns. When considering the First-Entry dataset, the differences are less distinct, but even in this ideal case, the inferiority of Knock Codes remains.

Surprisingly, increasing the size of the keyspace by enlarging the grid size to 2x3 offers only little security gain. Moreover, in some cases increasing the grid size may even *decrease* security. This is most apparent when considering a throttled attacker. After 10 guesses, 31.5% of the larger 2x3 codes are guessed compared to 28.0 % for the control 2x2 codes. A similar observation can be made after 30 guesses, 53.4 % of larger 2x3 codes are guessed compared to 51.3 % of control 2x2 codes.

Future works needs to examine why larger Knock Codes performed so poorly, but a similar phenomenon was observed by Aviv et al. with increasing Android patterns from 3x3 to 4x4 grid sizes [4]. Aviv et al. conjectured, and we do so here as well, that there may be a false sense of security that the larger set of choices offers, whereby users believe their individual choice matters less in the face of the increased number of possibilities. Analyzing other grid sizes, such as 3x2 or 3x3, would offer additional insight; nevertheless, it is interesting to see that providing more complexity in how to select Knock Codes does not increase the security.

Finally, we observed strong security improvements with the introduction of a blocklist. As compared to the con-2x2, the blocklist cuts the success rate of an attacker within the first 30 attempts by 30 % to 50 % and increases the guesswork by ~1.5 bits when considering an offline attacker. While the blocklist clearly encouraged more diverse choices, it also had the side effect of increasing user frustration and usability, as we describe later in Section 7.

6.2 Simulated Attacker Strength

We are also interested in modeling a more realistic, limited-knowledge attacker that has access to a subset of training data and attempts to guess some test set of unknown data: a *simulated attacker*.

A simulated attacker must model Knock Codes from a training set to predict a test set. We used a three-gram Markov model probability estimator for the likelihood of a given Knock Code, based on the empirical observations in the test set. This is a standard approach when analyzing user chose secrets, e.g., passwords [13,24], PINs [50], or Android Patterns [4,47]. In order to encode the start and end transitions, we defined special symbols for transitions to ending/starting nodes. This can be defined more formally:

$$x = \{x_{-(g-1)}, \ldots, x_{-1}, x_0, x_1, \ldots, x_n, \ldots, x_{n+g-1}\}$$

where x is the Knock Code of length n with first knock x_1, and g is the gram size. If $i \leq 0$ or $i > n$, then this indicates that x_i is a start or end transition state. These extra states are used to capture the early and late transitions taken by a user, for example, for the following Knock Code  , we would produce the following set of tri-grams, where $\perp$ is a start state and $\top$ is an end state: ($\perp$ $\perp$), (), (), (

Table 6: Guessing performance of a simulated attacker.

Dataset	Codes	Blocklist Hits No.	Blocklist Hits %	3 Guesses No.	3 Guesses %	10 Guesses No.	10 Guesses %	30 Guesses No.	30 Guesses %
All Control 2x2	**232**		-	**33**	**14 %**	**44**	**19 %**	**85**	**37 %**
Device Unlock	116		-	20	17 %	28	24 %	42	36 %
Banking App.	56		-	0	0 %	4	7 %	8	14 %
Shopping Cart	60		-	9	15 %	11	18 %	23	38 %
All Blocklist 2x2	**232**	**53**	**23 %**	**9**	**4 %**	**14**	**6 %**	**45**	**19 %**
Device Unlock	116	40	35 %	1	1 %	1	1 %	5	4 %
Banking App.	57	8	14 %	3	5 %	3	5 %	3	5 %
Shopping Cart	59	5	9 %	3	5 %	3	5 %	5	9 %
All Large 2x3	**238**		-	**24**	**10 %**	**62**	**26 %**	**97**	**41 %**
Device Unlock	119		-	6	5 %	37	31 %	44	37 %
Banking App.	63		-	1	2 %	6	10 %	15	23 %
Shopping Cart	56		-	5	9 %	10	18 %	15	27 %
All First-Entry 2x2	**464**		-	**42**	**9 %**	**83**	**18 %**	**127**	**27 %**
Device Unlock	232		-	31	13 %	47	20 %	84	36 %
Banking App.	113		-	5	4 %	16	14 %	27	24 %
Shopping Cart	119		-	12	10 %	20	17 %	29	24 %

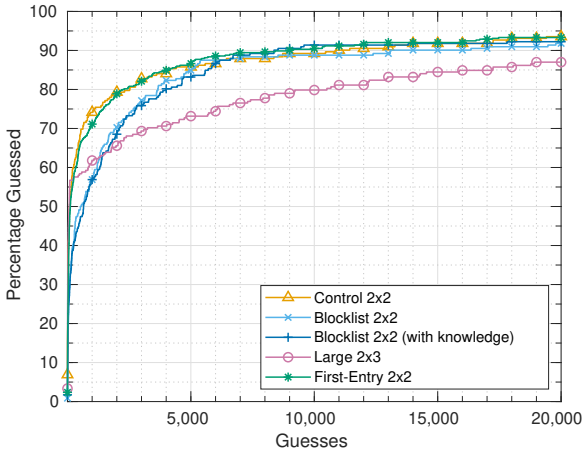
where $P(\cdot)$ is the probability function, $\mathsf{len}(x)$ is the length function, $\mathsf{start}(x)$ is the start function, and $\mathsf{end}(x)$ is the end function. These are our prior probabilities that capture the likelihood of a given length, start quadrant, and end quadrant. The transition probabilities are captured using the conditional probabilities of each transition between each sub-sequence of length g, given the prior state. As not all transitions are represented in our dataset, we used constant smoothing to avoid zero probabilities.

The simulated attackers guessing routine, given a training set, is to (1) create a Markov model of the training data; (2) guess patterns in frequency order of the training set, with ties broken by the likelihood estimation; and (3) guess from a set of additional Knock Codes (not in the training set) sorted based on the likelihood estimation. For (3), we generated a list of all length 6-to-8 Knock Codes for the 2x2 and 2x3 grid sizes, excluding those in our training set that were previously guessed. This accounted for 1,384,872 and 72,520,440 additional 2x2 and 2x3 Knock Codes that could be guessed, respectively. In our blocklist treatment, we assumed the attacker had knowledge of the blocklist.

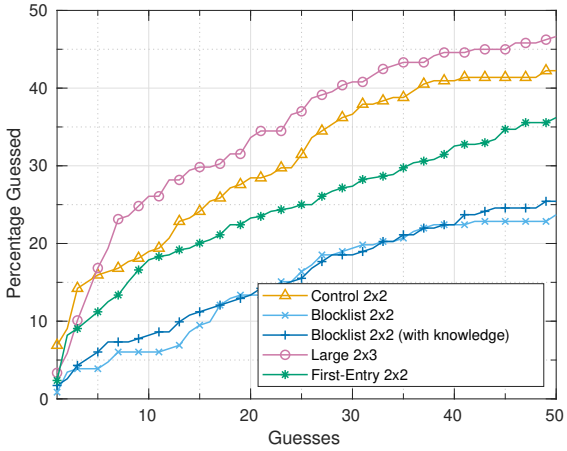
The results of our simulated attacker are presented in Table 6, and a graphical representation is provided in Figure 6. We report on the average of five randomized cross-fold validations. As expected, the simulated attacker performs worse than the perfect-knowledge attacker, but we find similar results comparing across treatments. Notably, the 2x3 Knock Codes offer little, or worse, security while there is marked improvement for the blocklist informed 2x2 Knock Codes.

7 Usability of Knock Codes

In this section, we focus on the usability metrics of Knock Codes. We first report results on the setup and recall times. Afterwards, we will focus on memorability and recall rates within our study, followed by the qualitative and quantitative responses to security and usability prompts.



(a) Guessing performance of an *unthrottled* attacker.



(b) Guessing performance of a *throttled* attacker.

Figure 6: Guessing performance of a simulated attacker against the different treatments based on the numbers of guesses.

7.1 Setup and Recall Times

Table 7 presents the average selection and recall times, as well as the number of attempts, needed to select a Knock Code. Outliers were removed using Tukey fences with $k = 1.5$.

Participants needed on average 16.2 s and 18.4 s to select and confirm a 2x2 and 2x3 Knock Code, respectively. This is faster than the blocklist treatment (22.5 s), where participants also had to make more attempts due to blocklisting (1.5 vs. 1.1 attempts). In comparison, setting up a 4- or 6-digit PIN takes on average only 7.9 and 11.5 seconds respectively [38] which is significantly faster than Knock Codes. While the described discrepancy between Knock Codes and PINs is distinct, the numbers for PINs may be lower since users are presumably more familiar with PINs as compared to Knock Codes. The differences may decrease with increased familiarity with Knock Codes.

Table 7: The average time and number of attempts required for setup and recall. The standard deviation is shown in brackets.

Treatment	Setup Time	Setup Attempts	Setup Time/Attempt	Recall Time	Recall Attempts	Recall Time/Attempt
Control 2x2	16.2 *s* (*7.7 s*)	1.1 (*0.4*)	15.6 *s* (*7.6 s*)	8.8 *s* (*4.5 s*)	1.1 (*0.7*)	7.2 *s* (*2.7 s*)
Blocklist 2x2	22.5 *s* (*13.7 s*)	1.5 (*1.1*)	18.3 *s* (*8.6 s*)	11.3 *s* (*6.7 s*)	1.2 (*0.8*)	7.4 *s* (*2.6 s*)
Large 2x3	18.4 *s* (*11.0 s*)	1.1 (*0.5*)	17.4 *s* (*8.4 s*)	8.4 *s* (*4.1 s*)	1.1 (*0.6*)	7.1 *s* (*2.6 s*)

In terms of the recall, which can be compared to unlocking a smartphone, the 2x2 (7.2 s per attempt) and 2x3 Knock Codes (7.1 s per attempt) are more efficient than Knock Codes selected with a blocklist (7.4 s per attempt). With 1.2 attempts per entry, it took participants 11.3 seconds to enter their Knock Codes for the blocklist treatment. Compared to entering an Android pattern (3.0 s) or a PIN (4.7 s) [27], clear usability issues with Knock Codes emerge as entering them is twice as slow. With greater use of Knock Codes, these differences may decrease, but it is unlikely that Knock Codes will be as efficient to enter as patterns or PINs.

7.2 Memorability

We will now go into more details on the memorability as it depicts an important benchmark for any authentication method. We analyzed the memorability of Knock Codes by looking at the recall rates at the end of the survey. While this is an imperfect measure for the memorability, as the survey took most participants less than 10 minutes to complete, it does speak to potential underlying usability issues, particularly if codes were not properly recalled in this short window.

We separated the recall rates based on each treatment. The con-2x2 treatment participants successfully recalled their codes 88.8 % of the time. The participants with the larger 2x3 grid had higher recall rates of 92.9 %, which may suggest an interesting usability vs. security trade-off as this group chose shorter and also some of the weakest Knock Codes. However, we did not find significant differences between the con-2x2 and big-2x3 recall rates using a χ^2 test. We would expect long term memorability rates to be equally high, but further study would be needed to confirm that conjecture.

The worst recall rate came from participants in the bl-2x2 treatment: 80.6 % successfully recalled their Knock Code, and the result was significantly different from the other two recall rates ($p < 0.0001$ for both comparison tests). This could be attributed to the impact of the blocklist, where participants who hit the blocklist had lower recall rates (66.0 %) than those that did not (84.9 %). Most likely, the blocklist affected users in two ways. First, participants who chose blocklisted codes were forced to select multiple codes until landing on one that was not blocklisted. The average number of blocklisting events per user who hit the blocklist was 1.4. Second, that final Knock Code chosen ended up being more complex (as evident in the prior section), and thus harder to recall. Again, this suggests a clear trade-off between usability and security.

We also analyzed the number of attempts to successfully recall a Knock Code. We found no statistical difference across all treatments between the attempts made in recalling the first or second scenario Knock Code correctly. In the big-2x3 treatment and the con-2x2 treatment, participants took on average 1.1 attempts when recalling a Knock Code correctly, with 3 attempts as the maximum. For the bl-2x2 group, users took on average 1.2 attempts to correctly recall a Knock Code, again having a maximum of 3 attempts. Again, we find bl-2x2‘s result to be significantly different in terms of the number of attempts made in the other treatments ($p < 0.001$ vs. big-2x3 and $p < 0.001$ vs. con-2x2), thus showing that the blocklist has an impact on recalling Knock Codes, even for those participants that eventually correctly do so. It is important to note though, that users had a maximum limit of 3 attempts to recall their code before we considered it “cannot be recalled” for the purpose of expediting the survey.

We also analyzed how participants failed to recall their Knock Codes by calculating the average edit distances between the submitted code and the true code for both recalls attempts, one for each scenario. We determined that there was no statistical difference between the average edit distances among treatments. The average edit distance between correct and incorrect recalls was 3.6, suggesting that when users get a code wrong, they get it wrong by a large margin, as the median length Knock Code is 6.

7.3 User Responses

Users provided their opinions and insights regarding Knock Codes’ usability and security. We coded these free responses using two independent reviewers where disputes in coding were resolved until consensus was reached. The specific codes and their frequencies are presented in the Appendices.

Overall Knock Codes were perceived positively by users, citing that they were “Easy,” “Quick,” and “Hard to Guess.” The uniqueness of Knock Codes also appealed to users who indicated they especially liked the fact that it is a “Discreet” and “Secure” authentication method which can be inconspicuous and hidden from others.

For many of the participants, this was a new method of authentication, and they employed various tactics when choosing their Knock Codes. We observed such strategies in determining memorable yet secure codes. To make the Knock Code more memorable, the majority of users opted to use some sort of “Pattern” or “Variation” that would be “Sim-

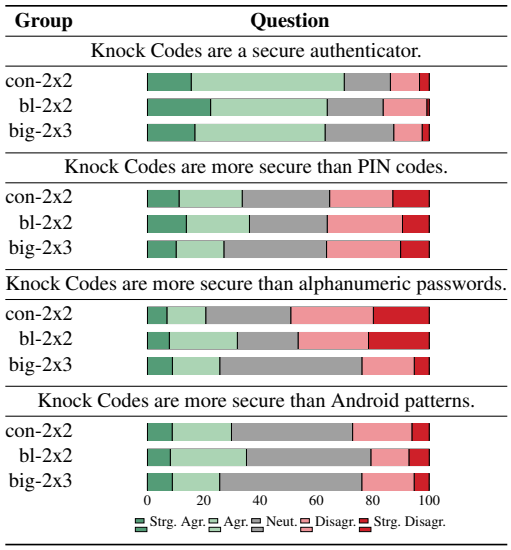


Figure 7: Likert response to comparisons to other mobile authentication methods.

ple." Other techniques users employed include "Directional," "Shape," "Game," and "Repeated." Often, users would create codes based on something "Personal" to them, such as the letter of a word that had meaning to the user.

While many users did not have a specific strategy for security and still focused on making their code "Easy to Remember" as the main priority, others determined that using "All Quadrants" or " Multiple Regions," as well as making the code "Long" or "Random" or being "Unexpected" and "Different" would secure their codes. Making their codes "Hard to Guess" often included attempts to obfuscate the number of clicks and the regions, using speed and potentially unpredictable tactics. Users continued to use similar tactics for memorability to double as security in their Knock Codes, for instance having "Repeated" regions.

Upon comparing Knock Codes with other forms of security, on average users found passwords, PINs, and Android patterns to be more secure than Knock Codes (see Figure 7). Overall, users found Knock Codes adequately secure, i.e., being difficult to hack, resistant to smudge attacks and shoulder surfing. However, they were not completely convinced about Knock Codes' security. Users expressed what they disliked overall, specifically that they found Knock Codes "Hard to Remember" and "Insecure," paving the way for an attacker to easily guess a Knock Code. They also found the interface provided "No improvement" and disliked how it was " Hard to type-in" the Knock Codes.

To have a more general opinion of the overall usability of Knock Codes, we employed the System Usability Scale (SUS). The full Likert responses are found in the Appendices. The average response for the con-2x2 treatment is 69.8, the big-2x3 is 68.1, and the bl-2x2 is 64.3. These scores are generally rated as "ok" or "marginal," with only the control treatment potentially offering some above-average usability.

8 Discussion

As reported above, while most participants offered some positive thoughts, their perception of the security of Knock Codes lagged behind other deployed options, and the SUS values for all schemes were "ok" or "marginal." There was some positive feedback on Knock Codes which suggests an openness to new designs in mobile authentication, particularly to authentication that can be entered while the phone screen is off. There was also increased perceptions of security from targeted attacks, e.g., via shoulder surfing [5, 19, 22]. It is reasonable to view Knock Codes as offering new design concepts that can ultimately improve mobile authentication.

However, we find that Knock Codes, as currently deployed, provide weaker security than other available knowledge-based, mobile unlock methods, such as 4-/6-digit PINs and Android patterns. This is far from the "perfect security" promised by LG's advertisement of Knock Codes. As such, we cannot recommend deploying Knock Codes in their current form as compared to alternative authentication options.

Our results also indicate that a straightforward improvement like increasing the grid size to 2x3 may offer little or worse security. Blocklisting common Knock Codes, on the other hand, does provide more resilience to a throttled attacker, as has been found in password authentication [25] and PINs [38]. Yet, blocklisting runs the risk of increasing user frustration during selection, but since selecting a Knock Code is a one-time event, the usability trade-off of adding a blocklist may be *extremely worthwhile* if Knock Codes continue to be available to LG users. It may also be worthwhile for designers to invest in other methods for improving Knock Code selection, e.g., forcing users to start or end at given quadrants, similar to SysPal [15], or using multi-touch, like chords [41].

9 Conclusion

We performed the first comprehensive user study and security analysis of user-chosen Knock Codes using a three-treatment, between groups study: a control 2x2 treatment, a blocklist 2x2 treatment, and a 2x3 treatment. We find that Knock Codes provide weaker security than other mobile unlock authentication, such as 4-digit PINs, 6-digit PINs, and Android pattern, and that increasing the grid size offered little (or worse) security outcomes, while the addition of a blocklist of common codes substantially increased the security against a throttled attacker. However, Knock Codes suffered in terms of usability, both in terms of entry/recall time and user perception.

Acknowledgments We wish to thank Timothy Forman, Maximilian Golla, and Genry Krichevsky for their assistance. This material is based upon work supported by the National Science Foundation under Grants Nos. 1845300 and 1617584; the research training group "Human Centered Systems Security" (NERD.nrw) sponsored by the state of North Rhine-Westphalia, Germany; and the Army Research Laboratory Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA).

References

[1] Daniel Amitay. Most Common iPhone Passcodes, June 2011. http://danielamitay.com/blog/2011/6/13/most-common-iphone-passcodes, as of June 11, 2020.

[2] Panagiotis Andriotis, Theo Tryfonas, and George Oikonomou. Complexity Metrics and User Strength Perceptions of the Pattern-Lock Graphical Authentication Method. In *Conference on Human Aspects of Information Security, Privacy and Trust*, HAS '14, pages 115–126. Springer, Heraklion, Crete, Greece, June 2014.

[3] Panagiotis Andriotis, Theo Tryfonas, George Oikonomou, and Can Yildiz. A Pilot Study on the Security of Pattern Screen-Lock Methods and Soft Side Channel Attacks. In *ACM Conference on Security and Privacy in Wireless and Mobile Networks*, WiSec '13, pages 1–6, Budapest, Hungary, April 2013. ACM.

[4] Adam J. Aviv, Devon Budzitowski, and Ravi Kuber. Is Bigger Better? Comparing User-Generated Passwords on 3x3 vs. 4x4 Grid Sizes for Android's Pattern Unlock. In *Annual Computer Security Applications Conference*, ACSAC '15, pages 301–310, Los Angeles, California, USA, December 2015. ACM.

[5] Adam J. Aviv, John T. Davin, Flynn Wolf, and Ravi Kuber. Towards Baselines for Shoulder Surfing on Mobile Authentication. In *Annual Conference on Computer Security Applications*, ACSAC '17, pages 486–498, Orlando, Florida, USA, December 2017. ACM.

[6] Adam J. Aviv, Katherine Gibson, Evan Mossop, Matt Blaze, and Jonathan M. Smith. Smudge Attacks on Smartphone Touch Screens. In *USENIX Workshop on Offensive Technologies*, WOOT '10, pages 1–7, Washington, District of Columbia, USA, August 2010. USENIX.

[7] Adam J. Aviv, Benjamin Sapp, Matt Blaze, and Jonathan M. Smith. Practicality of Accelerometer Side Channels on Smartphones. In *Annual Computer Security Applications Conference*, ACSAC '12, pages 41–50, Orlando, Florida, USA, December 2012. ACM.

[8] Chandrasekhar Bhagavatula, Blase Ur, Kevin Iacovino, Su Mon Kywey, Lorrie Faith Cranor, and Marios Savvides. Biometric Authentication on iPhone and Android: Usability, Perceptions, and Influences on Adoption. In *Workshop on Usable Security*, USEC '15, San Diego, California, USA, February 2015. ISOC.

[9] Joseph Bonneau. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In *IEEE Symposium on Security and Privacy*, SP '12, pages 538–552, San Jose, California, USA, May 2012. IEEE.

[10] Joseph Bonneau, Sören Preibusch, and Ross Anderson. A Birthday Present Every Eleven Wallets? The Security of Customer-Chosen Banking PINs. In *Financial Cryptography and Data Security*, FC '12, pages 25–40, Kralendijk, Bonaire, February 2012. Springer.

[11] John Brooke. SUS: A Quick and Dirty Usability Scale. *Usability Evaluation in Industry*, pages 189–194, 1996.

[12] Liang Cai and Hao Chen. TouchLogger: Inferring Keystrokes on Touch Screen from Smartphone Motion. In *Workshop on Hot Topics in Security*, HotSec '11, Berkeley, California, USA, August 2011. USENIX.

[13] Claude Castelluccia, Markus Dürmuth, and Daniele Perito. Adaptive Password-Strength Meters from Markov Models. In *Symposium on Network and Distributed System Security*, NDSS '12, San Diego, California, USA, February 2012. ISOC.

[14] Ivan Cherapau, Ildar Muslukhov, Nalin Asanka, and Konstantin Beznosov. On the Impact of Touch ID on iPhone Passcodes. In *Symposium on Usable Privacy and Security*, SOUPS '15, pages 257–276, Ottawa, Canada, July 2015. USENIX.

[15] Geumhwan Cho, Jun Ho Huh, Junsung Cho, Seongyeol Oh, Youngbae Song, and Hyoungshick Kim. SysPal: System-Guided Pattern Locks for Android. In *IEEE Symposium on Security and Privacy*, SP '17, pages 338–356, San Jose, California, USA, May 2017. IEEE.

[16] N.L. Clarke and S.M. Furnell. Authentication of Users on Mobile Telephones – A Survey of Attitudes and Practices. *Computers & Security*, 24(7):519–527, October 2005.

[17] Comscore, Inc. Top OEMs - Share of Smartphone Subscribers 3 Month Avg. Ending Nov. 2019 vs. 3 Month Avg. Ending Sep. 2019, September 2019. https://www.comscore.com/Insights/Rankings#tab_mobile_smartphone_oems, as of June 11, 2020.

[18] Nik Cubrilovic. RockYou Hack: From Bad To Worse, December 2009. https://techcrunch.com/2009/12/14/rockyou-hack-security-myspace-facebook-passwords/, as of June 11, 2020.

[19] Alexander De Luca, Marian Harbach, Emanuel von Zezschwitz, Max-Emanuel Maurer, Bernhard Ewald Slawik, Heinrich Hussmann, and Matthew Smith. Now You See Me, Now You Don't: Protecting Smartphone Authentication from Shoulder Surfers. In *ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2937–2946, Toronto, Ontario, Canada, April 2014. ACM.

[20] Alexander De Luca, Roman Weiss, and Heiko Drewes. Evaluation of Eye-Gaze Interaction Methods for Security Enhanced PIN-Entry. In *Australasian Computer-Human Interaction Conference*, OZCHI '07, pages 199–202, Adelaide, Australia, November 2007. ACM.

[21] Travis Deyle and Volker Roth. Accessible Authentication via Tactile PIN Entry. *Computer Graphics Topics*, 3:24–26, 2006.

[22] Malin Eiband, Mohamed Khamis, Emanuel von Zezschwitz, Heinrich Hussmann, and Florian Alt. Understanding Shoulder Surfing in the Wild:Stories from Users and Observers. In *ACM Conference on Human Factors in Computing Systems*, CHI '17, pages 4254–4265, Denver, Colorado, USA, May 2017. ACM.

[23] Alain Forget, Sonia Chiasson, and Robert Biddle. Shoulder-Surfing Resistance with Eye-Gaze Entry inCued-Recall Graphical Passwords. In *ACM Conference on Human Factors in Computing Systems*, CHI '10, pages 1107–1110, Atlanta, Georgia, USA, April 2010. ACM.

[24] Maximilian Golla and Markus Dürmuth. On the Accuracy of Password Strength Meters. In *ACM Conference on Computer and Communications Security*, CCS '18, pages 1567–1582, Toronto, Ontario, Canada, October 2018. ACM.

[25] Hana Habib, Jessica Colnago, William Melicher, Blase Ur, Sean M. Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Password Creation in the Presence of Blocklists. In *Workshop on Usable Security*, USEC '17, San Diego, California, USA, February 2017. ISOC.

[26] Marian Harbach, Alexander De Luca, and Serge Egelman. The Anatomy of Smartphone Unlocking: A Field Study of Android Lock Screens. In *ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 4806–4817, San Jose, California, USA, May 2016. ACM.

[27] Marian Harbach, Emanuel von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. 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 '14, pages 213–230, Menlo Park, California, USA, July 2014. USENIX.

[28] Patrick Kelley, Saranga Kom, Michelle L. Mazurek, Rich Shay, Tim Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Julio López. Guess Again (and Again and Again): Measuring Password Strength by Simulating Password-Cracking Algorithms. In *IEEE Symposium on Security and Privacy*, SP '12, pages 523–537, San Jose, California, USA, May 2012. IEEE.

[29] Daniel V. Klein. "Foiling the Cracker": A Survey Of, and Improvements To, Password Security. In *UNIX Security Workshop*, UNIX '90, pages 5–14, Portland, Oregon, USA, August 1990. USENIX.

[30] Knock Lock. Knock Lock Screen - Applock, 2020. https://play.google.com/store/apps/details?id=com.knocklock.applock&hl=en_US, as of June 11, 2020.

[31] Katharina Krombholz, Thomas Hupperich, and Thorsten Holz. Use the Force: Evaluating Force-Sensitive Authentication for Mobile Devices. In *Symposium on Usable Privacy and Security*, SOUPS '16, pages 207–219, Denver, Colorado, USA, July 2016. USENIX.

[32] Ravi Kuber and Shiva Sharma. Toward Tactile Authentication for Blind Users. In *ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '10, pages 289–290, Orlando, Florida, USA, October 2010. ACM.

[33] Ravi Kuber and Wai Yu. Feasibility Study of Tactile-Based Authentication. *International Journal of Human-Computer Studies*, 68(3):158–181, March 2010.

[34] Ravi Kuber and Wai Yu. Toward Tactile Authentication for Blind Users. In *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*, EuroHaptics '10', pages 314–319, Amsterdam, Netherlands, July 2010. Springer.

[35] William C. Lindsey and Chak Ming Chie. A Survey of Digital Phase-Locked Loops. *Proceedings of the IEEE*, 69(4):410–431, April 1981.

[36] Marte Løge, Markus Dürmuth, and Lillian Røstad. On User Choice for Android Unlock Patterns. In *European Workshop on Usable Security*, EuroUSEC '16, Darmstadt, Germany, July 2016. ISOC.

[37] Shushuang Man, Dawei Hong, and Manton Matthews. A Shoulder-Surfing Resistant Graphical Password Scheme – WIW. In *International Conference on Security and Management*, SAM '03, pages 105–111, Las Vegas, Nevada, USA, June 2003. CSREA Press.

[38] Philipp Markert, Daniel V. Bailey, Maximilian Golla, Markus Dürmuth, and Adam J. Aviv. This PIN Can Be Easily Guessed: Analyzing the Security of Smartphone Unlock PINs. In *IEEE Symposium on Security and Privacy*, SP '20, pages 1525–1542, San Francisco, California, USA, May 2020. IEEE.

[39] Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur. Measuring Password Guessability for an Entire University. In *Conference on Computer and Communications Security*, CCS '13, pages 173–186, Berlin, Germany, October 2013. ACM.

[40] William Melicher, Darya Kurilova, Sean M. Segreti, Pranshu Kalvani, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L. Mazurek. Usability and Security of Text Passwords on Mobile Devices. In *ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 527–539, Santa Clara, California, USA, May 2016. ACM.

[41] Ian Oakley, Jun Ho Huh, Junsung Cho, Geumhwan Cho, Rasel Islam, and Hyoungshick Kim. The Personal Identification Chord: A Four Button Authentication System for Smartwatches. In *ACM Asia Conference on Computer and Communications Security*, ASIA CCS '18, pages 75–87, Incheon, Republic of Kore, June 2018. ACM.

[42] Salil Prabhakar, Sharath Pankanti, and Anil K Jain. Biometric Recognition: Security and Privacy Concerns. *IEEE Security & Privacy*, 1(2):63–69, March 2003.

[43] Florian Schaub, Ruben Deyhle, and Michael Weber. Password Entry Usability and Shoulder Surfing Susceptibility on Different Smartphone Platforms. In *International Conference on Mobile and Ubiquitous Multimedia*, MUM '12, pages 13:1–13:10, Ulm, Germany, December 2012. ACM.

[44] Youngbae Song, Geumhwan Cho, Seongyeol Oh, Hyoungshick Kim, and Jun Ho Huh. On the Effectiveness of Pattern Lock Strength Meters: Measuring the Strength of Real World Pattern Locks. In *ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2343–2352, Seoul, Republic of Korea, April 2015. ACM.

[45] Team Counterpoint. US Smartphone Market Share: By Quarter, November 2019. https://www.counterpointresearch.com/us-market-smartphone-share/, as of June 11, 2020.

[46] Kevin C. Tofel. LG G2 and G Flex Phones Getting the Knock Code Wake and Unlock Feature, March 2014. https://gigaom.com/2014/03/25/lg-g2-and-g-flex-phones-getting-the-knock-code-wake-and-unlock-feature/, as of June 11, 2020.

[47] Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. Quantifying the Security of Graphical Passwords: The Case of Android Unlock Patterns. In *ACM Conference on Computer and Communications Security*, CCS '13, pages 161–172, Berlin, Germany, November 2013. ACM.

[48] U.S. Census Bureau, Population Division. Annual Estimates of the Resident Population by Single Year of Age and Sex for the United States: April 1, 2010 to July 1, 2018 , 2018 Population Estimates, June 2019. https://factfinder.census.gov/bkmk/table/1.0/en/PEP/2018/PEPSYASEXN?#, as of June 11, 2020.

[49] U.S. Census Bureau, Population Division. Annual Estimates of the Resident Population for Selected Age Groups by Sex for the United States, States, Counties, and Puerto Rico Commonwealth and Municipios: April 1, 2010 to July 1, 2018, June 2019. https://factfinder.census.gov/bkmk/table/1.0/en/PEP/2018/PEPAGESEX?#, as of June 11, 2020.

[50] Ding Wang, Qianchen Gu, Xinyi Huang, and Ping Wang. Understanding Human-Chosen PINs: Characteristics, Distribution and Security. In *ACM Asia Conference on Computer and Communications Security*, ASIA CCS '17, pages 372–385, Abu Dhabi, United Arab Emirates, April 2017. ACM.

[51] Guixin Ye, Zhanyong Tang, Dingyi Fang, Xiaojiang Chen, Kwang In Kim, Ben Taylor, and Zheng Wang. Cracking Android Pattern Lock in Five Attempts. In *Symposium on Network and Distributed System Security*, NDSS '17, San Diego, California, USA, February 2017. ISOC.

[52] Man Zhou, Qian Wang, Jingxiao Yang, Qi Li, Feng Xiao, Zhibo Wang, and Xiaofen Chen. PatternListener: Cracking Android Pattern Lock Using Acoustic Signals. In *ACM Conference on Computer and Communications Security*, CCS '18, pages 1775–1787, Toronto, Ontario, Canada, October 2018. ACM.

APPENDICES

A Survey Material

A.1 Main Study

1. Device Usage Questions

When referring to "mobile devices" throughout this survey, consider these to include smartphones and tablet computers, such as iPhone and Android phones and tablets. Traditional laptop computers, two-in-one computers, like the Microsoft Surface, or e-readers, like the Amazon Kindle, are not considered mobile devices for the purposes of this survey.

1. How many mobile devices do you use regularly? (Including phones and tablets, but excluding laptops)
   ○ 0 ○ 1 ○ 2 ○ 3 ○ 4+ ○ Prefer not to say

2. What brand of smartphone do you use? (Select all that apply)
   □ Apple □ Samsung □ LG □ Google (Pixel/Nexus)
   □ Motorola □ ZTE □ I do not own a smartphone
   □ Other: _____

3. Select "No" as the answer to this questions:
   ○ Yes ○ No ○ Sometimes ○ Always

4. Which method(s) do you use to lock your mobile device(s)?(Select all that apply)
   □ 4-digit PIN □ 6-digit PIN □ PIN of other length
   □ Android Graphical Pattern □ LG Knock Codes
   □ Fingerprint □ Face □ None □ Other: _____

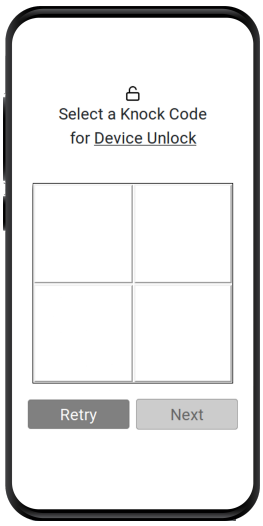
Where indicated, the text and the graphics on the following pages changed depending on the assigned grid size.

2. What are Knock Codes?

Knock Codes are an authentication method used to unlock your smartphone, much like a PIN. To unlock the phone, the user enters their knock Code by tapping different regions (or quadrants) of a [2x2|2x3] grid on the smartphone display. The grid may or may not be displayed at the time of entry, for example, below is a video of someone entering a Knock Code without a grid displayed.



As part of this survey, you will be asked to select your own Knock Codes using an on-screen approximation of a smartphone. You will enter your codes by clicking on different regions of a [2x2|2x3] grid with your mouse. Below is an image of the [2x2|2x3] grid and smartphone approximation.



con-2x2 & bl-2x2



big-2x3

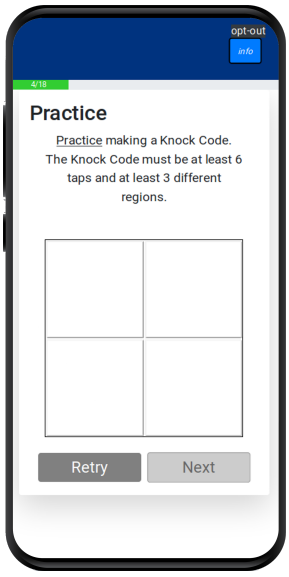
There are some rules! When selecting a Knock Code it must:

1. Use *at least* 3 regions of the grid.
2. Use *at least* 6 total knocks.

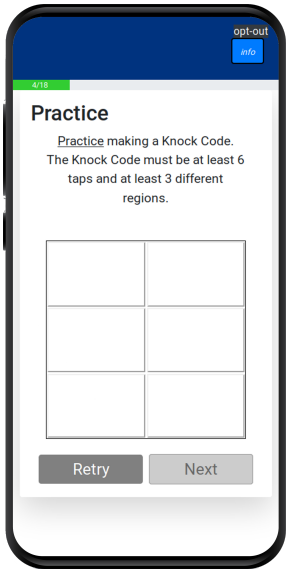
On the next page, you will have a chance to practice entering Knock Codes after which you will proceed with the rest of the survey.

Participants performed a practice run of using the interface. After completion, they were given the option to continue.

3. Practice



con-2x2 & bl-2x2



big-2x3

4. Scenarios
For the remainder of this survey, you will be asked to create Knock Codes for different scenarios.
Importantly, you will need to recall these codes later. So choose something that is secure and memorable. However, we ask that you DO NOT write down your codes or use other aids to help you remember.
I understand that I should not write down my codes or use other aids to assist in the survey:
∘ I understand

You will be asked to create Knock-Knock Codes for the following scenarios.

All participants were assigned to Device Unlock, and then one of either Banking App or Shopping Cart.

🔓 - *Device Unlock* Create a Knock Code you would use to **unlock your smartphone or tablet**.
💼 - *Banking App* Create a Knock Code you would use to secure access to your **mobile banking application**.
🛒 - *Shopping Cart* Create a Knock Code you would use to protect your **Amazon shopping cart**.

I understand that I should not write down my codes or use other aids to assist in the survey:
∘ I understand

Steps 5, 6, and 7 were done twice. First for the Device Unlock, then for the banking or shopping scenario.

5. Selection
Select a Knock Code for [SCENARIO]

6. Confirmation
Confirm the Knock Code for [SCENARIO]

7. Thinking about the Knock Code you just chose, answer the following questions.

1. I feel this Knock Code provides adequate security for this scenario.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree
2. It was difficult to choose this Knock Code.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree
3. What strategy did you use to make your code **more secure**?
   Answer: _____
4. What strategy did you use to make your code **more memorable**?
   Answer: _____

8. Please answer the following questions/prompts.
Select your agreement/disagreement with the following statements

1. Knock Codes are a secure authenticator.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree
2. Knock Codes are more secure than PIN codes.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree
   ∘ Do not know what a PIN code is
3. Knock Codes are more secure than alphanumeric passwords.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree
   ∘ Do not know what an alphanumeric passwords is
4. Knock Codes are more secure than Android patterns.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree
   ∘ Do not know what an Android pattern is

   Provide general feedback on the following questions
5. What are some aspects you **like** about Knock Codes? (use N/A if you do not wish to answer)
   Answer: _____
6. What are some aspects you **do not like** about Knock Codes? (use N/A if you do not wish to answer)
   Answer: _____

9. Please answer the following questions/prompts.
Select your agreement/disagreement with the following statements

1. I think that I would like to use Knock Codes frequently.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree
2. I found Knock Codes unnecessarily complex.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree
3. I thought Knock Codes were easy to use.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree
4. I think that I would need the support of a technical person to be able to use Knock Codes.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree
5. I found the various functions in Knock Codes were well integrated.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree

6. I thought there was too much inconsistency in Knock Codes.
   ○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree

7. I would imagine that most people would learn to use Knock Codes very quickly.
   ○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree

8. Select Agree as the answer to this question.
   ○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree

9. I found Knock Codes very cumbersome to use.
   ○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree

10. I felt very confident using Knock Codes.
    ○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree

11. I needed to learn a lot of things before I could get going with Knock Codes.
    ○ Strongly Agree ○ Agree ○ Neither agree nor disagree ○ Disagree ○ Strongly disagree

10. Recall
Recall your Knock Code for [SCENARIO]

11. Demographic Questions

1. What is your age range:
   ○ 18-24 ○ 25-29 ○ 30-34 ○ 35-39 ○ 40-44 ○ 45-49 ○ 50-54 ○ 55-59 ○ 60-64 ○ 65+ ○ Prefer not to say

2. With what gender do you identify:
   ○ Male ○ Female ○ Non-Binary/Third Gender ○ Not described here ○ Prefer not to say

3. What is your dominant hand?
   ○ Left handed ○ Right handed ○ Ambidextrous ○ Prefer not to say

4. Where you live is best described as:
   ○ Urban ○ Suburban ○ Rural ○ Prefer not to say

5. What is the highest degree or level of school you have completed?
   ○ Some high school ○ High school ○ Some college ○ Trade, technical, or vocational training ○ Associate's Degree ○ Bachelor's Degree ○ Master's Degree ○ Professional Degree ○ Doctorate ○ Prefer not to say

6. Which of the following best describes your educational background or job field?
   ○ I have an education in, or work in, the field of computer science, computer engineering or IT.
   ○ I do not have an education in, nor do I work in, the field of computer science, computer engineering, or IT.
   ○ Prefer not to say

Please indicate if you've honestly participated in this survey and followed instructions completely. You will not be penalized/rejected for indicating "No" but your data may not be included in the analysis:
○ Yes ○ No

A.2 Preliminary Study

1. Demographic Questions

1. What is your age range:
   ○ 18-24 ○ 25-29 ○ 30-34 ○ 35-39 ○ 40-44 ○ 45-49 ○ 50-54 ○ 55-59 ○ 60-64 ○ 65+ ○ Prefer not to say

2. With what gender do you identify:
   ○ Male ○ Female ○ Non-Binary/Third Gender ○ Not described here ○ Prefer not to say

3. What is your dominant hand?
   ○ Left handed ○ Right handed ○ Ambidextrous ○ Prefer not to say

4. Where you live is best described as:
   ○ Urban ○ Suburban ○ Rural ○ Prefer not to say

5. What is the highest degree or level of school you have completed?
   ○ Some high school ○ High school ○ Some college ○ Trade, technical, or vocational training ○ Associate's Degree ○ Bachelor's Degree ○ Master's Degree ○ Professional Degree ○ Doctorate ○ Prefer not to say

6. Which of the following best describes your educational background or job field?
   ○ I have an education in, or work in, the field of computer science, computer engineering or IT.
   ○ I do not have an education in, nor do I work in, the field of computer science, computer engineering, or IT.
   ○ Prefer not to say

2. Device Usage Questions
When referring to "mobile devices" throughout this survey, consider these to include smartphones and tablet computers, such as iPhone and Android phones and tablets. Traditional laptop computers, two-in-one computers, like the Microsoft Surface, or e-readers, like the Amazon Kindle, are not considered mobile devices for the purposes of this survey.

1. How many mobile devices do you use regularly? (Including phones and tablets, but excluding laptops)
   ○ 0 ○ 1 ○ 2 ○ 3 ○ 4+

2. What brand of smartphone do you use? (Select all that apply)
   □ Apple □ Samsung □ LG □ Google (Pixel/Nexus)
   □ Motorola □ ZTE □ Other: _____

3. Select "No" as the answer to this questions:
   ○ Yes ○ No ○ Sometimes ○ Always

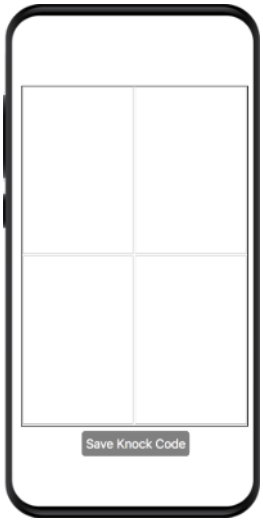
4. Which method(s) do you use to lock your mobile device(s)?(Select all that apply)
   □ 4-digit PIN □ 6-digit PIN □ PIN of other length
   □ Android Graphical Pattern □ LG Knock Codes
   □ Fingerprint □ Face □ Other: _____

3. What are Knock Codes?

Knock Codes are an authentication method used to unlock your smartphone, much like a PIN. To unlock the phone, the user enters their knock Code by tapping different regions (or quadrants) of a 2x2 grid on the smartphone display. The grid may or may not be displayed at the time of entry, for example, below is a video of someone entering a Knock Code without a grid displayed.



As part of this survey, you will be asked to select your own Knock Codes using an on-screen approximation of a smartphone. You will enter your codes by clicking on different regions of a 2x2 grid with your mouse. Below is an image of the 2x2 grid and smartphone approximation.



There are some rules! When selecting a Knock Code it must:

1. Use *at least* 3 regions of the grid.
2. Use *at least* 6 total knocks.

On the next page, you will have a chance to practice entering Knock Codes after which you will proceed with the rest of the survey.

4. Practice

Participants performed a practice run of using the interface. After completion, they were given the option to continue.

5. Scenarios

For the remainder of this survey, you will be asked to create Knock Codes for different scenarios.

Importantly, you will need to recall these codes later. So choose something that is secure and memorable. However, we ask that you DO NOT write down your codes or use other aids to help you remember.

I understand that I should not write down my codes or use other aids to assist in the survey:
○ I understand

You will be asked to create Knock-Knock Codes for the following scenarios.

All participants created Device Unlock, and then one of either Banking App or Shopping Cart. The order was randomized.

- *Device Unlock* Create a Knock Code you would use to **unlock your smartphone or tablet**.
- *Banking App* Create a Knock Code you would use to secure access to your **mobile banking application**.
- *Shopping Cart* Create a Knock Code you would use to protect your **Amazon shopping cart**.

I understand that I should not write down my codes or use other aids to assist in the survey:
○ I understand

Steps 5, 6, and 7 were done twice. For the Device Unlock and for the banking or shopping scenario. The order was randomized.

6. Selection

Select a Knock Code for [SCENARIO]

7. Confirmation

Confirm the Knock Code for [SCENARIO]

8. Thinking about the Knock Code you just chose, answer the following questions.

1. I feel this Knock Code provides adequate security for this scenario.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree

2. It was difficult to choose this Knock Code.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree

3. What strategy did you use to make your code **more secure**?
   Answer: _____

4. What strategy did you use to make your code **more memorable**?
   Answer: _____

9. Please answer the following questions/prompts.
Select your agreement/disagreement with the following statements

1. Knock Codes are a secure authenticator.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree

2. Knock Codes are more secure than PIN codes.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree
   ∘ Do not know what a PIN code is

3. Knock Codes are more secure than alphanumeric passwords.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree
   ∘ Do not know what an alphanumeric passwords is

4. Knock Codes are more secure than Android patterns.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree
   ∘ Do not know what an Android pattern is

   Provide general feedback on the following questions

5. What are some aspects you **like** about Knock Codes? (use N/A if you do not wish to answer)
   Answer: _____

6. What are some aspects you **do not like** about Knock Codes? (use N/A if you do not wish to answer)
   Answer: _____

10. Please answer the following questions/prompts.
Select your agreement/disagreement with the following statements

1. I would like to use Knock Codes frequently.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree

2. Knock Codes are unnecessarily complex.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree

3. Knock Codes are easy to use.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree

4. I would need the support of a technical person to be able to use Knock Codes.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree

5. I would make a lot of mistakes if I were to use Knock Codes.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree

6. Most people would learn to use Knock Codes very quickly.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree

7. Select Agree as the answer to this question.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree

8. I found Knock Codes very cumbersome to use.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree

9. I would neet to practice Knock Codes more before I could use them.
   ∘ Strongly Agree ∘ Agree ∘ Neither agree nor disagree
   ∘ Disagree ∘ Strongly disagree

11. Recall
Recall your Knock Code for [SCENARIO]

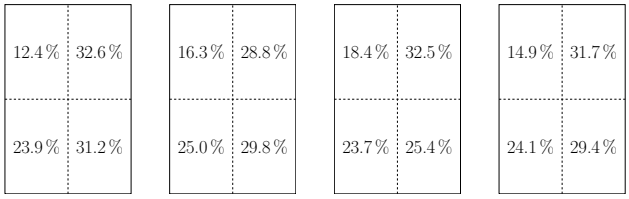
Please indicate if you've honestly participated in this survey and followed instructions completely. You will not be penalized/rejected for indicating "No" but your data may not be included in the analysis:
∘ Yes ∘ No

B Additional Figures & Tables



(a) Device Unlock (b) Shopping Cart (c) Banking Application (d) Overall

Figure 8: Frequency of start quadrants per scenario in the preliminary study.



(a) Device Unlock (b) Shopping Cart (c) Banking Application (d) Overall

Figure 9: Frequency of end quadrants per scenario in the preliminary study.



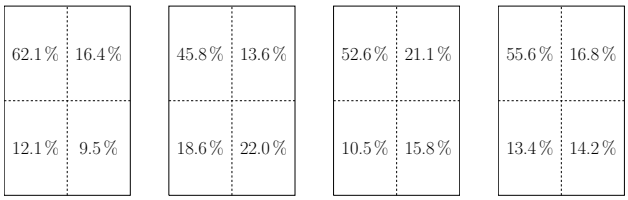
(a) Device Unlock (b) Shopping Cart (c) Banking Application (d) Overall

Figure 10: Frequency of start quadrants per scenario for the control treatment.



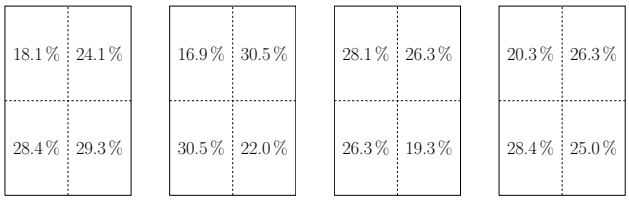
(a) Device Unlock (b) Shopping Cart (c) Banking Application (d) Overall

Figure 11: Frequency of end quadrants per scenario for the control treatment.



(a) Device Unlock (b) Shopping Cart (c) Banking Application (d) Overall

Figure 12: Frequency of start quadrants per scenario for the blocklist treatment.



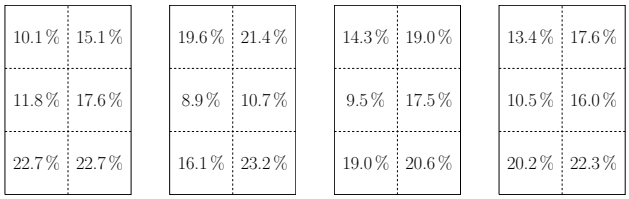
(a) Device Unlock (b) Shopping Cart (c) Banking Application (d) Overall

Figure 13: Frequency of end quadrants per scenario for the blocklist treatment.



(a) Device Unlock (b) Shopping Cart (c) Banking Application (d) Overall

Figure 14: Frequency of start quadrants per scenario for the big treatment.



(a) Device Unlock (b) Shopping Cart (c) Banking Application (d) Overall

Figure 15: Frequency of end quadrants per scenario for the big treatment.

Table 8: Comparison of the guessing metrics for a perfect-knowledge attacker between the scenarios.

	Dataset	Online Guessing (Success %)			Offline Guessing (bits)			
		λ_3	λ_{10}	λ_{30}	H_∞	$\widetilde{G}_{0.1}$	$\widetilde{G}_{0.2}$	$\widetilde{G}_{0.5}$
Control	Device Unlock*	17.9 %	37.5 %	73.2 %	3.81	3.81	4.10	4.93
	Banking App.	10.7 %	30.4 %	66.1 %	4.81	4.81	4.81	5.31
	Shopping Cart*	21.4 %	42.9 %	78.6 %	2.81	2.81	3.75	4.63
Blocklist	Device Unlock*	10.7 %	25.0 %	60.7 %	4.81	4.81	5.19	5.53
	Banking App.*	8.9 %	21.4 %	57.1 %	4.22	5.20	5.52	5.67
	Shopping Cart*	12.5 %	26.8 %	62.5 %	4.22	4.58	4.99	5.45
Large	Device Unlock*	17.9 %	41.1 %	76.8 %	3.81	3.99	4.20	4.79
	Banking App.*	12.5 %	32.1 %	67.9 %	4.22	4.58	4.68	5.21
	Shopping Cart	16.1 %	38.0 %	73.2 %	3.81	3.99	4.40	4.96
1st-Entry	Device Unlock*	16.1 %	33.9 %	69.6 %	3.81	3.99	4.40	5.12
	Banking App.*	12.5 %	28.6 %	64.3 %	4.22	4.58	4.81	5.38
	Shopping Cart.*	16.1 %	33.9 %	69.6 %	3.81	3.99	4.40	5.12

∗: For a fair comparison we downsampled all marked datasets to the size of the smallest datasets (56 Knock Codes).

Table 9: Qualitative codebook from post selection usability and security response.

Question	Code	Freq.	Description	Participant Sample
Security	RANDOM	78	Randomized use of quadrants and taps	"I tried to use random blocks to make it harder to guess."
	EASY TO REMEMBER	70	Prioritized memorability over security	"I was more concerned with it being easy to remember than security."
	LONG	56	Made codes longer as a means of security	"I tried to lengthen it to make it harder to crack."
	ALL QUADRANTS	52	Used all quadrants in the provided grid	"Using all the squares on all of the regions"
	UNEXPECTED	44	Avoided predictable patterns	"I tried to make it slightly more unpredictable than I normally would."
	NONE	42	Did not use any strategy for security	"Since this is not for my device I did not try to make it that secure. If it were my device I would write it down and it would be extensive."
	MULTIPLE QUADRANTS	40	Used a variety of quadrants, not necessarily all	"I tried to use multiple squares more than once to make it more secure."
	HARD TO GUESS	38	Chose a code that is difficult to guess	"Something I didn't think anyone could guess."
	DIFFERENT	37	Using a different code than the first one	"I needed it to be drastically different then [sic] the first code."
Memorability	PATTERN	104	Visualized a sequence or pattern	"Not overly random but three blocks of two patterns"
	SIMPLE	100	Used simple methods	"I used something that wasn't to [sic] complicated"
	EASY TO REMEMBER	77	Focused on overall memorability	"Something easy for me to remember but hard for someone else"
	DIRECTIONAL	76	Went in a specific sequence or order	"I used a specific direction as my way to remember like opening a box or lifting a lid."
	REPEATED	55	Tapped same quads multiple times	"I started at the top left quadrant and went clockwise."
	PERSONAL	52	Associated code with something personal	"I assigned numbers to the quadrants and input a date I'd remember."
	NONE	51	Had no strategy	"Didn't use one."
	VARIATION	40	Altered previous codes	"I used a combination that was similar to my other code but with a Twist."
	SHAPE	38	Followed a specific shape	"I patterned it off of a shape I would remember. In this case it was an underlined x."
	GAME	18	Used or made a game out of the sequence	"I tried to imagine a song pattern like Simon says."
Like	EASY	75	Found usability to be simple/straightforward	"Simple to input doesn't need much screen confirmation."
	HARD TO GUESS	42	Considered it a complex authentication	"I like how you can switch the codes up to many different patterns. It really makes it harder for people to guess what it is."
	DISCREET	40	Liked that it was/can be hidden and discrete	"You can be surreptitious and lock or unlock things without seeming like you are."
	QUICK	39	Found it to be efficient and quick	"It seems very convenient it can be quick and it gets old typing in my pin so much."
	FUN	32	Found it fun to use	"I like that they are unique and I like entering them it is enjoyable."
Dislike	HARD TO REMEMBER	124	Found it difficult to recall codes	"It's seems hard to remember the different patterns"
	INSECURE	90	Found it to be a less complex authentication	"Same thing as a pin without the numbers and with less combination possibilities."
	HARD TO TYPE	19	Found it difficult to input	"I could easily forget or tap the wrong location especially if there is no grid. Also it doesnt seem as fast as using a pattern to unlock like I currently do.."
	NONE	16	Had no issues	"Can't think of anything I overly dislike."
	NOT AN IMPROVEMENT	7	Considered it not better than other existing authentication methods	"There is absolutely no reason to use them for me or most people. They are hard to remember and not any different from a pin code."

Security: What strategy did you use to make your code **more memorable**?
Memorability: What strategy did you use to make your code **more secure**?
Like: What are some aspects you **like** about Knock Codes?
Dislike: What are some aspects you **do not like** about Knock Codes?

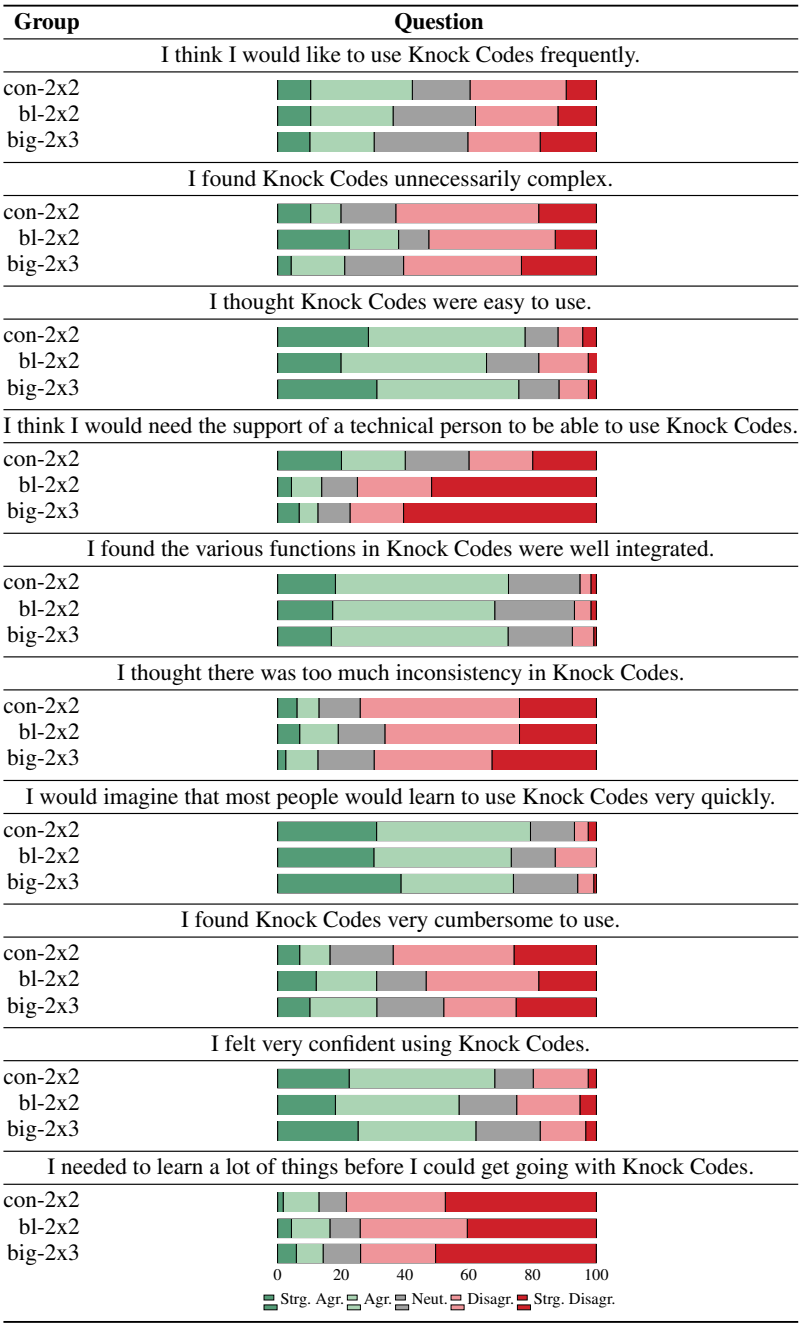


Figure 16: Responses to SUS questions: Averages are con-2x2 (69.8), bl-2x2 (65.3), and big-2x3 (68.1)

An Empirical Study of Wireless Carrier Authentication for SIM Swaps

Kevin Lee Ben Kaiser Jonathan Mayer Arvind Narayanan

Department of Computer Science and Center for Information Technology Policy
Princeton University

Abstract

We examined the authentication procedures used by five prepaid wireless carriers when a customer attempted to change their SIM card. These procedures are an important line of defense against attackers who seek to hijack victims' phone numbers by posing as the victim and calling the carrier to request that service be transferred to a SIM card the attacker possesses. We found that all five carriers used insecure authentication challenges that could be easily subverted by attackers. We also found that attackers generally only needed to target the most vulnerable authentication challenges, because the rest could be bypassed. Authentication of SIM swap requests presents a classic usability-security trade-off, with carriers underemphasizing security. In an anecdotal evaluation of postpaid accounts at three carriers, presented in Appendix A, we also found—very tentatively—that some carriers may have implemented stronger authentication for postpaid accounts than for prepaid accounts.

To quantify the downstream effects of these vulnerabilities, we reverse-engineered the authentication policies of over 140 websites that offer phone-based authentication. We rated the level of vulnerability of users of each website to a SIM swap attack, and have released our findings as an annotated dataset on issms2fasecure.com. Notably, we found 17 websites on which user accounts can be compromised based on a SIM swap alone, i.e., without a password compromise. We encountered failures in vulnerability disclosure processes that resulted in these vulnerabilities remaining unfixed by nine of the 17 companies despite our responsible disclosure. Finally, we analyzed enterprise MFA solutions from three vendors, finding that two of them give users inadequate control over the security-usability tradeoff.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

1 Introduction

Mobile devices serve many purposes: communication, productivity, entertainment, and much more. In recent years, they have also come to be used for personal identity verification, especially by online services. This method involves sending a single-use passcode to a user's phone via an SMS text message or phone call, then prompting the user to provide that passcode at the point of authentication. Phone-based passcodes are frequently used as one of the authentication factors in a multi-factor authentication (MFA) scheme and as an account recovery mechanism.

To hijack accounts that are protected by phone-based passcode authentication, attackers attempt to intercept these passcodes. This can be done in a number of ways, including surveilling the target's mobile device or stealing the passcode with a phishing attack, but the most widely reported method for intercepting phone-based authentication passcodes is a SIM swap attack. By making an unauthorized change to the victim's mobile carrier account, the attacker diverts service, including calls and messages, to a new SIM card and device that they control.

SIM swap attacks allow attackers to intercept calls and messages, impersonate victims, and perform denial-of-service (DoS) attacks. They have been widely used to hack into social media accounts, steal cryptocurrencies, and break into bank accounts [1–3]. This vulnerability is severe and widely known; since 2016 NIST has distinguished SMS-based authentication from other out-of-band authentication methods due to heightened security risks including "SIM change" [4].

SIM swap procedures have valid purposes: for example, if a user has misplaced their original device or acquired a new device that uses a different size SIM card slot than the device it is replacing. In these cases, customers contact their carrier (often by calling the carriers' customer service line) to request a SIM card update on their account. The customer is then typically presented with a series of challenges that are used to authenticate them. If the customer is successfully authenticated, the customer service representative (CSR) proceeds to update the SIM card on the account as requested.

We examined the types of authentication mechanisms in place for such requests at five U.S. prepaid carriers——AT&T,

T-Mobile, Tracfone, US Mobile, and Verizon Wireless——by signing up for 50 prepaid accounts (10 with each carrier) and subsequently calling in to request a SIM swap on each account.[1] Our key finding is that, at the time of our data collection, all five carriers used insecure authentication challenges that could easily be subverted by attackers. We also found that in general, callers only needed to successfully respond to one challenge in order to authenticate, even if they had failed numerous prior challenges in the call. Within each carrier, procedures were generally consistent, although on nine occasions across two carriers, CSRs either did not authenticate the caller or leaked account information prior to authentication. These findings are consistent with a policy that overemphasizes usability at the expense of security.

Our testing results offer insight into the security policies at major U.S. prepaid mobile carriers with implications for the personal security of the millions of U.S.-based customers they serve. We also offer recommendations for carriers and regulators to mitigate the risks of SIM swap attacks.

Next, we evaluated the authentication policies of over 140 online services that offer phone-based authentication to determine how they stand up to an attacker who has compromised a user's phone number via a SIM swap. Our key finding is that 17 websites across different industries have implemented authentication policies with logic flaws that would enable an attacker to fully compromise an account with just a SIM swap.

Finally, we analyzed enterprise MFA apps offered by Duo Security, Okta, and Microsoft, to further understand the downstream impact of SIM swaps. Our finding is that Duo enables SMS-based MFA by default (and makes it difficult to disable), which introduces security risks. The default authentication policies at Duo and Okta sit on opposite ends of the security-usability tradeoff, with Duo overemphasizing usability by default and Okta overemphasizing security.

Responsible disclosure and responses. In July 2019 we provided an initial notification of our findings to the carriers we studied and to CTIA, the U.S. trade association representing the wireless communications industry. In January 2020, T-Mobile informed us that after reviewing our research, it had discontinued the use of call logs for customer authentication.[2]

We reported our MFA configuration findings to the 17 vulnerable websites in January 2020 (Section 7.3). We document the widespread failures we encountered in the vulnerability disclosure processes established by companies, including the fact that many companies have no process to report security policy vulnerabilities as opposed to software bugs. As a consequence, nine of the 17 websites remain vulnerable, which cumulatively have billions of users.

[1] Unlike a postpaid account, registering a prepaid account does not require a credit check, making it easy for one researcher to sign up for multiple accounts. Authentication procedures may differ for postpaid accounts.

[2] Some carriers asked the customer for information that can be obtained from call logs for authentication, such as the phone number of the last placed or received call. The use of call logs—whether incoming or outgoing—for authentication is insecure because attackers can call the victim or trick the victim into placing a call.

2 Background

2.1 SIMs and number portability

Wireless service to a mobile device is tied to that device's SIM card. Wireless carriers keep track of the mapping between phone numbers and SIMs to ensure that calls, messages, and data connections are routed to the correct customer. Generally, the mapping from a phone number to a SIM is a one-to-one relationship: a phone number can only be associated with a single SIM at any given point in time and vice versa.

SIM cards further the bring-your-own-device (BYOD) policy that exists at many carriers today: users are usually free to bring their own devices to the network, provided that the device is not locked to another carrier and that the customer purchases a new SIM card. Similarly, if a user were to ever switch devices, they could easily remove their existing SIM card and insert it into the new device. The customer could also purchase a new inactive SIM card, provide a CSR at the mobile provider with the new card's Integrated Circuit Card Identifier (ICCID), and migrate the service over to the new SIM before inserting it into the new device. From then, service on the original device would be disconnected, and all connections would move over to the new device with the now-activated SIM.

In the U.S., customers also have the option of taking their phone numbers with them whenever they switch carriers; a user seeking to move their number to a new provider would provide their old account details to their new provider, who would in turn request the number from the original provider. After validating the request, the original provider would push their number over to the new carrier. Local number portability—as this is called—is regulated by the Federal Communications Commission (FCC), allowing customers to switch carriers while retaining their original numbers for little to no cost.

There are two scenarios in which an account holder would need to change the SIM card in their device: a SIM swap or a *port out*. In a SIM swap, the account and phone number stay with the original carrier, and only the SIM card is changed. In a port out, the number is transferred to a new account at a new carrier. Both types of account changes involve switching SIM cards; SIM swaps use cards from the same carrier whereas port outs use cards from different carriers.

We study SIM swaps due to their relative simplicity; we cannot be confident that the authentication procedures for SIM swaps and port outs are the same. It is worth noting the distinction that SIM swaps typically take no more than two hours (and are often instantaneous), while port outs can take several days.

Carrying out an unauthorized SIM swap or port out to

hijack a victim's phone number is obviously unlawful—at minimum a violation of the Computer Fraud and Abuse Act (CFAA) and possibly wire fraud or wiretapping. Authorities and companies have posted advisories against using SMS for two-factor authentication (2FA), most notably in 2016 when the National Institute of Standards and Technology (NIST) initially declared SMS-based authentication to be deprecated in its draft of *Digital Identity Guidelines* [4]. NIST slightly softened its stance a year later by categorizing SMS-based authentication as "restricted"—an authentication factor option that carries known risks [5]. The rise in SIM swap scams has recently led organizations like the Better Business Bureau (BBB) to issue warnings to consumers against using their phone numbers for authentication [6].

2.2 Phone-based authentication

Phone-based passcodes are a common authentication technique. They are typically used as one of multiple authentication factors, as a backup authentication option, or as an account recovery method. A passcode can be transmitted to a user's phone via an SMS text message, a phone call, an email, or an authenticator app. The Internet Engineering Task Force (IETF) has published standards for generating, exchanging, and verifying passcodes as part of an authentication procedure [7, 8].

We distinguish passcodes delivered by SMS and phone calls from the other phone-based passcode authentication methods (authenticator apps and email passcodes). The former are susceptible to SIM swap and port out vulnerabilities because they are tied to a phone number and the associated cellular service; the latter are not. In the balance of the paper, we consider only passcode authentication via SMS and phone call and use the terms "SMS-based authentication" and "SMS-based MFA" to describe these methods.

3 Threat model

We assumed a weak threat model: our simulated attacker knew only information about the victim that would be easily accessible without overcoming any other security measures. Specifically, our attacker knew the victim's name and phone number. We also assumed that the attacker was capable of interacting with the carrier only through its ordinary customer service and account refill interfaces, and for purposes of one attack, that the attacker could bait the victim into making telephone calls to a chosen number. Other than providing scripted answers and persisting through failed authentication challenges, the research assistants (RAs) simulating our attacker used no social engineering tactics. As we will show later, this weak attacker was able to defeat several different authentication challenges used by carriers.

We note that many realistic adversaries could gain access to additional information that could be used to bypass challenges. They could also seem more credible by spoofing the victim's caller ID or escalating the request to management, none of which were included in our method. By assuming such a conservative threat model, we provide a lower bound on real-world attacker success rates.

4 Method

The goal of a SIM swap attack is to convince the carrier to update the SIM card associated with a victim's account, thereby diverting service from the victim's SIM and phone to a new SIM and phone in the adversary's possession.

In our study, we sought to reverse-engineer the policies for SIM swaps at five U.S. carriers—AT&T, T-Mobile, Tracfone, US Mobile, and Verizon Wireless. We answer the following questions:

1. What are the authentication procedures that prepaid carriers use for SIM swaps? Are they consistent within carriers? Are they consistent across carriers?
2. Do SIM swap authentication procedures withstand attack?
3. What information would an attacker need about their victim to perform a SIM swap attack? Can the attack be perpetrated using only easily acquirable information?

Tracfone and US Mobile are mobile virtual network operators (MVNOs), meaning that they do not own their own wireless network infrastructure and instead contract access to the infrastructure of other networks. The MVNO marketplace is diverse: there are dozens of companies in the U.S. serving a combined subscriber base of over 36 million. Tracfone is a 20-year-old company that currently services over 25 million customers; US Mobile is a much smaller and newer provider, founded in 2014 and serving just 50,000. The difference in their age could suggest different policies for authenticating customers, so we included them in our study.

We created 10 simulated identities for our study and assigned each a name, date of birth, geographical location, and email address. For each identity, we registered prepaid accounts at all five carriers, using SIM cards we had purchased from electronics stores. The accounts were funded with prepaid refill cards purchased at local retail outlets; in a few cases we used one-time virtual debit cards instead. Due to the possibility that carriers log seen phones, we did not reuse devices between experiments; that is, each identity was assigned a unique "victim phone" and "adversary phone," for a total of 20 devices. For each account, we spent at least a week making and receiving phone calls and text messages to generate usage history. At the end of this phase, we hired research assistants (RAs)—who had been designated as the account owners at signup—to call the customer service number for the carrier and request that the SIM card on the account be updated to a new SIM card in our possession. We placed each call from a device that was not registered to the account being studied. During the call, we took notes on what pieces of authenticating information the CSR requested and whether or

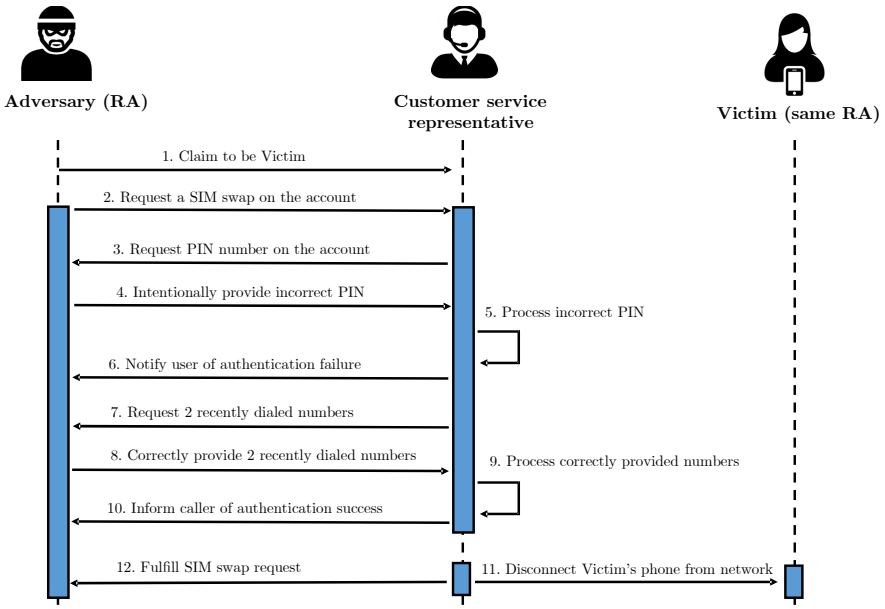


Figure 1: An example scenario from following our call script. The adversary (the research assistant) intentionally fails the first authentication scheme, but correctly answers the second one because of its inclusion in the threat model. The victim (the same research assistant) receives a notification about an account change when the SIM swap is complete.

not the swap was ultimately successful. We did not record or transcribe the calls.

On the calls, all RAs followed the same script: they informed the CSR that their SIM appeared to be faulty because service on the device was intermittent, but that they had a new SIM card in their possession they could try to use. They then responded to any authentication challenges the CSR posed. If the RA could not answer an authentication challenge correctly within the capabilities of the simulated attacker (see Section 3), the RA was instructed to claim to have forgotten the information or to provide incorrect answers. When providing incorrect answers to personal questions such as date of birth or billing ZIP code, RAs would explain that they had been careless at signup, possibly having provided incorrect information, and could not recall the information they had used. An example scenario from following our call script is shown in Fig. 1.

If the SIM swap was successful, we inserted the new SIM into a different device—the "adversary-controlled phone"—and proceeded to make a test call. We also made a test call on the original device to ensure that cell service had been successfully diverted. If the CSR had insisted on remaining on the line until the swap was completed, we gave a verbal confirmation and then ended the call. The experiments ran from May through July of 2019.

In all cases, the same RA simulated both the attacker and the victim, so there were no unauthorized transfers. The accounts were at all times controlled by the research team. RAs were paid standard institutional RA rates. While the purpose of the study was to understand carrier policies and practices, out of an abundance of caution we sought and obtained approval from Princeton University's Institutional Review Board. We provide additional details about mitigating risks in our study in Appendix B.

Our initial IRB application was submitted and approved in March of 2019 and April of 2019, respectively. We provided initial notification to the carriers we studied and CTIA on July 25, 2019. We presented our findings in-person to major carriers and CTIA in September 2019.

5 Results

We documented how the mobile carriers we studied authenticate prepaid customers who make SIM swap requests. We observed providers using the following authentication challenges:

- **Personal information**: street address, email address, date of birth
- **Account information**: last 4 digits of payment card number, activation date, last payment date and amount
- **Device information**: IMEI (device serial number), ICCID (SIM serial number)
- **Usage information**: recent numbers called (call log)
- **Knowledge**: PIN or password, answers to security questions
- **Possession**: SMS one-time passcode, email one-time passcode

	Personal Information			Account Information			Device Information		Usage Information	Knowledge		Possession	
	Street Address	Email Address	DOB	Last 4 of CC	Activation Date	Last Payment	IMEI	ICCID	Recent Numbers	PIN or Password	Security Questions	SMS OTP*	Email OTP
AT&T					✔	✔	✔	✔	✔	✔		✔	
T-Mobile									✔	✔		✔	✔
Tracfone	✔	✔	✔				✔	✔		✔	✔	✔	
US Mobile	✔	✔		✔				✔					
Verizon						✔	✔	✔	✔	✔		✔	

*We represent SMS OTP as a secure authentication factor because 1) we assume that a carrier sends the SMS OTP exclusively over its own network as a service message, such that the passcode is not vulnerable to routing attacks, and 2) we assume that if an attacker already has the ability to hijack a victim's SMS, a SIM swap does not provide the attacker with additional capabilities.

- generally accepted in the computer security research field
- had not been previously tested but we demonstrate is insecure (for reasons explained below)
- known to have security shortcomings (also for reasons described below)

Table 1: Authentication methods that we observed at each carrier. A checkmark means that a type of information was a component of at least one pathway for SIM swap customer authentication; it does not mean that a type of information was necessary or by itself sufficient for SIM swap customer authentication.

Table 1 presents the authentication methods that we observed at each carrier. Green represents secure authentication methods, red fields contain methods with known vulnerabilities, and yellow represents authentication methods that had not been previously documented and that we demonstrated are insecure. A checkmark in a cell indicates that on at least one call to the carrier's customer service, while attempting a SIM swap, a CSR requested that information to authenticate the subscriber. In other words, a checkmark means that a type of information was a component of at least one pathway for SIM swap customer authentication; *a checkmark does not mean that a type of information was necessary or by itself sufficient for SIM swap customer authentication.*

Although within each carrier the set of authentication mechanisms used by the 10 CSRs were mostly consistent, there was no particular pattern in which they were presented to us. The one exception, however, was T-Mobile: the order of PIN, OTP, and call log was consistent through all 10 calls. Further, providers that support PIN authentication (AT&T, T-Mobile, Tracfone, and Verizon) always used that mechanism first.

Our key findings are as follows:

1. **Mobile carriers use insecure methods for authenticating SIM swaps.**

   a. **Last payment.** We found that authenticating customers via recent payment information is easily exploitable. AT&T, T-Mobile, Tracfone, and Verizon use payment systems that do not require authentication when using a refill card. An attacker could purchase a refill card at a retail store, submit a refill on the victim's account, then request a SIM swap using the known refill as authentication.

   b. **Recent numbers.** We also found that using information about recent calls for authentication is exploitable. Typically CSRs requested information about *outgoing* calls. Consider the hypothetical following attack scenario: Using only the victim's name and phone number, our simulated adversary could call the victim and leave a missed call or message that would prompt the victim into returning the call to a number known to the attacker. This call would then appear on the outgoing call log and the attacker could use it for authentication. CSRs appeared to also have the discretion to allow authentication with *incoming* call information, as this occurred four times between AT&T, T-Mobile, and Verizon. An attacker can trivially generate incoming call records by calling the victim.

   c. **Personal information.** We found that Tracfone and US Mobile allowed personal information to be used for authentication. While our simulated attacker did not use this information, it would likely be readily available to real attackers (e.g., via data aggregators) and is often public, so it offers little guarantee of the caller's identity. We note that for over a decade, FCC rules have prohibited using "readily available biographical information" to authenticate a customer requesting "call detail information."[3]

   d. **Account information.** We found that AT&T, US Mobile, and Verizon allowed authentication using account information. As with personal information, this information would often be readily available to an adversary. Receipts (whether physical or electronic), for example, routinely include the last four digits of a payment card number. We note that PCI DSS, the industry standard for protecting payment card information, does not designate the last four digits of a payment card as "cardholder data" or "sensitive authentication data" subject to security requirements [9]. As for the activation date associated with an account, that information may be readily available from business records (e.g., via a data aggregator), inferable

[3] 47 C.F.R. § 64.2010.

by website or mobile app logs (e.g., via User-Agent logs), or inferable via mobile app API access (e.g., via the usage stats API on Android or the health APIs on Android and iOS). We note that FCC rules also prohibit using "account information" to authenticate a customer requesting "call detail information."[4]

e. **Device information.** We found that all carriers except for T-Mobile use device information for authentication. These authentication methods included the customer's IMEI (device serial number) and ICCID (SIM serial number). Both the IMEI and ICCID are available to malicious Android apps, and IMEIs are also available to adversaries with radio equipment.

f. **Security questions.** We found that Tracfone used security questions for authentication. We also found that T-Mobile, Tracfone, and Verizon prompted users to set security questions upon signup. Prior research has demonstrated that security questions are an insecure means of authentication, because answers that are memorable are also frequently guessable by an attacker [10–12].

2. **Some carriers allow SIM swaps without authentication.** Tracfone and US Mobile did not offer any challenges that our simulated attacker could answer correctly. Yet, CSRs at these carriers allowed us to SIM swap without ever correctly authenticating: six times at Tracfone and three times at US Mobile.

3. **Some carriers disclose personal information without authentication, including answers to authentication challenges.**

   - **AT&T.** In one instance, the representative disclosed the month of the activation and last payment date and allowed multiple tries at guessing the day. They also guided us in our guess by indicating whether we were getting closer or further from the correct date.
   - **Tracfone.** In one instance, the representative disclosed the service activation and expiration dates. Neither are used for customer authentication at Tracfone.
   - **US Mobile.** In three instances, the representative disclosed the billing address on the account prior to authentication. In one instance, a portion of the address was leaked. In one instance, part of the email address was disclosed. In three instances, the representative disclosed portions of both the billing address and email address.

In addition to learning the carriers' authentication policies, we also documented whether the swap was successful or not. The outcomes are shown in Table 2.

[4] *Id.*

	AT&T	T-Mobile	Tracfone	US Mobile	Verizon
Success	10	10	6	3	10
Failure	0	0	4	7	0

Table 2: The outcomes of our SIM swap requests. Note that our attempts at major carriers were all successful.

	Recently dialed numbers	Last payment details	No authentication
AT&T	2	8	0
T-Mobile	10	0	0
Tracfone	0	0	6
US Mobile	0	0	3
Verizon	9	1	0

Table 3: The authentication scheme that was used to authenticate the calls on successful attempts.

In our successful SIM swaps, we were able to authenticate ourselves with the carrier by passing at most one authentication scheme. For instance, Verizon—a provider that uses call log verification—allowed us to SIM swap once we provided two recently dialed numbers, despite us failing all previous challenges, such as the PIN. Some CSRs at Tracfone and US Mobile also forgot to authenticate us during our calls, but they were able to proceed with the SIM swap, indicating that back-end systems do not enforce authentication requirements before a customer's account can be changed. Table 3 details the exact authentication challenge that was exploited in each successful call.

Devices transmit identifying information to the network, namely the International Mobile Equipment Identity (IMEI), which is unique to the device. Therefore carriers could presumably detect that we were not only switching SIM cards, but devices as well. This never presented an issue across our 50 calls; in three cases, the CSR noted verbally that the device IMEI had changed, but did not intervene or flag the account.

Our key finding is that all three major carriers in our study used manipulable information—call logs and/or payment information—for authentication. Carriers may have changed their customer authentication practices since our testing. We requested that they update us if they did.

6 Discussion

6.1 Weak authentication mechanisms

It has long been known that carriers' authentication protocols are subject to social engineering or subversion using stolen personal information [13, 14]. We found an additional, more severe vulnerability: carriers allow customers to authenticate using information that can be manipulated without authenticating.

In our experiments, several carriers relied on call log verification as an authentication method, asking us to provide recently dialed phone numbers (T-Mobile asked only for the

last four digits of one recently dialed number; Verizon required two full phone numbers). An adversary could easily obtain these records by baiting victims into calling numbers that he knows about. As an example, the adversary could first send an intentionally vague text message claiming to be an institution that the victim frequents (e.g., her school, bank, or healthcare provider) with a callback number. The victim might then call the number to learn more details. As long as the call connects, an outgoing call to this number will be logged in the victim's call record. The adversary can then provide that number as a correct response to the challenge when requesting a SIM swap at the carrier. Another attack that achieve the same result is the "one-ring" scam, in which the attacker hangs up just as the victim's phone starts ringing; the victim—upon seeing the missed call—will call back out of curiosity. To make matters worse, in four instances between AT&T, T-Mobile, and Verizon, we were able to succeed call record verification by providing incoming numbers. This means that the adversary would not even need the victim to place a call; as long as the victim picks up the initial call from the adversary, a valid record in the call log would be generated.

The second manipulable authentication challenge we saw in our experiments is payment record verification. In these cases, we were asked to provide details about the most recent payment on the accounts. Most of the carriers in our study—including all of the major carriers—allow for payments to be made over the phone. None of these payment systems require any authentication when making these payments using a refill card, even when calling from a third-party number. To obtain payment information, an adversary can first purchase a refill card for the victim's mobile carrier at, for example, a convenience store. After dialing into the payment system, he can enter the victim's phone number and redemption code on the refill card to add value to her account. Once the payment is accepted, the adversary—now with complete knowledge of the most recent payment—can call the carrier to request a SIM swap and successfully pass payment record verification. This attack has an even lower barrier to entry than call log verification because it requires no action from the victim. Although it does require the attacker to spend a small amount of money, minimum required payments are typically quite low (between $5-30 in our experiments). As shown in Table 1, two of the five carriers in our study (both major carriers) support payment record verification. For AT&T, payment record verification was used consistently in all 10 calls. Only US Mobile did not allow for unauthenticated refills to be made; they only supported online refills which required account authentication.

Tracfone and US Mobile—the MVNOs—did not use any manipulable information for authentication and thus had fewer successful swaps. However, nearly all of their authentication challenges came from public records. A dedicated adversary would plausibly be able to obtain a victim's DOB, address, email address, or answers to security questions through online profiles, and thus be able to successfully authenticate at the carriers. Even then, we were still able to succeed at Tracfone and US Mobile in instances where CSRs skipped authentication, which suggests that policies for customer authentication at those carriers might not be as rigorous as those at other carriers.

In all instances of unauthenticated information leakage, the customer service representatives had released parts of the answer—either the email address, billing address, activation date, or payment date—as hints and said we would be authenticated once we remembered the whole response. This suggests that sensitive account details are stored in the clear and visible to CSRs, who are thus susceptible to social engineering attacks.

6.2 Severity

It has long been known that mobile subscribers are at risk of SIM swap attacks [15–17]. Our research demonstrates that insecure means of customer authentication are still widely used by mobile carriers. This exposes customers to severe risks: denial of service, interception of sensitive communications, and impersonation, which can lead to further account compromises.

As mentioned above, an attacker who hijacks a victim's phone number could intercept authentication passcodes sent by SMS or phone call. Phone-based passcode authentication as a second factor or account recovery method is ubiquitous on the internet, including at financial institutions and cryptocurrency exchanges where access to online accounts confers access to funds. Since reports about bank theft stemming from SIM swap attacks appear regularly in the media, we consider this a high severity vulnerability [18, 19].

At the recommendation of wireless carriers, we conducted an additional round of data collection to understand how customers could protect themselves against SIM swap attacks. We signed up for one additional prepaid account each with AT&T, T-Mobile, and Verizon; after one week, we called to inquire about and enable any safeguards against SIM swaps and port outs, citing T-Mobile's `NOPORT` as an example.[5] None of the carriers had additional protection features beyond the ones we had set in our initial study. We placed these calls in September 2019.[6] This additional result indicated that prepaid customers not only were vulnerable to SIM swap attacks, but also were not capable of easily employing any mitigation.

We studied prepaid accounts because they can be registered without undergoing a credit check, enabling us to scale the

[5] `NOPORT` is a T-Mobile option that heightens authentication requirements for port out requests [20]. While `NOPORT` would not itself protect against SIM swap attacks, at least as currently implemented, we referenced it during our calls with CSRs. During the course of our additional data collection, we also found that T-Mobile did not offer `NOPORT` for prepaid accounts.

[6] Verizon has since implemented an opt-in feature called `Number Lock`, which prohibits port out requests unless switched off [21]. The feature is exclusive to postpaid accounts.

number of test accounts. Prepaid plans accounted for 21% of U.S. wireless connections in Q3 2019, or about 77 million connections [22].[7] Compared to postpaid accounts, these contract-free plans are less expensive and do not require good credit, so they are more attractive to (and are often marketed to) low-income customers. Based on our experimental results for prepaid accounts, as well as our anecdotal evaluation of postpaid accounts (presented in Appendix A), we hypothesize that current customer authentication practices disproportionately place low-income Americans at risk of SIM swap attacks.

Anecdotally, during this study, one of the authors themselves fell victim to an account hijacking via a SIM swap attack. After initial unsuccessful attempts to authenticate himself to the carrier using personal and knowledge-based information, he escalated the issue to the carrier security team. From there, he was able to leverage our findings by requesting to authenticate via recently dialed numbers—a method which we knew the carrier supported although it had not been offered in this instance.

7 Analysis of phone-based authentication

Software tokens and SMS-based passcodes delivered by SMS or call have become popular authentication schemes for online services [25, 26]. SMS-based passcodes as a second authentication factor are an especially common option, as they make the security of MFA available to any user with an SMS-enabled phone.

We aimed to reverse-engineer the authentication policies of popular websites and determine how easy it is for an attacker to compromise a user's account on the website provided they have successfully carried out a SIM swap.

7.1 Method

We started with the dataset used by `TwoFactorAuth.org`, an open-source project to build a comprehensive list of sites that support MFA. Anyone can contribute MFA information about websites to the database, while the owner—a private developer—acts as the moderator. In the dataset, over 1,300 websites are grouped by categories including healthcare, banking, and social media. The available methods are also listed under each website in the dataset. As of late 2019, 774 of the sites in the dataset support MFA; of those, 361 support SMS-based MFA. The 361 websites that support SMS-based authentication are of interest to us. Of these, 145 were accessible for our analysis; the rest required ID verification, enterprise signups, payment, or were duplicate entries (e.g., the Xbox site uses Microsoft's login system). We used a snapshot of the dataset from November 1, 2019.

The `TwoFactorAuth.org` dataset lists the available authentication factors for each website, but it does not include information about how authentication can be configured or how different authentication factors are presented to the user (e.g., which are recommended or set as defaults). To compile this information, we signed up for accounts at each website and traversed their authentication flows. To the best of our knowledge, we contribute the first dataset that shows how MFA is implemented in practice.

At each website, we created a user account and provided all requested personal information. After signing up, we enrolled in MFA using the recommended configurations at each site, opting for schemes that were mandated, listed first, or had conspicuous labeling. We then examined other possible MFA configurations, if available, taking note of schemes that were mandatory, linked, or automatically activated. Between each configuration setup, we also looked at account recovery options. We took screenshots of the authentication options, enrollment process, login procedures, and account recovery procedures at all websites. We tested each configuration on a new browser session with no previous site data.

We classified configurations into three categories: secure, insecure, and doubly insecure. A doubly insecure configuration indicates that a SIM swap alone is enough for account compromise; the configuration uses both SMS-based MFA and SMS-based password recovery. An insecure configuration can only be compromised if the attacker knows the account password; these configurations offer SMS-based authentication but do not allow for SMS-based password recovery (the attacker could obtain the password via data dumps, social engineering, or compromising the victim's account recovery email). The secure configuration uses stronger authentication schemes, such as authenticator apps, and cannot be recovered or reset by SMS.

7.2 Results

Our key findings are as follows:

1. **The majority of websites default to insecure configurations.** Of the 145 websites, 83 (a majority) have recommended or mandated configurations that are insecure. For most of these websites, there are other secure schemes present; only 14 websites have SMS as their sole MFA option.
2. **Some websites are doubly insecure.** 17 websites allow doubly insecure configurations, 13 of which default to or recommend doubly insecure configurations.[8] Accounts of users who choose these configurations can be compromised with a SIM swap alone. That is, an attacker needs

[7] This figure is based on data from carriers' earnings and financial statements. Carriers may use slightly different terms and definitions; e.g., Verizon defines a "connection" as an individual line of service for a wireless device while T-Mobile defines a "customer" as a SIM card associated with a revenue-generating account [23, 24], a seemingly equivalent metric. These definitions explain how carriers appear to have a population penetration rate above 100%, as an individual can possess multiple wireless-connected devices.

[8] Additionally, 10 websites that have SMS-based password recovery from examining their account recovery pages, but could not sign up for accounts due to the aforementioned restrictions.

only the victim's phone number to reset the password and bypass SMS-based authentication. These websites span different industries, including finance (Paypal, Venmo, Taxact), travel (Finnair), commerce (Amazon, eBay), and social media (Snapchat). We initially redacted the names and other identifying information of these websites in our annotated dataset, while providing initial notification as part of the responsible disclosure process (Section 7.3).

Recall that the doubly insecure configuration is only possible if SMS-based account recovery is also available. We found 11 websites that use SMS-based password recovery, but switch to different recovery tools—such as email or manual review—when MFA is enabled. Similarly, we found two websites that switch when SMS-based MFA is enabled.

3. **Security is only as good as the weakest link.** 10 websites recommend secure authentication schemes but simultaneously suggest insecure methods, like SMS or personal knowledge questions, as backups. Since an attacker only needs to defeat one of the authentication schemes to defeat MFA, an insecure backup renders the configuration insecure. Eight websites with multiple authentication options also mandate initial enrollment in SMS before allowing users to switch to other MFA schemes. Six websites with multiple options mandate SMS in order to keep MFA enabled.

4. **Some websites give users a false sense of security.** Some services automatically enroll users in email- or SMS-based MFA using the email address or phone number on file, respectively, without any user input or notice. Seven websites enroll users in SMS-based MFA without notice, either with the account recovery number or a phone number a user must provide in order to sign up for a non-SMS-based 2FA method. Even if the user then signs up for another MFA method, they continue to be simultaneously enrolled in SMS-based MFA without being made aware of it. Thus even users who are educated about SIM swap risks may nonetheless be lulled into a false sense of security. At four of these websites, the automatic SMS 2FA enrollment renders the configuration doubly insecure. A user may believe that account compromise requires both a stolen password and a compromise of the authenticator app (e.g. via phone theft), but in fact, a SIM swap alone is sufficient.

5. **Some websites offer 1-step SMS OTP logins.** Seven websites also offer 1-step logins via an SMS OTP. eBay, for instance, will send users a temporary password via SMS if MFA is not enabled, and WhatsApp uses SMS OTP by default if MFA is not enabled.

The annotated dataset describing all of our findings is available at issms2fasecure.com.

7.3 Failures in vulnerability disclosure processes

We attempted to responsibly disclose the vulnerabilities we uncovered to the 17 affected websites. Only in 4 of the 17 cases did the process work as expected and result in bug fixes. We document the failures we encountered and call for improvements in vulnerability disclosure processes.

Method. In January 2020 we attempted to notify the 17 websites described above of the presence of doubly insecure configurations. We first looked for email addresses dedicated to vulnerability reporting; if none existed, we looked for the companies on bug bounty platforms such as HackerOne. Many companies outsource bug reporting to these third-party platforms in order to triage reports for relevance and novelty. Reports are screened by employees of the platform, who are independent from the company, and passed on to the company's security teams if determined to be in scope. If we were unable to reach a company through a dedicated security email or through bug bounty programs, as a last resort, we reached out through customer support channels.

Sixty days after our initial notifications, we re-tested the companies using the same method in Section 7.1, except for those that reported that they had fixed the vulnerabilities.

Outcomes. Three companies—Adobe, Snapchat, and eBay—acknowledged and promptly fixed the vulnerabilities we reported. In one additional case, the vulnerability was fixed, but only after we exhausted the three contact options listed above and reached out to company personnel via a direct message on Twitter.

In three cases—Blizzard, Microsoft, and Taxact—our vulnerability report did not produce the intended effect (as documented in the following paragraph), but in our 60-day re-test we found that the vulnerabilities had silently been fixed. We do not know whether the fixes were implemented in light of our research.

There were several failure modes, which were not mutually exclusive.[9] In five cases, personnel did not understand our vulnerability report, despite our attempts to make it as clear as possible, shown in Appendix C. For example, Microsoft claimed that SIM swaps are widely known, and did not appreciate that their insecure MFA configuration exacerbated the issue. In five cases, we received no response. Predictably, all four attempts to report security vulnerabilities through customer support channels were fruitless: either we received no response or personnel did not understand the issue. Three of the four reports we submitted to bug bounty programs also resulted in failures and were closed due to the absence of a bug (recall that our findings are not software errors, but rather, logically inconsistent customer authentication policies).[10] HackerOne employs mechanisms that restrict

[9] The counts in this paragraph are out of a total of 13 websites, including the three that silently fixed the vulnerabilities.

[10] We had unsuccessfully submitted our vulnerability reports to carriers

users from submitting future reports after too many closed reports [27], which could disincentivize users from reporting legitimate vulnerabilities [28].

We have listed all 17 responses in Appendix C. Unfortunately, nine of these websites are doubly insecure *by default* and remain so as of this writing. Among them are payment services PayPal and Venmo. The vulnerable websites cumulatively have billions of users.

We provide an up-to-date timeline of responses on this study's website at `issms2fasecure.com`.

7.4 Analysis of enterprise MFA solutions

Many organizations offer (or require) MFA to their personnel for accessing internal resources. Most of these MFA solutions are provided by third-party services and integrate with organizations' existing login pages. To further understand the downstream impact of SIM swaps, we examined the handling of SMS-based MFA by three such vendors: Duo Security, Okta, and Microsoft. We selected these solutions based on popularity reports by Gartner, a global technology research and advisory firm [29]. We focused on the security-usability tradeoff provided by these solutions.

Method. In addition to checking the documentation for how those services handle SMS-based MFA, we created fictitious organizations and signed up for administrator accounts at each service. Next, we invited a new user to our organization, and finished account setup—along with MFA enrollment—from the user view. Both services offer proprietary mobile apps that come with authentication prompts and authenticator passcodes (TOTP); we installed the apps when instructed. Our findings are as follows:

Findings: Duo Security MFA. We find that Duo automatically and silently enrolls the user in SMS-based MFA, despite the availability of stronger second factors, unnecessarily weakening security.

When a user enrolls in MFA, Duo requires them to specify the type of device they are adding. If the user elects to add a smartphone (which Duo recommends), she will be required to add a phone number.[11] The user will be automatically enrolled in SMS-based MFA, provided that the organization has enabled it (which is the default). The user is also automatically enrolled in two other MFA methods: push notifications and TOTP. Users are not informed of the authentication methods they have been enrolled in during setup.

Users can view their authentication methods after logging in for the first time by navigating to the MFA page. However, they cannot modify their authentication methods (e.g., disable SMS-based MFA) — only an administrator can do so. Intriguingly, we found that users can bypass the requirement to enter a phone number (while retaining the other authentication methods) by setting up their smartphones as tablets. However, this is undocumented.

Findings: Okta Adaptive MFA. Okta does not suffer from the abovementioned vulnerability. It uses a method-oriented enrollment process: users explicitly enroll in authentication methods without being asked to provide their device details.

However, only the proprietary app is enabled as a second factor by default, while all other authentication methods, including SMS, are disabled, which means that users are not given any choice of authentication methods and cannot choose to enroll in SMS-based 2FA. Unless an administrator changes this policy, users without smartphones — or who do not wish to install the app — are locked out of the system.

Findings: Microsoft Azure MFA. We find that Azure defaults to SMS-based MFA during enrollment, despite the availability of stronger second factors, potentially weakening security.

Azure—like Okta—uses a method-oriented enrollment process. With the default administrator settings, users are able to choose between SMS, push notifications, and TOTP, with SMS being the default. However, the UI is slightly confusing: users must first select the medium to receive authentication messages from a dropdown menu (e.g., "Authentication phone" for SMS, "Mobile app" for push notifications and TOTP). "Authentication phone" is the default menu option provided that the organization has enabled SMS-based MFA (which is the default), so a user may be unaware that stronger second factors are available.

The contrasting approaches by Duo and Okta, and their corresponding limitations — one weakens security, and the other hurts usability — suggests an underlying issue, which is that the MFA vendors seek to maximize administrators' control over configuration for the whole organization and minimize variation between users. Allowing users more control, while also giving them guidance about benefits and risks, may allow for a more nuanced security-usability tradeoff. Azure does give users such control, although it offers SMS as the default and the confusing user interface compounds this issue.

8 Recommendations

8.1 Recommendations for carriers

In evaluating existing and proposed authentication schemes, we looked to the framework proposed by Bonneau et al. to consider the usability, deployability, and security of these mechanisms [30]. We also discussed usability and deployability issues with wireless carriers and CTIA. We offer the following recommendations:

1. **Carriers should discontinue insecure methods of customer authentication.** Every mobile carrier in our study, with one exception, already offers secure methods

via HackerOne when possible (i.e. for AT&T, T-Mobile, and Verizon) before reaching out to CTIA. If we include those figures, six out of seven reports to bug bounty programs resulted in failures.

[11] Duo allows administrators to add devices for users in the admin interface, where the phone number requirement for smartphones is also present.

of customer authentication: password/PIN,[12] one-time passcode via SMS (to the account phone number or a pre-registered backup number), or one-time passcode via email (to the email address associated with the account). Abandoning insecure authentication schemes—personal information, account information, device information, usage information, and security questions—may inconvenience customers who are legitimately requesting a SIM swap, but preventing account hijacking attacks is crucial to customers' privacy and security. Moreover, legitimate SIM swap requests appear to be infrequent, occurring only when a user's SIM is damaged or lost, when a user acquires a new phone that is incompatible with their SIM, or in other rare cases. These requests may become even more infrequent going forward, as users are now waiting longer before switching their devices [32]. Thus, carriers should begin to phase out insecure authentication methods and develop measures to educate customers about these changes to reduce transition friction. Carriers should use data on the type and frequency of legitimate SIM swaps to assess the usability impact of authentication procedures.

2. **Implement additional methods of secure customer authentication.** We recommend that mobile carriers implement customer authentication for telephone support via a website or app login, or with a one-time password via a voice call. The methods do not require memorization or carrying extra devices and are easy to learn. They also should not pose significant costs to carriers because the infrastructure already exists; all carriers we examined support online accounts via websites and/or mobile applications.
3. **Provide optional heightened security for customers.** We recommend that carriers provide the option for customers to enable MFA for account change requests, as well as the option to disable account changes by telephone or at a store.
4. **Respond to failed authentication attempts.** If someone attempts to authenticate as a customer and is unsuccessful, we recommend that carriers notify the customer and heighten security for the account. An adversary should not be allowed to attempt multiple authentication methods or to repeatedly attempt authentication. Moreover, even if an adversary was able to successfully authenticate after failing previous attempts, carriers should not be convinced that the caller is who they claim to be. For instance, a customer who has forgotten their PIN, is unable to access their email and backup phone for an OTP, but can recall some call log information, is very unlikely to be the customer, but rather an adversary who is trying to authenticate using call log verification. If a customer who loses or has their phone stolen goes into a store and attempts to purchase a new device with the original number, they should not be allowed to authenticate with only a government-issued ID. IDs are open to forgery, and the absence of the original device—though unfortunate—should result in additional security measures being taken. In both scenarios, the carrier can respond in different ways, such as adding a 24 hour delay to a SIM swap request while notifying the customer via SMS or email, going further down the authentication flow, or denying the caller's request for a period of time. In other words, authentication should not be binary.
5. **Restrict customer support representative access to information before the customer has authenticated.** There is no need for representatives to access customer information before authentication, and providing such access invites deviation from authentication procedures and enables social engineering attacks. In all instances of unauthenticated information leakage in our study, the customer support representatives had released parts of the answer as hints and stated we would be authenticated once we remembered the whole response. This strongly suggests that sensitive account details are, for at least some carriers, visible to representatives prior to customer authentication.
6. **Publicly document customer authentication procedures.** Carriers should list all the ways customers can be authenticated over the phone in order to avoid uncertainties regarding risks and defenses. They also stand to benefit from informing their customers and homogenizing the authentication flow within and between carriers. In addition, carriers should maintain pages that explain SIM swap attacks and any available security countermeasures that they offer.
7. **Provide better training to customer support representatives.** Representatives should thoroughly understand how to authenticate customers and that deviations from authentication methods or disclosure of customer information prior to authentication is impermissible. That said, we emphasize that training alone is not sufficient—there should also be technical safeguards in place.

Taken collectively, these recommendations should decrease the number of unauthorized SIM swaps by improving user authentication.

8.2 Call for research: better design of customer service interfaces

It is essential that authentication procedures be consistent across callers and CSRs. This is challenging because CSRs may be susceptible to social engineering attacks (e.g., an adversary pretending to be a victim of domestic violence desperate to urgently regain control over their account). The

[12] A password or PIN that is easily guessed is not secure, of course. Carriers must have safeguards that prevent users from choosing weak PINs [31].

software used by CSRs play an integral role in keeping accounts secure, in particular:

1. Restrict CSR access to account information before the customer has authenticated
2. Present authentication mechanisms in a consistent order
3. Prohibit CSR bypass of user authentication

From our study, we believe that current customer support interfaces do not meet the above-mentioned requirements. That is, CSRs released parts of the answer as hints, authentication mechanisms were generally not presented in any particular order within and across carriers (with the exception of T-Mobile), and carriers allowed us to SIM swap without ever correctly authenticating in nine instances (Section 5). We are unable to find information about any software tools used by CSRs for authenticating customers.

An improved secure interface should complement improved CSR training procedures, and more importantly, be easy for CSRs to use. To our knowledge, CSRs themselves have never been subjects of study from a security and usability perspective. Just as the security community has realized the value of research on developers making security design decisions, CSRs should also be subjects of research, in order to effectively study security in practice [33, 34]. By studying workers' behaviors, the community can make recommendations on training procedures and interface design.

Our suggestions above can only be implemented with commitment from the carriers themselves. We call on carriers to collaborate with usable security researchers to study CSRs and their software tools. One important open research question is how carriers should respond to failed authentication attempts. Ignoring failures carries security risks (as we have documented) but an overly strict policy risks locking out customers. In the long term, carriers (and all other organizations that need to authenticate customers over the phone) should endeavor to develop an industry standard, informed by research, that is accessible to the community for scrutiny.

8.3 Recommendations for websites

Carriers are ultimately responsible for mitigating the authentication vulnerabilities that we have reported, but meanwhile, users of websites relying on SMS-based MFA continue to be at risk—in some cases severely (Section 7.2). We offer the following recommendations for websites to better protect their users from the effects of SIM swap attacks:

1. **Employ threat modeling to identify vulnerabilities.** Threat modeling is a fundamental information security technique that is used to identify vulnerabilities in a systematic way. It consists of a structured analysis of the application, the attacker, and the possible interactions between them. Many of our findings, especially the existence of doubly insecure websites, suggest a failure (or absence) of threat modeling.
2. **Implement at least one secure MFA option.** Websites without any other MFA options should roll out alternative options such as authenticator apps, and notify users when these options become available. Popular secure MFA options do not pose large usability hurdles. Reese et al. performed a usability lab study of five 2FA methods, including push notifications, SMS, TOTP, and U2F [35]. They found—with statistical significance—that push notifications, TOTP, and U2F have faster median authentication times and higher system usability scale (SUS) scores than those of SMS. Authenticator apps also have an added usability benefit over SMS-based MFA: the device need not be online to generate the one-time password.
3. **Eliminate or discourage SMS-based MFA.** Websites should not make SMS the default or recommended MFA option. Websites should highlight the dangers of SIM swaps, and label SMS as an option with known risks. As of 2019, only 15% of adults in the U.S. own non-smartphone cellular devices (compared to 81% of adults in the U.S. that own smartphones) [36]. As that share continues to decrease, websites should eliminate SMS-based MFA altogether.
4. **Improve vulnerability disclosure processes.** A bug bounty program is not a substitute for a robust security reporting mechanism, yet some companies are using it as such (Section 7.3). These third-party platforms appear to be overly strict with their triage criteria, preventing qualified researchers from communicating with the companies. Companies should maintain direct contact methods for security reporting procedures.

9 Conclusion

The theory and practice of user authentication has come a long way in the last decade. Yet these gains have been uneven. We found that five carriers in the United States continue to use authentication methods that are now known to be insecure, enabling straightforward SIM swap attacks. Further difficulties arise when security rests on interactions between independent systems. Phone-based authentication, and SMS in particular, has made rapid inroads because of convenience, but carriers don't adequately account for this scope creep in protecting against SIM swaps. Meanwhile, many online services view SIM swaps as "someone else's problem."

In addition to fixing the vulnerabilities we identified, our work suggests fruitful avenues for academia and industry: better quantifying the security-usability tradeoff in specific settings including over-the-phone authentication and enterprise authentication; studying user populations such as customer-service representatives and their user interfaces; and improving the vulnerability disclosure process for non-software vulnerabilities.

Acknowledgements

We are grateful to Mihir Kshirsagar for assisting with our vulnerability notification and presentation to carriers, to Ben Burgess for his advice on our experiment method, and to Arunesh Mathur for discussions on security-usability trade-offs. We also like to thank our research assistants who helped us carry out our experiments.

This work is supported by a grant from the Ripple University Blockchain Research Initiative.

References

[1] Brian Barrett. *How to Protect Yourself Against a SIM Swap Attack*. WIRED. Aug. 19, 2018. URL: https://www.wired.com/story/sim-swap-attack-defend-phone/ (visited on 12/01/2019).

[2] Brian Krebs. *Busting SIM Swappers and SIM Swap Myths*. Krebs on Security. Nov. 7, 2018. URL: https://krebsonsecurity.com/2018/11/busting-sim-swappers-and-sim-swap-myths/ (visited on 12/01/2019).

[3] Lorenzo Franceschi-Bicchierai. *How Criminals Recruit Telecom Employees to Help Them Hijack SIM Cards*. Motherboard. Aug. 3, 2018. URL: https://www.vice.com/en_us/article/3ky5a5/criminals-recruit-telecom-employees-sim-swapping-port-out-scam (visited on 12/01/2019).

[4] Paul A. Grassi et al. *DRAFT NIST Special Publication 800-63B Digital Authentication Guidelines. Authentication and Lifecycle Management*. June 24, 2016. URL: https://web.archive.org/web/20160624033024/https://pages.nist.gov/800-63-3/sp800-63b.html (visited on 02/15/2019).

[5] Paul A. Grassi et al. *NIST Special Publication 800-63B Digital Authentication Guidelines. Authentication and Lifecycle Management*. June 22, 2017. DOI: 10.6028/NIST.SP.800-63b.

[6] Better Business Bureau of Central Oklahoma. *BBB Warns About Cell Phone Porting Scams*. Feb. 6, 2018. URL: https://www.bbb.org/article/news-releases/17019-bbb-warns-about-cell-phone-porting-scams (visited on 12/01/2019).

[7] David M'Raihi et al. *HOTP: An HMAC-Based One-Time Password Algorithm*. Tech. rep. 4226. RFC Editor, Dec. 2005. 37 pp. DOI: 10.17487/RFC4226. URL: https://rfc-editor.org/rfc/rfc4226.txt.

[8] David M'Raihi et al. *TOTP: Time-Based One-Time Password Algorithm*. Tech. rep. 6238. RFC Editor, May 2011. 16 pp. DOI: 10.17487/RFC6238. URL: https://rfc-editor.org/rfc/rfc6238.txt.

[9] PCI Security Standards Council, LLC. *Payment Card Industry (PCI) Data Security Standard. Requirements and Security Assessment Procedures Version 3.2.1*. URL: https://www.pcisecuritystandards.org/documents/PCI_DSS_v3-2-1.pdf?agreement=true&time=1591585370712 (visited on 06/07/2020).

[10] John Podd, Julie Bunnell, and Ron Henderson. "Cost-Effective Computer Security: Cognitive and Associative Passwords". In: *Proceedings of the Sixth Australian Conference on Computer-Human Interaction (OZCHI)*. Nov. 1996. DOI: 10.1109/OZCHI.1996.560026.

[11] Stuart Schechter, A. J. Bernheim Brush, and Serge Egelman. "It's No Secret. Measuring the Security and Reliability of Authentication via "Secret" Questions". In: *Proceedings of the 30th IEEE Symposium on Security & Privacy (S&P)*. May 2009. DOI: 10.1109/SP.2009.11.

[12] Joseph Bonneau et al. "Secrets, Lies, and Account Recovery: Lessons from the Use of Personal Knowledge Questions at Google". In: *Proceedings of the 24th World Wide Web Conference (WWW)*. May 2015. DOI: 10.1145/2736277.2741691.

[13] Lorrie Cranor. *Your mobile phone account could be hijacked by an identity thief*. Tech@FTC. June 7, 2016. URL: https://www.ftc.gov/news-events/blogs/techftc/2016/06/your-mobile-phone-account-could-be-hijacked-identity-thief (visited on 06/07/2020).

[14] Federal Trade Commission. *Consumer Sentinel Network Data Book for January – December 2015*. URL: https://www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-january-december-2015/160229csn-2015databook.pdf (visited on 06/08/2020).

[15] Action Fraud. *Alert – how you can be scammed by a method called SIM Splitting*. May 9, 2014. URL: https://www.actionfraud.police.uk/alert/alert-how-you-can-be-scammed-by-a-method-called-sim-splitting (visited on 12/01/2019).

[16] Mateen Hafeez. *SIM fraud: Police zero in on public phone booth owners*. The Times of India. Aug. 9, 2008. URL: https://timesofindia.indiatimes.com/city/mumbai/SIM-fraud-Police-zero-in-on-public-phone-booth-owners/articleshow/3344515.cms (visited on 12/01/2019).

[17] Clayton Barnes. *Beware SIM card swop scam*. The Saturday Star. Jan. 7, 2008. URL: https://www.security.co.za/news/5907 (visited on 12/01/2019).

[18] Robert McMillan. *He Thought His Phone Was Secure; Then He Lost $24 Million to Hackers*. The Wall Street Journal. Nov. 8, 2019. URL: https://www.wsj.com/articles/he-thought-his-phone-was-secure-then-he-lost-24-million-to-hackers-11573221600 (visited on 12/01/2019).

[19] Jason Aten. *SIM Swapping Is the Biggest Security Threat You Face, and Almost No One Is Trying to Fix It. Here's Why It Matters*. Inc. Sept. 17, 2019. URL: https://www.inc.com/jason-aten/sim-swapping-is-one-of-biggest-cyber-security-threats-you-face-almost-no-one-is-trying-to-fix-it-heres-why-it-matter.html (visited on 12/01/2019).

[20] Lorenzo Franceschi-Bicchierai. *T-Mobile Has a Secret Setting to Protect Your Account From Hackers That It Refuses to Talk About*. Motherboard. Sept. 13, 2019. URL: https://www.vice.com/en_us/article/ywa3dv/t-mobile-has-a-secret-setting-to-protect-your-account-from-hackers-that-it-refuses-to-talk-about (visited on 01/06/2020).

[21] Verizon Wireless. *Transfer (port out) your number to another carrier FAQs. Number Lock*. URL: https://www.verizonwireless.com/support/port-out-faqs/ (visited on 04/06/2020).

[22] Frost & Sullivan TEAM Research. *Consumer Communication Services Tracker, Q3 2019*. Frost & Sullivan. July 5, 2019. URL: https://store.frost.com/consumer-communication-services-tracker-q3-2019.html (visited on 06/07/2020).

[23] T-Mobile US, Inc. *Q3 2019. Financial Results, Supplementary Data, Non-GAAP Reconciliations, Reconciliation of Operating Measures*. URL: https://s22.q4cdn.com/194431217/files/doc_financials/2019/q3/TMUS-09_30_2019-Financial-Results,-Supplemental-Data,-Non-GAAP-Reconciliations,-and-reconciliation-of-operating-measures-FINAL.pdf (visited on 06/07/2020).

[24] Verizon Communications. *2018 Annual Report*. URL: https://www.verizon.com/about/sites/default/files/2018-Verizon-Annual-Report.pdf (visited on 06/07/2020).

[25] Nitrokey. *DongleAuth.info*. URL: https://www.dongleauth.info/ (visited on 06/07/2020).

[26] Elie Bursztein. *The bleak picture of two-factor authentication adoption in the wild*. Elie. Dec. 21, 2018. URL: https://elie.net/blog/security/the-bleak-picture-of-two-factor-authentication-adoption-in-the-wild/ (visited on 06/07/2020).

[27] HackerOne. *Improving Public Bug Bounty Programs with Signal Requirements*. HackerOne Blog. Mar. 15, 2016. URL: https://www.hackerone.com/blog/signal-requirements (visited on 06/07/2020).

[28] Aron Laszka, Mingyi Zhao, and Jens Grossklags. "Banishing Misaligned Incentives for Validating Reports in Bug-Bounty Platforms". In: *Computer Security – European Symposium on Research in Computer Security (ESORICS) 2016*. Vol. 9879. Lecture Notes in Computer Science (LNCS). Sept. 2016. DOI: 10.1007/978-3-319-45741-3_9.

[29] Gartner, Inc. *Duo Security Competitors and Alternatives in User Authentication Reviews*. URL: https://www.gartner.com/reviews/market/user-authentication/vendor/duo-security/alternatives (visited on 02/25/2020).

[30] Joseph Bonneau et al. "The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes". In: *Proceedings of the 33rd IEEE Symposium on Security & Privacy (S&P)*. May 2012. DOI: 10.1109/SP.2012.44.

[31] Joseph Bonneau, Sören Preibusch, and Ross Anderson. "A Birthday Present Every Eleven Wallets? The Security of Customer-Chosen Banking PINs". In: *Proceedings of the 16th International Conference on Financial Cryptography and Data Security (FC)*. Vol. 7397. Lecture Notes in Computer Science (LNCS). Mar. 2012. DOI: 10.1007/978-3-642-32946-3_3.

[32] Linda Serges. *Q3 2018 Mobile Trade In Data: The iPhone Effect*. Hyla Blog. Oct. 25, 2018. URL: https://blog.hylamobile.com/q3-2018-mobile-trade-in-data-the-iphone-effect (visited on 12/01/2019).

[33] Rebecca Balebako et al. "The Privacy and Security Behaviors of Smartphone App Developers". In: *Proceedings of the Workshop on Usable Security (USEC)*. Feb. 2014. DOI: 10.14722/usec.2014.23006.

[34] Peter Leo Gorski et al. "Developers Deserve Security Warnings, Too: On the Effect of Integrated Security Advice on Cryptographic API Misuse". In: *Proceedings of the 14th Symposium On Usable Privacy and Security (SOUPS)*. Baltimore, MD, USA, Aug. 2018. URL: https://www.usenix.org/system/files/conference/soups2018/soups2018-gorski.pdf (visited on 06/08/2020).

[35] Ken Reese et al. "A Usability Study of Five Two-Factor Authentication Methods". In: *Proceedings of the 15th Symposium On Usable Privacy and Security (SOUPS)*. Santa Clara, CA, USA, Aug. 2019. URL: https://www.usenix.org/system/files/soups2019-reese.pdf (visited on 06/08/2020).

[36] Pew Research Center. *Mobile Fact Sheet*. Pew Research Center: Internet, Science & Tech. June 12, 2019. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ (visited on 06/07/2020).

[37] Daehyun Strobel. "IMSI catcher". MA thesis. Ruhr-Universität Bochum, July 13, 2007. URL: https://www.emsec.ruhr-uni-bochum.de/media/crypto/attachments/files/2011/04/imsi_catcher.pdf (visited on 06/08/2020).

[38] Chris Paget. "Practical cellphone spying". 31st Chaos Communication Congress (31C3). Aug. 2010. URL: http://index-of.es/Miscellanous/LIVRES/cellphonespying.pdf (visited on 06/08/2020).

[39] Altaf Shaik et al. *Practical Attacks Against Privacy and Availability in 4G/LTE Mobile Communication Systems*. Aug. 7, 2017. arXiv: 1510.07563 [cs.CR].

[40] Byeongdo Hong, Sangwook Bae, and Yongdae Kim. "GUTI Reallocation Demystified: Cellular Location Tracking with Changing Temporary Identifier". In: *Proceedings of the 25th Network and Distributed System Security Symposium (NDSS)*. Feb. 2018. DOI: 10.14722/ndss.2018.23349.

[41] Tyler Moore et al. "Signaling system 7 (SS7) network security". In: *The 2002 45th Midwest Symposium on Circuits and Systems (MWSCAS) Conference Proceedings*. Aug. 2002. DOI: 10.1109/MWSCAS.2002.1187082.

[42] Positive Technologies. *SS7 Security Report*. URL: https://positive-tech.com/storage/articles/ss7-security-report-2014-eng.pdf (visited on 06/07/2020).

[43] Karsten Nohl. "Mobile self-defense". 31st Chaos Communication Congress (31C3). Dec. 27, 2014. URL: https://fahrplan.events.ccc.de/congress/2014/Fahrplan/system/attachments/2493/original/Mobile_Self_Defense-Karsten_Nohl-31C3-v1.pdf (visited on 06/08/2020).

[44] Lily Hay Newman. *Fixing the Cell Network Flaw That Lets Hackers Drain Bank Accounts*. WIRED. May 9, 2017. URL: https://www.wired.com/2017/05/fix-ss7-two-factor-authentication-bank-accounts/ (visited on 12/01/2019).

[45] Silke Holtmanns and Ian Oliver. "SMS and one-time-password interception in LTE networks". In: *2017 IEEE International Conference on Communications(ICC 2017)*. Paris, France, May 2017. DOI: 10.1109/ICC.2017.7997246.

[46] Jessica Colnago et al. ""It's not actually that horrible": Exploring Adoption of Two-Factor Authentication at a University". In: *Proceedings of the 2018 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Apr. 2018. DOI: 10.1145/3173574.3174030.

[47] Catherine S Weir et al. "User perceptions of security, convenience and usability for ebanking authentication tokens". In: *Computers and Security* 28 (1–2 2009). DOI: 10.1016/j.cose.2008.09.008.

[48] Elissa M. Redmiles, Michelle L. Mazurek, and John P. Dickerson. "Dancing Pigs or Externalities?: Measuring the Rationality of Security Decisions". In: *EC '18: Proceedings of the 2018 ACM Conference on Economics and Computation*. June 2018. DOI: 10.1145/3219166.3219185.

	Account Information	Device Information		Usage Information	Knowledge	Possession
	Account Number	IMEI	ICCID	Recent Numbers	PIN or Password	SMS OTP*
AT&T	✔	✔	✔	✔	✔	✔
T-Mobile					✔	✔
Verizon					✔	

*We represent SMS OTP as a secure authentication factor because 1) we assume that a carrier sends the SMS OTP exclusively over its own network as a service message, such that the passcode is not vulnerable to routing attacks, and 2) we assume that if an attacker already has the ability to hijack a victim's SMS, a SIM swap does not provide the attacker with additional capabilities.

- generally accepted in the computer security research field
- had not been previously tested but we demonstrate is insecure (for reasons explained in Section 5)
- known to have security shortcomings (also for reasons described in Section 5)

Table 4: Authentication methods we observed at each postpaid carrier. A checkmark means that a type of information was a component of at least one pathway for SIM swap customer authentication; it does not mean that a type of information was necessary or by itself sufficient for SIM swap customer authentication.

A Authentication for postpaid accounts

After completing our data collection on prepaid accounts, engaging with industry stakeholders, and reviewing public disclosures about wireless carrier account security, it appeared likely that authentication practices for postpaid accounts differed from authentication practices for prepaid accounts. We therefore followed our study of prepaid accounts with a study of postpaid accounts at 3 carriers: AT&T, T-Mobile, and Verizon.

We used a similar method for studying the postpaid carriers. Rather than using generated identities, members of the research team signed up with their own credentials. This was to address the additional identify verification process present at postpaid signups. We used the same threat model and script; after one week of usage we called in to request a SIM swap. To the best of our ability, we enabled all available safeguards against SIM swaps at each carrier by configuring our online profiles and calling in soon after to request protections against SIM swaps.[13]

It is important to note that postpaid accounts require real-world identities. Ultimately, we were only able to sign up for one account per carrier using the identities of research personnel. Therefore, the results of this study of postpaid carriers should be interpreted anecdotally. Spotting an authentication factor in this very limited run is some evidence that it is a component of the carrier's customer authentication flow, but not spotting an authentication factor provides little information. In other words, we believe these results are best interpreted as somewhat unlikely to include false positives for authentication factors, but we cannot offer much confidence about false negatives.

The calls were made in December 2019. Our IRB application was submitted in September 2019 and approved in November 2019. Results of our findings are shown in Table 4.

[13] We also enabled the `NOPORT` option for T-Mobile, though our understanding is that the option only applies to port outs and not SIM swaps at present. Our understanding is also that T-Mobile does have additional protections against SIM swaps that can be associated with an account, but only after the account has been the victim of fraud.

B Ethical considerations

Working with our institution's IRB, we took steps to minimize the risk of harm to both research personnel and customer service representatives, primarily by protecting their privacy.

B.1 Minimizing the risk of harm to RAs

We took steps to protect the privacy of the research assistants we hired. During account setup, we were required to provide the name of the account owner. Since prepaid accounts do not require a real-world identity, our protocol allowed RAs to use a fictitious name on the account if they elected for it. We assigned names using an online name generator.

The accounts were at all times controlled by the research team, and only the RA who had been designated as the account owner would be allowed to view information on that account. That is, RAs were not allowed access to accounts assigned to other RAs. The accounts were funded through the duration of the study and closed at the end of the experiment.

B.2 Minimizing the risk of harm to CSRs

We took two preventive measures to minimize the risk of harm to the customer service representatives who handled our calls:

- **Calls were not recorded.** The study design was approved with the parameter that the study procedures not be recorded due to differing laws regarding recordings across the states. Instead, we took detailed notes about the carrier's policies and practices during the call. Our notes do not include references to time of conversation (timestamps), gender, or any other identifying information related to the CSRs.
- **Account information will remain unpublished.** We have not revealed the phone numbers used in our study in order to minimize risk to CSRs. Otherwise, carriers

would be able to track the service history on the accounts and potentially subject pertinent CSRs to disciplinary action (which would also be orthogonal to our study, since our research was designed to obtain information about corporate policies rather than about individuals).

We did not obtain the CSRs' informed consent before interacting with them, because our mitigations listed above ensure that the risks to them are minimal; they are simply carrying out their ordinary responsibilities. Furthermore, our study could not have been conducted with informed consent; firms might decline to participate or misrepresent their policies and practices. We obtained a waiver of consent from the IRB before carrying out our study. The Common Rule specifies a set of criteria for waiver, which we addressed in our IRB application.[14] While we did not debrief CSRs immediately after each SIM swap request, we provided an initial notification of our findings to the carriers we studied and to CTIA in July 2019 (even though our IRB did not impose an ex-post disclosure requirement).

C Website responses to vulnerability reports

In early January 2020, we attempted to notify each of the 17 websites described in Section 7.3 of the presence of doubly insecure configurations. We aimed to make as clear as possible the fact that our report was not merely a recapitulation of the already widely known possibility of SIM swaps and that our report was specific to the victim website's configuration. Shown below is a sample notification:

```
To whom it may concern,

This is a vulnerability disclosure arising out
of security research at Princeton University.
We are computer science researchers affiliated
with the Center for Information Technology
Policy.

example.com currently offers SMS as an account
recovery method. It also offers SMS as an
optional two-factor authentication (2FA)
method. It allows users to simultaneously
choose SMS for account recovery and 2FA. This
means that an attacker who hijacks a user's
phone number can take over their account on
example.com, without a password compromise. We
have attached screenshots that demonstrate this
vulnerability.

We studied the account security measures that
control SIM swaps at five major U.S. carriers.
We found that all five carriers use insecure
authentication challenges that can easily be
subverted, allowing attackers to take control
of a victim's phone number and intercept their
calls and messages.

We also studied 145 websites that offer
phone-based authentication and found 17
websites, including example.com, on which user
accounts can be compromised based on a SIM swap
alone. Currently, in our published dataset, we
have redacted your website's name and other
identifying information (row XYZ). We plan to
release the dataset with all website names in
30 days.

We recommend that you:

  • disable SMS-based account recovery if
    SMS-based 2FA is enabled.
  • recommend more secure 2FA options such as
    authenticator apps to users over SMS.

Please contact us if you have any questions
about our research or recommendations. If
you intend to take any actions to improve
user account security after learning of our
findings, we request that you notify us.
```

Table 5 describes all responses we have received at the time of writing (more than 30 days after initial notification). We coded the responses as follows:

- **"Closed as won't fix"**. The reviewers acknowledged the issue, but decided against mitigation.
- **"Closed as non-issue"**. The reviewers believed the current authentication policy to be adequate.
- **"Did not understand"**. The reviewers did not believe the report was relevant. This includes interpreting our report as customer feedback, and closing our report as out-of-scope.
- **"Fixed without reporting"**. The company mitigated the vulnerability but did not notify us. We discovered the patch during our 60-day re-test.
- **"No response"**. We did not receive any relevant correspondence at the time of writing.
- **"Reported as fixed"**. The reviewers reported to us—at or before the time of writing—that after reviewing our research, the company mitigated the vulnerability.
- **"Template acknowledgement"**. The reviewers acknowledged we had submitted a report on a possible vulnerability in the company's MFA implementation, but the acknowledgment provided no indication that they had read and understood our report. At the time of writing, we had not received any further correspondence.

[14] 45 C.F.R. § 46.116(f).

Website	Available platforms	Response(s)	Default configuration	Days to fix
Adobe	*Security email*, HackerOne	Reported as fixed	Secure	10
Amazon	*Security email*	Closed as won't fix	Doubly insecure	—
Aol (Verizon Media)	*Security email*	No response	Doubly insecure	—
Blizzard	*Security email*	Template acknowledgment; Fixed without reporting	Doubly insecure	—
eBay	*Internal bug bounty*	Reported as fixed	Secure	28
Finnair	*Customer support portal*	No response	Doubly insecure	—
Gaijin Entertainment	*Support email*	Did not understand	Doubly insecure	—
Mailchimp	*Security email*, *BugCrowd*	No response	Doubly insecure	—
Microsoft	*Security email*, internal bug bounty	Did not understand; Fixed without reporting	Doubly insecure	—
Online.net	*Security email**	Reported as fixed	Secure	18
Paypal	*HackerOne*	Did not understand	Doubly insecure	—
Snapchat	*HackerOne*	Reported as fixed	Doubly insecure	38
Taxact	*Support email*	Did not understand; Fixed without reporting	Doubly insecure	—
Venmo	*Support email*	No response	Doubly insecure	—
WordPress.com	*Support email*	No response	Doubly insecure	—
Yahoo (Verizon Media)	*HackerOne*	Did not understand	Doubly insecure	—
Zoho Mail	*Support email*, security email, internal bug bounty	Closed as non-issue**	Secure	—

Table 5: Responses from our vulnerability disclosure, detailed in Section 7.3. Contacted platforms are in italicized font. Only in four of the 17 cases did the process work as expected, resulting in fixes.

*Email address not publicly available, we were provided the address only after sending a Twitter direct message (DM) asking for a reporting address.
**Zoho claims that its current policy—which disallows the same number to be used for recovery and MFA—is secure and does not require any changes.

D Additional related work

SIM swapping is not the only means to intercept calls and SMS messages. There are man-in-the-middle (MITM) attacks that take advantage of weaknesses in mobile phone network infrastructure. For instance, IMSI-catchers [37] can be used to intercept nearby connections on certain older wireless protocols by posing as a mobile tower and forcing phones in the vicinity to connect to it. From there, the IMSI-catcher can force connected phones to use vulnerable encryption or none at all, rendering calls and SMS unprotected. IMSI-catchers take advantage of a weakness in design: legacy cellular networks do not support cell tower authentication. That is, nearby phones are forced to downgrade their connections in order to use legacy cellular network protocols. Though initially used by authorities only, IMSI-catchers can now be built with commercially available components and used by anyone [38].

In Long-Term Evolution (LTE) networks, mobile devices are assigned a Globally Unique Temporary ID (GUTI) in order to alleviate the location-tracking implications of IMSI-catchers. As the name suggests, an temporary identifier is assigned to the device by the access network. The GUTI is then periodically updated to inhibit device tracking. However, as there are no standard guidelines for when and how to update the GUTI, many carriers have been mishandling reallocations either by reusing the same GUTI or assigning predictable identifiers. Shaik et al. showed that repeated calls using Voice over LTE (VoLTE) could reveal a victim's location, since the same GUTI is reallocated [39]. Hong et al. showed that 19 out of 28 carriers across 11 countries were reallocating GUTIs in predictable ways; reallocated GUTIs contained patterns that could be linked back to the previous ones [40]. They also proposed a scalable unpredictable GUTI reallocation mechanism.

There are also weaknesses in the framework that enables carrier interoperability, namely the Signaling System 7 (SS7) protocol, which is designed to trust all requests. The weaknesses of SS7 have long been documented [41]; in 2014, researchers discovered how SMS can be intercepted using the SS7 protocol [42, 43]. Recently, criminals used an SS7 attack to intercept SMS MFA messages for bank accounts, resulting

in financial loss [44].

SS7 has been replaced with Diameter—an improved signaling protocol that supports encrypted requests—with the rollout of 4G and 5G networks, but there are still many carriers in the network that do not use authentication, leading researchers to discover new Diameter-based SMS attacks [45].

While IMSI-catchers and SS7 attacks represent significant threats to the security of mobile communications, SIM swap attacks are inexpensive, low-risk, and as we show, very effective for account hijacking attacks. This makes them attractive to a host of adversaries, including those for whom IMSI-catchers and SS7 attacks are out of reach. Thus, our study focuses on this urgent threat.

There has also been research on customer authentication in other industries. Bonneau et al. examined the use of personal knowledge questions at Google; they discovered that a significant portion of users (37%) provided false answers in order to make them "harder to guess" [12]. Personal knowledge questions among English-speaking users had low rates (60%) of success, as most users could not recall their answers when asked. Colnago et al. [46] observed the deployment of a software token 2FA system at Carnegie Mellon University, and found that while adopters found 2FA annoying, they found it fairly easy to use. The study also found that adopters who were forced to enroll in 2FA had a slightly negative perception of it, as opposed to adopters who were offered to enroll. Weir et al. examined user perceptions of security and usability in online banking, and found that nearly two-thirds of participants chose the device they perceived least secure (but most convenient) as their preference [47]. Redmiles et al. empirically examined the relationship between the proportion of users signing up for SMS-based 2FA based on perceived risk [48]. In the study, users of a testbed bank website were informed of the risks of account hackings and offered to enroll in SMS-based 2FA. Accounts were then randomly selected on a daily basis to be "hacked", weighted by their 2FA settings. The study found that participants were more likely to make these decisions when faced with higher risk.

Lessons Learnt from Comparing WhatsApp Privacy Concerns Across Saudi and Indian Populations

Jayati Dev
Indiana University
Bloomington, IN, USA
jdev@iu.edu

Pablo Moriano*
Oak Ridge National Laboratory
Oak Ridge, TN, USA
moriano@ornl.gov

L. Jean Camp
Indiana University
Bloomington, IN, USA
ljcamp@indiana.edu

Abstract

The purpose of this study is to understand the privacy concerns and behavior of non-WEIRD populations in online messaging platforms. Analysis of surveys ($n = 674$) of WhatsApp users in Saudi Arabia and India revealed that Saudis had significantly higher concerns about being contacted by strangers. In contrast, Indians showed significantly higher concerns with respect to social contact from professional colleagues. Demographics impinge privacy preferences in both populations, but in different ways. Results from regression analysis show that there are statistically significant differences between the privacy behaviors of Saudis and Indians. In both cases, privacy concerns were strongly correlated with their reported privacy behaviors. Despite the differences, we identified technical solutions that could address the concerns of both populations of participants. We close by discussing the applicability of our recommendations, specifically those on transparency and consent, to other applications and domains.

*Research performed while author was at Indiana University. This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

1 Introduction

WhatsApp is a multimedia messaging application with a range of capabilities beyond traditional text messaging: asynchronous chat, photo sharing, video sharing, synchronous voice and video chat, and location sharing [14]. WhatsApp supports two-person conversations, ad-hoc discussions, and larger long-lived structured groups. Due to the penetration of WhatsApp in several non-western countries and it being widely considered for peer-to-peer information sharing, there is an opportunity to study differences and similarities in privacy concerns across these nations and design corresponding privacy defaults for users. Here we focus on two populations where WhatsApp is the dominant social network platform: Saudi Arabia and India. We targeted the nation with the most intensive use of WhatsApp (Saudi Arabia) [33] and the nation where WhatsApp has its largest user base (India) [62]. In fact, WhatsApp is so widely used in Saudi Arabia and India that it is a frequent topic for national debate, having been condemned for being a tool for political propaganda [53] and for instigating mob violence with misinformation [60].

Prior research on social networking applications have found that privacy concerns vary significantly across nationalities [48, 69]. Yet much research has focused on WEIRD (Western, Educated, Industrialized, Rich and Democratic) populations [25]. While there have been independent studies on WhatsApp for Saudi Arabian users [57] and Indian users [18], there is not previous work that has directly compared the privacy choices of these two geographically and culturally distinct populations, without a comparison with western populations where privacy is more broadly studied. Having similar surveys about identical features with the same settings in two very different populations offers an opportunity for a comparison of privacy preferences and behaviors. Privacy concerns of non-WEIRD populations, including Saudis and Indians, are relatively understudied in comparison to WEIRD populations. Platform adoption is also different in non-WEIRD populations with WhatsApp more widely used in Saudi Arabia and India than in the United States or Europe. There is

a need for further studies to understand how origin reflects different priorities for managing boundaries in non-WEIRD populations. To the best of our knowledge, this is the first empirical study that compares users' concerns and attitudes on social networks between Saudis and Indians. To explore how privacy concerns influence privacy behavior, we formulated the following specific research questions:

RQ1: *To what extent are privacy concerns of Saudis different from Indians?*

We found that certain privacy concerns were similar between the participant groups, validating the work by Altman [7] who found similar privacy concerns across non-western cultures. For example, both populations were equally concerned about being added to conversation groups without consent. However, there were differences in other privacy concerns like being contacted by strangers (Saudis more concerned than Indians) and workplace acquaintances (Indians more concerned than Saudis). Even when nationality was the only measure of difference, these differences seemed to stem from cultural values often represented in previous research [51, 57].

RQ2: *To what extent does gender impinge on privacy concerns for both the populations?*

Gender has been found to be an important factor in privacy research since the 1980s [6, 59]. We consider gender separately to measure if it is an important factor in user privacy concern in our population samples as well. For women, the concern about being contacted by a stranger was found to be statistically significant greater for Saudis than for Indians. Conversely, concern towards being contacted by professional contacts for personal interactions is statistically significant greater for Indian women than for Saudi women.

RQ3: *How do privacy concerns affect privacy behavior for both of the populations?*

We analyzed the effect on privacy concern on selected behaviors for each of our participant groups. Privacy behavior was operationalized in terms of feature settings and profile information sharing boundaries which are common for most messaging platforms including WhatsApp.

We found that while privacy concerns did influence privacy behavior, they did so differently in the two groups. For instance, Saudi participants who expressed concerns over being contacted by strangers also reported using the blocking feature. Indian participants who had greater concerns about being added to a group chose to hide when they were last online and saw messages from others. Comparisons about the use of WhatsApp features and profile settings offers insight into if and how individuals are influenced by their nationality, and demographic differences. Informed by prior cross-cultural research on non-western populations, we contribute to the literature on (i) cross cultural concerns of mobile messaging applications, which are dominant in non-western populations and (ii) the differences in privacy behavior of these large populations driven by concerns. In addition to the specific cross-cultural studies, we also hope to contribute to the discussion about methods of research on privacy perceptions and behavior, e.g. [20, 55] by operationalizing behavior in terms of feature and profile settings. The comparison between the two populations using a similar experimental approach may provide general insights on privacy decision-making, which can further inform design considerations on platforms like WhatsApp.

We analyzed survey responses of 674 participants who self-identified as Saudis and Indians. While these participants might or might not be currently residing in the respective countries, they identify themselves as being Saudi or Indian citizens. We used snowball sampling; asking Saudis to share the survey with other Saudis and Indians to share the survey with other Indians. This helped ensure that we captured as many participants possible who were culturally homogeneous.

In the following section, we present an overview of other cross-population privacy studies. Section 3 details our recruitment, data compilation, and analysis procedures. Section 4 presents the results of the analyses of individual populations, and comparisons between them. We then discuss the implications of the results in Section 5, making recommendations for changes in WhatsApp and supporting these with our analysis. We conclude the paper in Section 6, illustrating avenues for future work.

2 Related Work

Prior work has long considered if privacy is a cultural phenomenon. As early as 1977, Altman recognized "privacy [as] a universal process that involves culturally unique regulatory mechanisms" [7]. Thus, there are aspects of privacy that are pervasive across cultures and those that are culturally distinct. However, due to the concentration of technology in western populations, practical reasons have made privacy studies to be largely geographically constrained.

For example, a study of 201 Facebook users in the United Kingdom found that participants' perceived risk of sharing information on Facebook was a significant predictor of privacy concerns and precautionary behaviors [68]. King, Lampinen and Smolen report privacy attitudes to be a consequence of previous events rather than overall risk perception [34]. Lewis, Kaufman and Christakis argue that privacy behaviors are a result of *'social influence'* and *'personal incentives'* [39] such as peer attitudes and nationality biases. This allows an opportunity to study population intrinsic privacy concerns as a predictor of privacy behavior. In privacy research, however, WEIRD populations are not necessarily representative of other populations [25]. It is reasonable to evaluate if privacy research on WEIRD populations predicts findings from South-East Asian and Middle-Eastern populations given that studies of offline risks have consistently found strong evidence that the tolerance for risk [29] and the cultural framing

of risk [9] vary significantly across nations.

2.1 Cross-Cultural Studies in Privacy

The privacy community has been increasingly interested in privacy concerns and attitudes across cultures, often comparing non-western populations against western populations like [44], [69], and [32]). Cvrcek et al.'s study of privacy valuation across Europe found significant differences between Greek, Belgian, Czech, German, and Slovak populations in terms of location privacy indicating the importance of studying culturally varied populations [17]. Further afield, privacy risk perception of German participants were found to be higher than American participants, and both were higher than their Chinese counterparts [29]. A study of 92 participants in three countries found that generally American respondents were the most privacy concerned, followed by the Chinese while Indians showed the least concern [69]. The difference in the WEIRD populations were partly credited to the presence of data protection laws, but nationality also played a role [11, 16, 47]. These smaller studies were also followed by larger studies on privacy concerns of Internet users across different cultural and political settings like [10] and [40] as well as development of universal privacy frameworks [67].

Prior work has also attempted to study individual populations in-depth in Japan [3], Saudi Arabia [1, 73], Bangladesh [4] and India [19]. If privacy attitudes are primarily a function of national attitudes, then examination of privacy in different populations is needed to provide the support for different populations. A core motivation of our work has been to contribute to this rising body of work by not only confirming that there are differences, but also confirming that there are similarities—both of which can be useful in making actionable privacy controls.

2.2 Privacy and Gender

The way different cultures treat men and women also has been found to have an effect on their privacy perceptions [7]. Historically, gender has been considered as "a key social variable in the availability of certain forms of individual and group privacy" [6]. While both men and women are equally subject to invasions of privacy, how these invasions have an effect upon them can vary. Complex gender norms can spill over to cyberspace from the physical world that have a greater impact on women than men (e.g., stalking [38, 57] and family expectations [4]). Female internet users were also "disproportionately [more] prone" to online harm since they formed the greater population of online consumers [8].

Accordingly, there have been a rising number of studies on understanding how gender impinges privacy concerns. A study on American teens found that privacy concerns about receiving unknown emails were higher for female high schoolers [72]. Similarly, greater privacy concerns in women caused them to have enhanced "privacy protection behavior" on Facebook [28]. Few studies have also been focusing on the impact of gender in non-western populations in South-east Asia [59]. An important aspect of privacy is its formulation as a form of 'modesty', where in some cultures, privacy is a way to protect what the society might objectively consider as immoral [70]. Since women tended to have greater privacy concerns and enforced privacy preserving behavior, gender was also a significant consideration in our study on Saudi and Indian populations where culture has a more patriarchal grounding [5, 54].

2.3 Privacy Behavior Against Concerns

Our third research question has been grounded in previous work on privacy behavior resulting from privacy concerns. Research in risk perceptions on various other social media platforms (including Friendster, MySpace, and Facebook) has reported weak correlations between user's privacy choices and their online behavior [2]. Most of the users were unable to or uninterested in addressing privacy settings to control information sharing. The source of this 'privacy paradox' was investigated in a study of 232 Facebook users, where the perceived risk of sharing information was found to be the most important determinant of privacy behaviors. Privacy preferences, measured using a standard Likert scale, were found to be significant but to have the least impact on behavior [22]. These findings were also supported across cultures (e.g. China) [52].

Patil and Kobsa have similarly argued that people are more privacy concerned about specific factors like accessibility of information to strangers, content of the messages in communication, and reliability of the service [50]. Following their example, we have also considered information sensitivity and stranger contact concern as factors influencing privacy behavior. In cases where privacy protecting behaviors are present, they suggest that this is a result of *'impression management'*, specifically in messaging apps [36] at workplaces. Privacy concerns were found to vary based on data type as well as data content. For example, perceptions and valuation of location sharing as a privacy risk vary across contexts and between individuals, and nations [15]. Hence, we also consider if users were concerned with being contacted by colleagues over mobile messaging applications outside of workplace and if it changed how they managed information over messaging platforms (like restricting media download).

2.4 Mobile Messaging Platforms

In this section we discuss WhatsApp as an example of a mobile messaging platfrom. In most studies, e.g. [69], Facebook has largely been the dominant platform studied in cross-cultural privacy research. While Facebook does have a fairly large user base across countries, it is often not the dominant platform used by the majority of the population. With the rise of smartphones, mobile phone based messaging applications like WhatsApp have been increasingly adopted in non-WEIRD populations instead of Facebook [62]. One of

the countries in our study, Saudi Arabia, has the highest WhatsApp market penetration, with 78% of the population using WhatsApp [63]. WhatsApp is also treated as a credible source of information for law enforcement [42]. It is integrated into daily life, in educational institutions, political groups [12] and places of employment, making information dissemination over WhatsApp an important domain to study [13, 58]. Similarly, in India, reacting to the brutality of the photos distributed over WhatsApp, five people were incorrectly identified as kidnappers and killed by the residents of isolated towns [24]. In response, WhatsApp has implemented tagging to indicate that the message was forwarded and limited the ability to forward a specific message to five people to prevent mass forwarding [46]. Though the latter event happened after the breadth of our study, it highlights how WhatsApp is an important focal point in behavior over social media, especially for these understudied populations.

Despite the ubiquity of WhatsApp in daily life in many non-WEIRD countries, it is only recently being considered in social networking research [49]. For example, the ever popular Vinco's *Annual World Map of Social Networks* does not even consider WhatsApp as a social network, but rather as a messaging platform.[1] Given the range of services and group management functionalities of WhatsApp, privacy evaluations of social use are worthwhile. While it is true that it might not forever be the dominant messaging application in non-WEIRD countries like Saudi Arabia and India, research on privacy concerns of these populations may be applicable to competing or future platforms as well.

3 Methods

For this study, we used data collected through a survey instrument targeted at WhatsApp users, above the age of 18, who identified themselves as either Saudis or Indians based on nationality. The instrument was initially developed as a bilingual self-reported survey for Saudis in Arabic and English. We adopted the English version for the Indian population and added questions on Location and use of Live Status, but did not translate the survey into any of the major languages or dialects in India due to the fact that there are too many languages that could effect interpretation of the translated text. English is used for all official government communications in India according to the Official Languages Act of 1963 [65] which makes it convenient for population sampling. While the survey contains both quantitative and qualitative responses, we focus here on the quantitative results in order to gain empirically grounded insights of privacy concerns.

3.1 Population Sampling

WhatsApp is the most widely used instant messaging platform in both Saudi Arabia and India [33, 62]. This allowed us an opportunity for convenience sampling, given researchers from the aforementioned countries. The survey was done in two phases, the first targeted at Saudi users in 2015 and the second targeted at Indian users in 2017. The study was approved by the University's Institutional Review Board (IRB) for both surveys. We conducted snowball sampling to recruit respondents who were culturally homogeneous. It also would have been difficult and possibly infeasible to reach Saudi women with a survey distributed in the United States. We designed the second study (with Indian users) to enable comparison with the first. Given we were excluding Saudis/Indians living outside their country for years (because privacy perceptions can change depending on the country of residence) we retained the snowball sampling method for both similarity and recruitment. Since the data were collected in two phases over a gap of two years, we removed additional features available on WhatsApp to make the variables in the Saudi and Indian groups consistent, as detailed in Section 3.4. The initial survey instrument sampled 820 WhatsApp users; 452 from Saudi Arabia, 146 from non-Saudi Arabs, and 222 from India. The 146 non-Saudi Arabs were excluded from analysis, making the total number of participants 674. This was because non-Saudi Arabs can have membership from 21 Arab countries of Algeria, Bahrain, Comoros, Djibouti, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tunisia, United Arab Emirates, and Yemen in addition to the Palestinian Territories. This would increase the variability since we did not have enough number of users from each of these countries. This gave us two datasets of 452 (Saudi) and 222 (Indian) users respectively.

3.2 Sample Size and Variables

Both Saudi and Indian population datasets were then combined for all 674 valid responses from users who identified themselves as Saudi or Indian nationals respectively. We had 452 Saudi participants (we needed 385 participants for a 95% confidence level and 5% margin of error) and 222 Indian participants (we needed 271 participants for a 95% confidence level and 5% margin of error). Table 1 shows the demographic distribution of both populations across gender, age and educational qualification.

Since our dataset was obtained from a prior survey of Saudi respondents that contained only binary gender, we had to remove non-binary or undeclared responses from seven users in the survey to maintain consistency. Gender categories have been a long standing debate [64] and a follow-up study that accounts specifically for privacy concern variations based on gender would be invaluable. We assigned a variable to every question in our complete dataset, resulting in 19 variables, across four categories: privacy concerns, general usage, demographics, and feature and profile information settings. The number of independent variables in the survey was 12 (Table 2 contains 11 and nationality) for consistency across variables

[1] https://vincos.it/world-map-of-social-networks/

that can be compared for analysis. Questions that were present in one dataset but absent in the other were removed. We also removed responses which had missing values for any of the questions that remained in our dataset.

Demographic Details		Saudis	Indians
Gender	Male	159	140
	Female	293	75
Age (years)	18-24	99	80
	25-30	225	123
	31-40	103	17
	41-50	21	2
	over 50	4	0
Education	High School or Less	30	4
	Some College	38	2
	Bachelors	58	90
	Masters or Professional	18	114
	Doctoral	2	9
		452	222

Table 1: Demographic distribution of the populations broken down into two samples.

Recall the three research questions were about the differences and similarities of privacy concerns (***RQ1***), how demographics impinge these concerns (***RQ2***), and how these concerns affect privacy behavior (***RQ3***). The survey had five questions for each population into the 'Privacy Concerns' category. In order to answer if these concerns were similar or different based on demographics (***RQ2***), we included demographic queries: age, gender, and education. Gender has been found to be a significant factor in use of security technology, and has a significant impact on privacy behavior [21]. Expertise has been found to be a major determinant of security behaviors [20, 56]. Expertise also clearly impinges risk awareness. The differences, or lack thereof, in privacy and security papers sometimes is embedded in expertise, such as works where student samples have predominantly male technology experts and predominantly female non-experts. However, here gender is the only explicit variable, based in no small part on the differences in findings in the comparative works discussed in Section 2. In addition a study of privacy concerns on Facebook that addressed use and risk perceptions found gender to be a significant variable in WEIRD population samples [21].

In order to measure if privacy behavior changed according to privacy concerns for each population (***RQ3***), privacy behavior of participants was operationalized as use of features on WhatsApp. This includes blocking others from accessing personal devices, and limiting the transmitting of profile information. Table 3 lists the dependent variables in our study. We have four dependent variables as 'Feature Settings'. These features include *Blocking*, *Auto Download*, *Location*, and *Notification*. The responses to feature usage questions are binary. Three additional variables for profile settings, *Profile Photo*, *Last Seen*, and *Status* were also included because they signified access control to an individual's profile at three levels - Nobody, My Contacts, and Everyone. These seven variables served as dependent variables in our study. Similarly, Table 2 lists the independent variable categories, 'Privacy Concerns', 'Usage', and 'Demographics' further described below.

1. Privacy Concern Variables: Users were queried about their privacy concerns while using WhatsApp. The survey did not contain generic questions on privacy, but specific questions in the context of WhatsApp. This, however, does not limit the questions to the existence of WhatsApp. The variables under 'Privacy Concerns' like *Sensitive Data*, *Professional Contact*, *Targeted Ads*, *Group Add Ask*, and *Stranger Contact Concern* can be applied across all mobile messaging platforms with similar functionality which ensures that questions on privacy concerns remain relevant beyond the scope of our study.

2. Usage Variables: These variables measure the usage habits of users on WhatsApp. This includes information on the operating system (*Platform*), frequency of usage (*Frequency*) and length of usage (*Length*).

3. Demographic Variables: The demographic variables we considered were *Age*, *Gender*, and level of education (*Education*).

Apart from these, we had a separate variable for nationality that distinguished between Indian and Saudi users. The list of independent and dependent variables and their relationship with the research questions we investigated in this paper are in Table 8 (Appendix).

3.3 Analysis

For our first research question (***RQ1***) we compared privacy concerns between Saudis and Indians using the Mann–Whitney–Wilcoxon (MWW) test. In doing so, we compared the responses to their privacy-related concern questions ('Privacy Concern' variables). All of the five variables were measured on a three-point Likert scale for the Saudi population. The MWW test is a non-parametric test of the null hypothesis that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample [43]. This makes it suitable for dealing for the Likert-scale data used for quantifying privacy concerns in the two samples [71].

We also use the MWW test to answer our second research question (***RQ2***) pertaining to gender differences in privacy concerns. We maintain the separation between the populations, with gender as the additional control variable.

For our third research question (***RQ3***), where we measure the concern factors that influence privacy behavior, we per-

Q. No.	Description	Type
	Privacy Concerns	
1	*Sensitive Data:* I frequently use WhatsApp to send/share private or sensitive chats/media.	Likert
2	*Professional Contact:* Do you use WhatsApp to communicate formally or informally with your professional contacts, like your boss or coworkers?	Likert
3	*Targeted Ads:* Are you concerned that since Facebook bought WhatsApp, targeted ads might start appearing in WhatsApp?	Likert
4	*Group Add Ask:* When adding me to a group chat, I would like the app to (ask/ not ask) me before adding.	Likert
5	*Stranger Contact Concern:* Are you concerned that anyone who has your phone number is able to contact you and see the activity shared publicly using WhatsApp?	Likert
	Usage	
6	*Platform:* Which operating system do you currently use for your primary smartphone?	Categorical
7	*Frequency:* On average, how often do you use WhatsApp?	Categorical
8	*Length:* How long have you been using WhatsApp?	Categorical
	Demographics	
9	Age	Categorical
10	Gender	Boolean
11	Education	Categorical

Table 2: Privacy concerns, usage and demographic factors which are independent variables for both population samples.

formed an exploratory factor analysis (EFA) along with logistic and and ordered logistic regression. EFA was used to find influential underlying factors from a set of observed independent variables. We used the Privacy Concerns, Usage, and Demographics variables to form a new set of variables that would be independent of each other. EFA extracts the maximum variance from all the variables and groups them under a common score. A latent factor representation of the independent variables allows us to deal with multicollinearity. When the degree of collinearity is high between independent variables, it becomes difficult to estimate the relationship between each independent variable and the dependent variable, as well as, the overall precision of the estimated coefficients.

EFA helps in finding the relationship between independent variables in terms of a smaller set of factors. We tested the adequacy of conducting EFA for both samples using the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy [26] and the Bartlett's test of Sphericity [66]. The KMO measure is a statistic that indicates the proportion of variance in the dependent variables that might be caused by underlying factors. The Bartlett's test of Sphericity measures the hypothesis that the correlation matrix is an identity matrix, which would indicate that the independent variables are unrelated and therefore unsuitable for EFA. We used orthogonal rotation with the varimax method to force the latent factors to remain uncorrelated. We considered an item to be loaded on a factor if its loading exceeded 0.3 and repeated it for each sample. Finally, we used logistic and ordered logistic regression to analyze the influence of the independent variables in the privacy attitudes of both samples. This results inform the discussion about how nationality-based mental models have an impact of the privacy preferences (measured through feature and profile setting choices) of users. We used the *psych* package in R to run the EFA and regression analyses.

3.4 Considerations and Limitations

In the two years between the studies there were no significant user interface changes; however, WhatsApp did add new features. These features were *Live Location* (real-time location sharing), document sharing, making phone calls, and end-to-end encryption for text messages. We have tried to account for the changes in features by measuring respondents' settings on these features independently. We acknowledge that there is a possibility of a shift in general privacy attitudes over two years of the survey. We also noted that there were no changes in WhatsApp during this time period that addresses our recommendations and yet hope that these may be adopted. WhatsApp had a $1 yearly subscription fee beginning in 2013. This was eliminated in 2016 [61] which might have an effect on its adoption in both countries.

This study is focused on participants with different origins, but not necessarily culture. The distinction between culture and origin is profoundly important, nuanced, subject to a vast literature [27]. This discussion is beyond the scope of this work but could be the subject of further research.

Due to the use of snowball sampling, our sample is not statistically representative of the populations. However, it does provide valuable insight into the privacy concerns of respondents. Snowball sampling was responsible for the demographic skew even as it helped reach our participants. De-

mographics is one of the many factors that affect findings like workplace contact concerns. It is reasonable to assert that a representative sample would include the issues we address (like the prevalent use of WhatsApp in workplaces) and our results would still hold. Nevertheless, a more representative sample that accounts for nuanced cultural differences within countries would be beneficial for further research and would likely expand the recommendations.

4 Findings

Following the quantitative analysis above, we categorized our findings in three sections. First, we report privacy concerns based on nationality. Next, we report privacy concerns based on gender. Finally, we present how privacy behavior of users was dependent on privacy concerns.

4.1 Privacy Concerns Based on Nationality

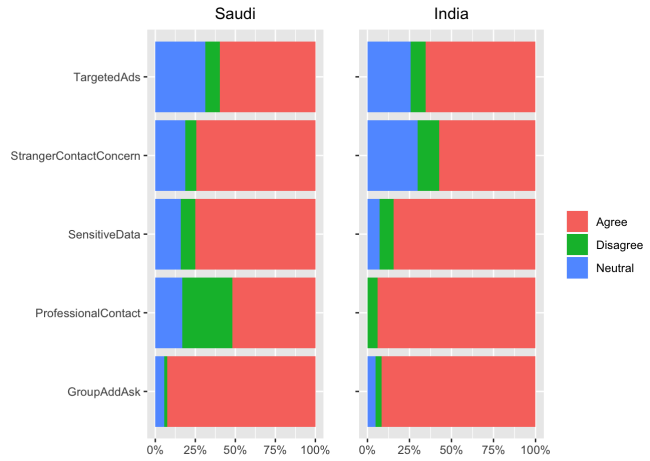
Overall, we noticed that while Saudi and Indian participants have different privacy concerns towards WhatsApp, there are concerns that are also similar for both population samples. Figure 1 shows the differences regarding the privacy concerns between Saudis and Indians. Specifically, we found that the higher concerns towards being contacted by strangers is statistically significant for Saudis than for Indians ($W = 37,254, p < .001$). On the other hand, we found that Indians tend to be more privacy concerned than Saudis regarding being contacted by professional contacts. Indian respondents also seemed to have greater privacy concerns about being subjected to targeted advertisements over WhatsApp. Concerns towards being contacted by a professional contact ($W = 66,299, p < .001$) and targeted advertisements ($W = 56,408, p < .001$) that uses data from their conversations are significantly greater for Indians than for Saudis respectively.

However, we did not find significant differences between concerns about *Sensitive Data* and *Group Add Ask*. Their concerns over sharing sensitive information and preferences for being asked before someone adds them to a group were similar. Respondents from both samples shared data which they believed was sensitive, and expressed displeasure at being added to WhatsApp groups without consent.

4.2 Privacy Concerns Based on Gender

Figure 2 compares the privacy concerns between Saudis and Indians by gender. We tested the same five categories pertaining to privacy concerns (*Targeted Ads*, *Stranger Contact Concern*, *Sensitive Data*, *Professional Contact*, and *Group Add Ask*) as seen in Table 2, splitting the two samples by gender and using the MWW test. We found that privacy concerns between genders within the same sample (for both Saudis and Indians) are not statistically significant. In other words,

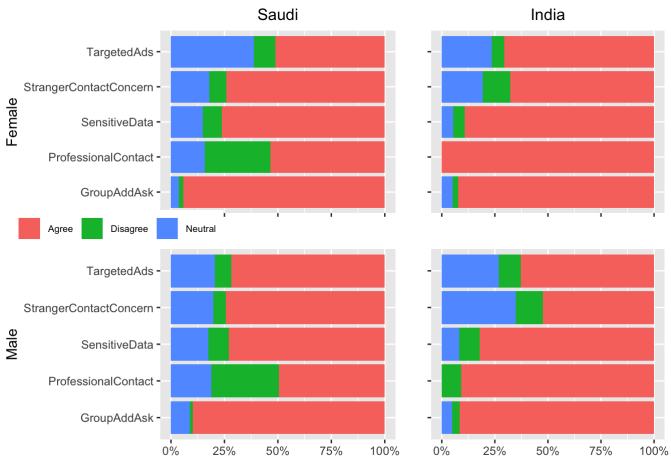
Figure 1: Privacy concerns of respondents by nationality.



within the same sample, both men and women expressed the same level of privacy concerns.

However, when we compared the two samples by gender individually, the concerns were different. For females, we found that the concern towards being contacted by a stranger is statistically significant greater for Saudis than for Indians ($W = 11,648, p = .015$). Similarly, we found that the concern towards being contacted for professional contact is statistically significant greater for Indians females than for Saudis females ($W = 6,569, p < .001$). We also compare

Figure 2: Privacy concerns of Saudi and Indian respondents by gender.



the privacy concerns between samples for males. We found that the concern towards being contacted by a stranger is statistically significant greater for Saudi males than for Indian males ($W = 12,222, p < .001$). Conversely, we found that the concern towards being contacted for professional contact is statistically significant greater for Indians than for Saudis ($W = 8,545, p < .001$). Overall, results across gen-

Q. No.	Description	Type
	Feature Settings	
1	*Blocking:* Did you use the Blocked feature in WhatsApp to block any person from contacting with you?	Boolean
2	*Auto Download:* Did you disable the auto-download feature on WhatsApp?	Boolean
3	*Location:* Have you previously shared your location using WhatsApp?	Boolean
4	*Notification:* Have you enabled WhatsApp to send you notifications when there is a new message?	Boolean
	Profile Information	
6	*Profile Photo:* What is your setting? (Everyone, My Contacts, Nobody)	Categorical
7	*Last Seen:* What is your setting? (Everyone, My Contacts, Nobody)	Categorical
8	*Status:* What is your setting? (Everyone, My Contacts, Nobody)	Categorical

Table 3: Feature settings and profile information which are dependent variables for both population samples.

ders between different samples are consistent for females and males. Thus, the privacy concerns reflected by overall participants were also reflected by gender with no gender differences within each sample.

We did not find significant differences between *Sensitive Data* and *Group Add Ask* privacy concerns when comparing genders between the two samples.

4.3 Comparing Privacy Behavior Based on Privacy Concerns

Following comparison between Saudi and Indian users of WhatsApp against their privacy concerns, we tested to see if privacy behavior in each sample differed based on privacy concerns. Privacy behavior was measured in terms of user behavior in using WhatsApp features like *Blocking* (restricting access to self), *Auto-Download* (allowing automatic download of media files), *Location* (sharing static map coordinates), and *Notification* (enabling notifications on device from WhatsApp when user receives a message). We used three additional features pertaining to user profile information - *Profile Photo*, *Last Seen* (when the user had last checked their WhatsApp messages) and *Status* (static description of users about themselves). These had three levels of access control - Everyone, Contacts only, and Nobody. We removed *Live Location* (ability to share location continuously instead of static coordinates), *Read Receipts* (blue ticks showing when messages have been delivered and seen), and *Live Status* (instant story posts visible for 24 hours) which were not present across both surveys.

4.3.1 Exploratory Factor Analysis (EFA)

Saudi Users: For Saudis, the variable to subject ratio was 1:41.1 (452/11 = 41.1). This shows that the number of participants per question was adequate to obtain quality in the factor solution [35]. The Kaiser-Meyer-Olkin (KMO) test statistic for sampling adequacy was 0.50 which makes this dataset acceptable for EFA. In addition, the Bartlett's Test of Sphericity [66] revealed that the correlation matrix came from independent samples ($\chi^2 = 177.1$, $df = 55$, $p < .05$), and further indicated that the factor analysis was justified by the properties of the correlation matrix. Therefore, EFA is considered as an appropriate technique for further analysis of this sample.

	Age & Education	Usage Platform	Group Permission	Gender	Information Sensitivity	Targeted Ads
Age	0.962					
Education	0.490					
Platform		0.992				
Group Add Ask			0.671			
Gender				0.565		
Sensitive Data					0.418	
Frequency					0.316	
Targeted Ads						0.461
Stranger Contact						
Concern						
Length						
Professional Contacts						
Eigenvalues	1.465	1.284	1.213	1.101	1.03	1.011
% of Total Variance	10.7	9.1	4.8	3.9	3.7	2.6

Table 4: Rotated varimax factors from the factor analysis of privacy concerns of Saudi participants.

We identified and extracted factors based on the Kaiser's criterion for eigenvalues, i.e., we choose all factors with an eigenvalue greater than 1 as a measure of reliability [31]. Figure 3 shows the scree plot of successive eigenvalues. Specifically, it shows that after the sixth the total variance accounts for smaller amounts. The rotated factor loadings along with the eigenvalues are illustrated in Table 4.

The six selected factors predicted 34.8% of the variance, including: variables that are related to demographics (i.e., *Age and Education*), operating system type (i.e., *Usage Platform*), privacy issues for being added to a group chat (i.e., *Group Permission*), *Gender*, privacy issues to send private data and frequency (i.e., *Information Sensitivity*), and being targeted by advertisements (i.e., *Targeted Ads*). Each of these components accounted for 10.7%, 9.1%, 4.8%, 3.9%, 3.7% and 2.6% of variance respectively.

Indian Users: For Indians, the variable to subject ratio was 1:20.1 (221/11 = 20.1). Similar to Saudi respondents, the number of participants per question was adequate to obtain quality

Figure 3: Scree plot showing eigenvalues for EFA among Saudi users. There are six values that are greater than 1.

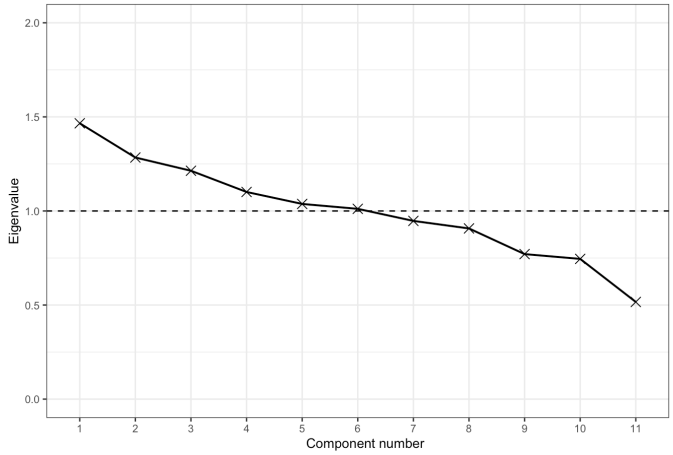
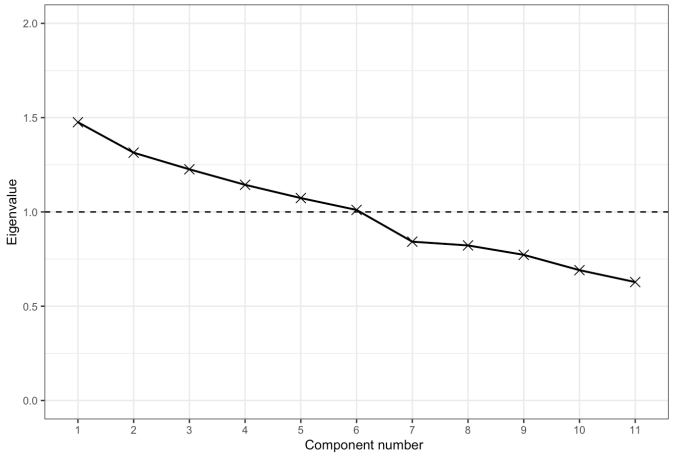


Figure 4: Scree plot showing eigenvalues for EFA among Indian users. There are six values that are greater than 1.



in the factor solution [35]. The KMO test statistic was 0.50 which makes this dataset acceptable for EFA. The Bartlett's Test of Sphericity revealed that the correlation matrix came from independent samples ($\chi^2 = 70.5$, $df = 55$, $p < .05$), and further indicated that the factor analysis was justified by the properties of the correlation matrix. Therefore, EFA is considered as an appropriate technique for further analysis of this sample as well.

We retained only the factors which had eigenvalues greater than 1 in accordance with Kaiser's criterion like we did for the Saudi samples [31]. As can be seen Figure 4 shows the scree plot of successive eigenvalues, of which we selected six factors. The rotated factor loadings are illustrated in Table 5.

The six selected factors predicted 34% of the variance, including: variables that are related to privacy regarding *Sensitive Data*, type of operating system (i.e., *Usage Platform*), frequency of using WhatsApp (i.e., *Usage Frequency*), privacy issues for being added to a group chat and education (i.e., *Education and Group Permission*), privacy issues related to being targeted by advertisements and age (i.e., *Age and Targeted Ads*), and privacy issues regarding being contacted by professional contacts (i.e., *Professional Contact*). Each of these components accounted for 9.1%, 5.8%, 5.2%, 4.7%, 4.7% and 4.5% of variance, respectively.

	Sensitive Data	Usage Platform	Usage Frequency	Education & Group Permission	Age & Targeted Ads	Professional Contacts
Professional Contact						0.676
Targeted Ads					0.609	
Age					0.340	
Group Add Ask				0.504		
Education				0.353		
Frequency			0.587			
Platform		0.741				
Sensitive Data	0.981					
Stranger Contact						
Concern						
Length						
Gender						
Eigenvalues	1.475	1.314	1.073	1.010	1.143	1.225
% of Total Variance	9.1	5.8	5.2	4.7	4.7	4.5

Table 5: Rotated varimax factors from the factor analysis of privacy concerns of Indian participants.

After EFA for each sample, we tested the six factors obtained in each case against privacy behavior of respondents.

4.3.2 Findings from Regression Analysis

We used factor scores (i.e., each example weight into its factor loading) from the EFA as the predictors of privacy attitudes using multiple logistic regression (for *Blocking*, *Auto Download*, *Location*, and *Notification*) and multiple ordinal logistic regression (for *Profile Photo*, *Last Seen*, and *Status*) based on the type of dependent variable. In doing so, each example contribution to the factor score depends on how strongly it relates with the factor. The reported regression coefficients are based on the logarithm of the odds. Odds are the probability of an event occurring divided by the probability of the event not occurring. In our results, that means that for every one unit of gain in the independent variable, the logarithm of odds of the dependent variable increases by the correspondent coefficient. For details about the interpretation of the regression coefficients, we suggest the reader the work in [23].

Saudi Users: Table 6 shows the coefficients of the regression model for Saudi respondents with statistically significant p-values. As seen in the table, most factors only effect one of the feature settings. Given that the level of significance is (*) for $p < .05$, (**) for $p < .01$, and (***) for $p < .001$, we observe the following.

Gender is statistically significant and contributes (0.498) to the overall inclination to use the *Blocking* feature. Thus, female Saudi users were more likely to use this feature to block access to them on WhatsApp if it was from a person they did not know.
When respondents were more likely to be asked before being

added to a group, the likeliness that they would not download content from another user was higher, indicated by a high coefficient in case of *Auto Download* (0.416) (note that the question for *Auto Download* is framed in terms of disabling the feature).
Age and Education (-0.2789) as well as privacy concerns about receiving targeted advertisements (-0.3067) were linked to user choice of turning notifications on or off. Older and more educated users were more likely to turn off notifications. Similarly, users with higher concerns about targeted advertisements turned their notifications off. This indicates that users who had more experience and knowledge of WhatsApp, coupled with an aversion to targeted advertisements were more likely to avoid notifications.
Privacy concerns did not significantly affect *Location* settings and profile settings like *Profile Photo*, *Last Seen*, and *Status* for Saudi users.
Usage variables like type of operating system (*Usage Platform*) did not effect privacy behavior. Similarly, willingness to share sensitive data frequently (*Information Sensitivity*) did not effect privacy behavior.

Indian Users: Table 7 shows the coefficients of regression analysis for Indian respondents. We found that privacy concerns did effect privacy behavior in case of Indian users as well, but their privacy behavior was very different from Saudis.
Privacy concerns or demographics did not effect feature settings like *Blocking* (which was effected by gender in case of Saudis) and *Location* sharing. However, unlike Saudi users, privacy concerns did effect how Indian users changed their profile settings like *Profile Photo* and *Last Seen*. *Last Seen* had a high value of regression coefficient (0.32010) and was more likely to be hidden by more educated users who were concerned about being added to a group with consent (*Education and Group Permission*). A possibility is that they chose to hide when they were last online when more people had access to them.
Privacy concerns about being contacted by colleagues from a workplace (*Professional Contact*) had a higher coefficient for *Auto Download* (0.35605) and Notification (0.420984) settings.
Similar to Saudi users, concerns about sensitive data did not play a role in effecting privacy behavior. More broadly, education and group permissions were two overlapping factors that effected privacy behavior in both samples. While gender and targeted advertisements played a more important role for Saudi users, privacy behavior of Indian users depended on professional contacts. Profile settings were effected by privacy concerns more for Indian users than Saudi users.

5 Discussion and Implications

Our findings reify and add to previous results on the relationship between culture and privacy. The major goal of this study was to highlight both similarities and differences between two culturally distinct non-WEIRD populations. Our findings expand existing literature on cross-cultural privacy in the SOUPS community. We discuss practical recommendations for more inclusive privacy design choices for non-WEIRD countries; for example, stranger contact concerns and work-personal boundaries.

We found that **participants within each individual sample had similar privacy concerns**. Neither gender nor origin alone were significant determinants; however, **gender effected privacy controls among Saudi users**. Not only did female users had greater privacy concerns about being contacted by strangers over their Indian counterparts, but also this concern effected the use of the blocking feature (to prevent strangers from contacting them).

We have also shown that privacy concerns were dependent on origin, but there were aspects common across the non-WEIRD samples in the study. **Both participant groups shared sensitive content over WhatsApp. Furthermore, this sharing was not found to have an effect on their privacy behavior**. Both groups also had similar group privacy concerns and disliked being added to a group without their consent. However, this was expressed in different ways in different samples. While Saudis restricted **content** from auto-downloading, Indians restricted **visibility access** (being seen by other people that they were online) by hiding their *Last Seen*.

In terms of privacy behavior, neither Saudi nor Indian users changed settings based on the sensitivity of information **content**, but rather information **recipient**. This was particularly true for restricting stranger contact (among Saudis) and preventing use of WhatsApp for professional contacts (among Indians).

Based on our findings, we make the following design recommendations to the already significant security and privacy features of WhatsApp (and future social messaging platforms). While these are specific to WhatsApp, the feature settings offered by WhatsApp and the privacy concern variables that have been operationalized in our study (*Targeted Ads*, *Stranger Contact Concern*, *Sensitive Data*, *Professional Contact*, and *Group Add Ask*) can be observed in other mobile messaging applications like Signal as well.

5.1 Offer an Option for Permissions-Based Contact

Being contacted by strangers was disliked by Saudi female respondents. While we do not suggest a gender-based control, adhering to stronger privacy concerns might improve the privacy of the overall platform. One of the peculiarities of WhatsApp is that contact information of users is easily accessible if they are in a common group, even after blocking particular users. A problem that might happen in large groups in that strangers can contact users over other communication

Factors	Blocking	Auto Download	Location	Notification	Profile Photo	Last Seen	Status
Intercept	1.208***	0.427***	0.722***	1.291***			
Age and Education	-0.024	0.088	-0.157	-0.279*	0.013	-0.021	-0.058
Usage Platform	-0.139	0.117	0.117	-0.217	-0.027	0.002	-0.068
Group Permission	0.173	0.416***	0.034	-0.241	-0.109	0.060	0.069
Gender	0.498***	0.158	-0.165	-0.167	-0.035	-0.143	-0.041
Information Sensitivity	0.044	-0.177	-0.027	0.138	0.169	0.065	-0.038
Targeted Ads	0.073	0.161	0.079	-0.307*	-0.152	-0.165	-0.205

Statistical significance levels are indicated as: (*) for $p < .05$, (**) for $p < .01$, (***) for $p < .001$.

Table 6: Significant regression co-efficient values for privacy behavior measured against the six EFA factors as dependent variables for Saudi respondents.

Factors	Blocking	Auto Download	Location	Notification	Profile Photo	Last Seen	Status
Intercept	1.292***	-0.384*	1.244***	2.303***			
Sensitive Data	-0.159	0.194	0.082	-0.452	0.152	-0.016	-0.096
Usage Platform	-0.164	0.031	-0.056	-0.169	0.353*	0.018	0.253
Usage Frequency	-0.153	0.079	0.254	0.006	-0.098	-0.038	-0.032
Education and Group Permission	-0.141	-0.017	0.051	0.0368	0.304	0.320*	0.215
Age and Targeted Ads	0.101	0.170	-0.267	-0.168	0.169	0.025	0.084
Professional Contact	-0.438	0.356*	-0.043	0.421*	0.081	0.184	-0.137

Statistical significance levels are indicated as: (*) for $p < .05$, (**) for $p < .01$, (***) for $p < .001$.

Table 7: Significant regression co-efficient values for privacy behavior measured against the six EFA factors as dependent variables for Indian respondents.

channels if they happen to be in a common group. While the current version of WhatsApp asks the user if they want to receive communication from someone not in their contact list, it does not prevent strangers from reaching users over other communication channels like text messages or phone calls.

A way to enforce this would be to replace contact information sharing with username (or similar) sharing. This would protect both users who are stranger-averse and those being cyber-bullied. A way to ensure advanced permission-based contact would be to allow a mechanism for **cooperative blocking**. Due to the nature of WhatsApp as a messaging platform, there is not a centralized way to report platform abuse. Blocking is limited to individuals even though WhatsApp is used as a social platform with large groups. Allowing communities to self-organize and block individuals collectively would enhance the usability of the platform and the autonomy of users. This is also a practice that can be extended to messaging applications which do not have a permission-based contact mechanism in place.

5.2 Choice and Consent in Joining Groups

Respondents from both the samples in our dataset wanted to be asked before being added to a group. Given that users restricted content and access based on group permissions, it is likely that this is more than a social construct and privacy controls that allow users the ability to consent before being added to a group must exist. Messaging applications at large would benefit from this consent process as group communication become increasingly prevalent.

5.3 Option for Group Types

Saudi users seemed to use WhatsApp in a more personal setting given that being contacted by colleagues was not significantly higher. On the other hand, WhatsApp is frequently used in India for contacting colleagues in workplaces and other professional contacts. Our findings indicate that privacy behavior on WhatsApp was effected by user choice to interact with professional contacts. This is possibly because users have different self-presentation for their personal and professional lives [51]. Our findings amplify previous research on WhatsApp as well where qualitative research has stressed the importance of "communication places" to separate group interaction over WhatsApp [45].

Enabling easy *segregation of users into high level groups such as work and family* and having audience based information boundaries would ensure that users are able to share selectively without worrying about boundary management between close members and co-workers (especially with more broadcast features like *Live Status*). Such group-based access control can be applied to other messaging applications as well.

While the above design implications are yet to be tested, they extend directly from our findings and serve as some of the

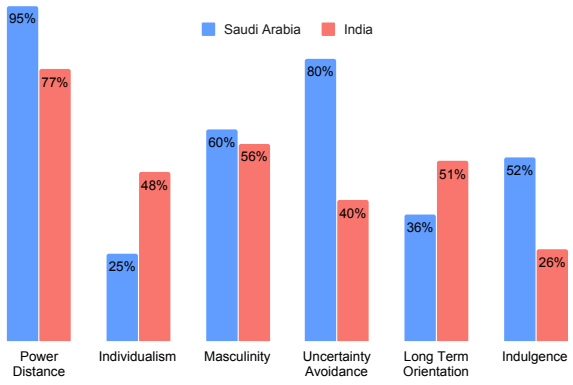
possibilities to make WhatsApp (and possibly other similar messaging applications) more privacy-sensitive and to cater to complex privacy expectations, enhance risk communication and improve trust. Privacy behavior variables like *Blocking*, *Auto Download*, *Location*, *Notification*, *Profile Photo*, *Last Seen*, and *Status* are present in other messaging platforms like Facebook Messenger and Signal as well, and can be similarly studied to view the effect of privacy concerns over behavior.

Though measuring the different dimensions of culture elaborated in Hofstede's work is not covered in the breadth of this study, the influence of nationality seems to hint at the underlying cultural values that affect both privacy concerns and privacy behavior. Figure 5 shows the Hofstede's cultural values measured in terms of power distance, individualism, masculinity, uncertainty avoidance, long term orientation, and indulgence [27] (adopted from Hofstede Insights [30]). Some of these factors may help explain the privacy behavior that we observed in our study. For example, the greater likelihood of Saudis to avoid uncertainty in social situations might drive their reluctance to interact with strangers. Similarly, higher sense of individualism among Indians might explain why they did not want to be contacted by colleagues beyond their workplace. However, these are only conjectures and a future study could help explain the same.

6 Conclusion

Our findings indicate that privacy concerns had both similarities and differences between Saudi and Indian users, both of which were non-WEIRD populations. These privacy concerns combined with demographics like gender affected the privacy behavior of users on WhatsApp in very specific and distinct ways.

Figure 5: Hofstede's cultural dimensions in Saudi Arabia and India (estimates).



Privacy behavior had differences between populations. However, there were also similarities. This allows an opportunity for mobile messaging platforms to enforce both universal and culture-specific boundary management. Privacy behavior was also socially situated, with Indian participants most likely to hide change feature settings to restrict content from professional contacts rather than friends or family. **Most participants in both populations wanted to be able to control the content and recipient (with a greater focus on recipients) of their shared information**.

A core observation, and one which calls for more research is that WhatsApp is experienced as a social network application rather than a messaging application. The embedded use of large groups and workplace-linked contact norms indicate that user perception of WhatsApp is very socially grounded. This leaves an opportunity for a more nuanced re-examination of how privacy settings are implemented. However, our findings cannot be generalized across all non-western populations. A more large scale study with different nationalities across different messaging platforms would be a richer description of privacy concerns in mobile applications. Nevertheless, the results serve to inform the importance of inclusiveness in design choices for privacy–impinging technologies that reach across the globe.

We hope that a brief comparative study would highlight some of the more culturally and socially grounded privacy choices that non-WEIRD populations make. An in-depth cultural study would add to the findings in explaining the norms behind why these privacy choices come into practice in such populations. Religion, political climate, economic models, personal freedom, and societal norms among many others could influence the way people interact with others on mobile messaging platforms, of which WhatsApp is a widely used example.

We used datasets from two populations available to us in order to study cross-cultural privacy concerns and behavior. Studies which include participants from multiple countries, would inform a richer perspective on how different cultures value privacy. In addition, studying the different cultural aspects of privacy might be useful not only to make messaging platforms like WhatsApp sensitive to privacy preferences but also identify areas where cultures reconcile, and create a shared notion of privacy defaults. As mobile messaging platforms increase, largely due to low data usage in non-WEIRD countries, future studies could address different aspects of lesser studied populations to informa privacy choices.

Another aspect that would benefit from follow-up research would be the effect of technical expertise on privacy choices in non-WEIRD populations. Technical expertise [20, 56] and innate privacy sensitivity [37, 41] have long been hypothesized as a measure of privacy behavior and could be implemented in future work.

Acknowledgments

We would like to thank Sanchari Das and Yasmeen Rashidi for their valuable support in implementation of the surveys and data collection. This research was supported in part by the National Science Foundation under CNS 1565375.

References

[1] Norah Abokhodair and Sarah Vieweg. Privacy & Social Media in the Context of the Arab Gulf. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, pages 672–683, 2016.

[2] Alessandro Acquisti and Ralph Gross. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. In *International Workshop on Privacy Enhancing Technologies*, pages 36–58. Springer, 2006.

[3] Andrew A Adams, Kiyoshi Murata, and Yohko Orito. The Japanese Sense of Information Privacy. *AI & society*, 24(4):327–341, 2009.

[4] Syed Ishtiaque Ahmed, Md Romael Haque, Jay Chen, and Nicola Dell. Digital Privacy Challenges with Shared Mobile Phone Use in Bangladesh. 2017.

[5] Madawi Al-Rasheed. *A Most Masculine State: Gender, Politics and Religion in Saudi Arabia*. Number 43. Cambridge University Press, 2013.

[6] Anita L Allen and Erin Mack. How Privacy Got its Gender. *N. Ill. UL Rev.*, 10:441, 1989.

[7] Irwin Altman. Privacy Regulation: Culturally Universal or Culturally Specific? *Journal of social issues*, 33(3):66–84, 1977.

[8] Ann Bartow. Our Date, Ourselves: Privacy, Propertization, and Gender. *USFL Rev.*, 34:633, 1999.

[9] Ulrich Beck and Cordula Kropp. Environmental Risks and Public Perceptions. In *Handbook on Environment and Society*, chapter 41, pages 1205–1216. Sage, 2007.

[10] Steven Bellman, Eric J Johnson, Stephen J Kobrin, and Gerald L Lohse. International Differences in Information Privacy Concerns: A Global Survey of Consumers. *The Information Society*, 20(5):313–324, 2004.

[11] Susan Bennett et al. GDPR: Change to European Privacy Laws and its Impact on Australian Businesses. *Governance Directions*, 70(2):85, 2018.

[12] Victor S. Bursztyn and Larry Birnbaum. Thousands of Small, Constant Rallies: A Large-Scale Analysis of Partisan WhatsApp Groups. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, page 484–488, New York, NY, USA, 2019. Association for Computing Machinery.

[13] Josemar Alves Caetano, Gabriel Magno, Marcos Gonçalves, Jussara Almeida, Humberto T. Marques-Neto, and Virgílio Almeida. Characterizing Attention Cascades in WhatsApp Groups. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 27–36, New York, NY, USA, 2019. Association for Computing Machinery.

[14] Karen Church and Rodrigo de Oliveira. What's Up with Whatsapp? Comparing Mobile Instant Messaging Behaviors with Traditional SMS. In *Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '13, page 352–361, New York, NY, USA, 2013. Association for Computing Machinery.

[15] Sunny Consolvo, Ian E Smith, Tara Matthews, Anthony LaMarca, Jason Tabert, and Pauline Powledge. Location Disclosure to Social Relations: Why, When, & What People Want to Share. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 81–90. ACM, 2005.

[16] Marc Cornock. General Data Protection Regulations (GDPR) and Implications for Research. *Maturitas*, 111:A1 – A2, 2018.

[17] Dan Cvrcek, Marek Kumpost, Vashek Matyas, and George Danezis. A Study on the Value of Location Privacy. In *Workshop on Privacy in Electronic Society*, pages 109–118. ACM, 2006.

[18] Jayati Dev, Sanchari Das, and L Jean Camp. Poster: Understanding Privacy Concerns of WhatsApp Users in India. In *Symposium and Bootcamp on Hot Topics in the Science of Security*, page 28. ACM, 2018.

[19] Jayati Dev, Sanchari Das, and L Jean Camp. Privacy Practices, Preferences, and Compunctions: WhatsApp Users in India. In *International Symposium on Human Aspects of Information Security & Assurance (HAISA)*, page 135, 2018.

[20] Serge Egelman and Eyal Peer. Predicting Privacy and Security Attitudes. *SIGCAS Comput. Soc.*, 45(1):22–28, February 2015.

[21] Vaibhav Garg, Kevin Benton, and L Jean Camp. The Privacy Paradox: a Facebook Case Study. In *2014 TPRC conference paper*, 2014.

[22] Vaibhav Garg and L Camp. Ex Ante vs Ex Post: Economically Efficient Sanctioning Regimes for Online Risks. *Research Conference on Communication, Information and Internet Policy*, 2013.

[23] David A Grimes and Kenneth F Schulz. Making Sense of Odds and Odds Ratios. *Obstetrics & Gynecology*, 111(2):423–426, 2008.

[24] Swati Gupta and Bard Wilkinson. WhatsApp India: Five Lynched after Online Child Kidnap Rumors. https://www.cnn.com/2018/07/02/asia/india-lynching-whatsapp-intl/index.html, July 2018. (Accessed on 02/15/2019).

[25] Joseph Henrich, Steven J Heine, and Ara Norenzayan. Most People are not WEIRD. *Nature*, 466(7302):29, 2010.

[26] Brent Dale Hill. *Sequential Kaiser-meyer-olkin Procedure as an Alternative for Determining the Number of Factors in Common-factor Analysis: a Monte Carlo Simulation*. PhD thesis, Oklahoma State University, 2011.

[27] Geert Hofstede. Cultural Dimensions. *LIUC (IT)*, 13, 2003.

[28] Mariea Grubbs Hoy and George Milne. Gender Differences in Privacy-Related Measures for Young Adult Facebook Users. *Journal of Interactive Advertising*, 10(2):28–45, 2010.

[29] Christopher Hsee and Elke U Weber. Cross-cultural Differences in Risk Perception, but Cross-Cultural Similarities in Attitudes Towards Perceived Risk. *Management Science*, 44(9):1205, 1998.

[30] Hofstede Insights. Cross-Country Comparison. https://www.hofstede-insights.com/country-comparison/, April 2020. (Accessed on 04/02/2020).

[31] Henry F Kaiser. The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, 20(1):141–151, 1960.

[32] Naz Kaya and Margaret J Weber. Cross-cultural Differences in the Perception of Crowding and Privacy Regulation: American and Turkish Students. *Journal of environmental psychology*, 23(3):301–309, 2003.

[33] Simon Kemp. Saudi Arabia Social Media Statistics 2018. https://www.globalmediainsight.com/blog/saudi-arabia-social-media-statistics/, January 2018. (Accessed on 02/15/2019).

[34] Jennifer King, Airi Lampinen, and Alex Smolen. Privacy: Is There an App for That? In *Symposium on Usable Privacy and Security*, page 12. ACM, 2011.

[35] Paul Kline. *An Easy Guide to Factor Analysis*. Routledge, 2014.

[36] Alfred Kobsa, Sameer Patil, and Bertolt Meyer. Privacy in Instant Messaging: An Impression Management Model. *Behaviour & Information Technology*, 31(4):355–370, 2012.

[37] Ponnurangam Kumaraguru and Lorrie Faith Cranor. Privacy Indexes: A Survey of Westin's Studies. *Carnegie Mellon University*, 2005.

[38] Rebecca K Lee. Romantic and Electronic Stalking in a College Context. *Wm. & Mary J. Women & L.*, 4:373, 1997.

[39] Kevin Lewis, Jason Kaufman, and Nicholas Christakis. The Taste for Privacy: An Analysis of College Student Privacy Settings in an Online Social Network. *Journal of Computer-Mediated Communication*, 14(1):79–100, 2008.

[40] Yao Li, Alfred Kobsa, Bart P Knijnenburg, and MH Carolyn Nguyen. Cross-cultural Privacy Prediction. *Proceedings on Privacy Enhancing Technologies*, 2017(2):113–132, 2017.

[41] Naresh K Malhotra, Sung S Kim, and James Agarwal. Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research*, 15(4):336–355, 2004.

[42] Sara Malm. Saudi Arabian Man gets 40 Lashes for Insulting ex-Wife on WhatsApp. https://www.dailymail.co.uk/news/article-6388353/Saudi-Arabian-man-gets-40-lashes-insulting\/-ex-wife-WhatsApp.html, November 2018. (Accessed on 02/15/2019).

[43] Patrick E McKnight and Julius Najab. Mann-Whitney U Test. *The Corsini Encyclopedia of Psychology*, pages 1–1, 2010.

[44] Patricia Brierley Newell. A Cross-cultural Comparison of Privacy Definitions and Functions: A Systems Approach. *Journal of Environmental psychology*, 18(4):357–371, 1998.

[45] Midas Nouwens, Carla F. Griggio, and Wendy E. Mackay. "WhatsApp is for Family; Messenger is for Friends": Communication Places in App Ecosystems. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 727–735, New York, NY, USA, 2017. Association for Computing Machinery.

[46] Times of India. WhatsApp to Cap Message Forwarding to 5 Chats Globally. https://timesofindia.indiatimes.com/business/india-business/whatsapp-to-cap-message-forwarding-to-5-\/chats-globally/articleshow/67623233.cms, January 2019. (Accessed on 01/25/2019).

[47] Department of Justice. Privacy Act of 1974, 1974.

[48] Daniela Seabra Oliveira, Jeremy Epstein, James Kurose, and Anderson Rocha. Cybersecurity and Privacy Issues in Brazil: Back, Now, and Then [Guest Editors' Introduction]. *IEEE Security & Privacy*, 16(6):10–12, 2018.

[49] Kenton P. O'Hara, Michael Massimi, Richard Harper, Simon Rubens, and Jessica Morris. Everyday Dwelling with WhatsApp. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work Social Computing*, CSCW '14, page 1131–1143, New York, NY, USA, 2014. Association for Computing Machinery.

[50] Sameer Patil and Alfred Kobsa. Instant Messaging and Privacy. In *Human-Computer Interaction*, volume 4, pages 85–88, 2004.

[51] Sameer Patil, Alfred Kobsa, Ajita John, and Doree Seligmann. Comparing Privacy Attitudes of Knowledge Workers in the US and India. In *International Conference on Intercultural Collaboration*, pages 141–150. ACM, 2010.

[52] Iryna Pentina, Lixuan Zhang, Hatem Bata, and Ying Chen. Exploring Privacy Paradox in Information-sensitive Mobile App Adoption: A Cross-cultural Comparison. *Computers in Human Behavior*, 65:409–419, 2016.

[53] Sankalp Phartiyal and Aditya Kalra. Indian Political Parties Abuse WhatsApp Service Ahead of Elections. https://www.reuters.com/article/us-india-whatsapp/indian-political-parties-abuse-whatsapp-\/service-ahead-of-election-executive-\/idUSKCN1PV1E3, February 2019. (Accessed on 02/15/2019).

[54] Bandana Purkayastha, Mangala Subramaniam, Manisha Desai, and Sunita Bose. The Study of Gender in India: A Partial Review. *Gender & Society*, 17(4):503–524, 2003.

[55] Prashanth Rajivan and Jean Camp. Influence of Privacy Attitude and Privacy Cue Framing on Android App Choices. In *Authentication Workshop of the 12th Symposium on Usable Privacy and Security*. USENIX Association, 2016.

[56] Prashanth Rajivan, Pablo Moriano, Timothy Kelley, and L Jean Camp. Factors in an End-user Security Expertise Instrument. *Information & Computer Security*, 25(2):190–205, 2017.

[57] Yasmeen Rashidi, Kami Vaniea, and L Jean Camp. Understanding Saudis Privacy Concerns when using WhatsApp. In *Workshop on Usable Security (USEC)*, 2016.

[58] Gustavo Resende, Johnnatan Messias, Márcio Silva, Jussara Almeida, Marisa Vasconcelos, and Fabrício Benevenuto. A System for Monitoring Public Political Groups in WhatsApp. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, WebMedia '18, page 387–390, New York, NY, USA, 2018. Association for Computing Machinery.

[59] Nithya Sambasivan, Garen Checkley, Amna Batool, Nova Ahmed, David Nemer, Laura Sanely Gaytán-Lugo, Tara Matthews, Sunny Consolvo, and Elizabeth Churchill. "Privacy is not for Me, it's for Those Rich Women": Performative Privacy Practices on Mobile Phones by Women in South Asia. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 127–142, 2018.

[60] Balla Satish. How WhatsApp Helped Turn an Indian Village into a Lynch Mob. https://www.bbc.com/news/world-asia-india-44856910, July 2018. (Accessed on 02/15/2019).

[61] Paul Sawers. Why WhatsApp Scrapped Its $1 Annual Subscription Fee. https://venturebeat.com/2016/01/18/whatsapp-subscription/, January 2016. (Accessed on 05/26/2020).

[62] Manish Singh. WhatsApp Hits 200 Million Active Users in India. http://mashable.com/2017/02/24/whatsapp-india-200-million-active-users/#Dka5Ao6c5sqW, February 2017. (Accessed on 05/26/2020).

[63] Statista. WhatsApp: Mobile Usage Penetration by Country. https://www.statista.com/statistics/291540/mobile-internet-user-whatsapp/. (Accessed on 02/23/2020).

[64] Ryan Theodosia. Being (Nonbinary) Transgender in a Census Year, May 2019.

[65] MS Thirumalai and B Mallikarjun. The Official Language Act, 1963 (As Amended, 1967). *Language in India: Strength for Today and Bright Hope for Tomorrow*, 2002.

[66] Sigmund Tobias and James E Carlson. Brief Report: Bartlett's Test of Sphericity and Chance Findings in Factor Analysis. *Multivariate Behavioral Research*, 4(3):375–377, 1969.

[67] Blase Ur and Yang Wang. A Cross-cultural Framework for Protecting User Privacy in Online Social Media. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 755–762. ACM, 2013.

[68] Paul van Schaik, Jurjen Jansen, Joseph Onibokun, Jean Camp, and Petko Kusev. Security and Privacy in Online Social Networking: Risk Perceptions and Precautionary Behaviour. *Computers in Human Behavior*, 78:283–297, 2018.

[69] Yang Wang, Gregory Norice, and Lorrie Faith Cranor. Who is Concerned About What? A Study of American, Chinese and Indian Users Privacy Concerns on Social Network Sites. In *International Conference on Trust and Trustworthy Computing*, pages 146–153. Springer, 2011.

[70] Alan F Westin. Privacy and Freedom. *Washington and Lee Law Review*, 25(1):166, 1968.

[71] K Yoshida. Performing Multiple t-tests on Different Variables Between the Same Two Groups. 2013.

[72] Seounmi Youn and Kimberly Hall. Gender and Online Privacy Among Teens: Risk Perception, Privacy Concerns, and Protection Behaviors. *Cyberpsychology & behavior*, 11(6):763–765, 2008.

[73] Norhayati Zakaria, Jeffrey M Stanton, and Shreya TM Sarkar-Barney. Designing and Implementing Culturally-sensitive IT Applications: The Interaction of Culture Values and Privacy Issues in the Middle East. *Information Technology & People*, 16(1):49–75, 2003.

A Appendix: Survey Instrument (Excerpt) and Table of Variables

Please note that the numbers are for ordering the options, and are not necessarily used in the same order for the analysis

- What is your age?
  (1) Less than 18 years (2) 18-24 years (3) 25-30 years (4) 31-40 years (5) 41-50 years (6) More than 50 years
- Do you have a WhatsApp account?
  (1) Yes (2) No
- Which operating system do you currently use for your primary smartphone?
  (1) Android (2) iOS (3) Windows (4) Blackberry (5) Symbian (6) Other. Specify.
- Are you concerned that since Facebook bought WhatsApp, targeted ads might start appearing in WhatsApp?
  (1) Definitely yes (2) Probably yes (3) Might or might not (4) Probably not (5) Definitely not
- How long have you been using WhatsApp?
  (1) Less than 1 year (2) 1-2 years (3) 2-3 years (4) 3-4 years (5) 4-5 years (6) More than 5 years
- Do you use the latest updated version of WhatsApp?
  (1) Yes (2) No (3) I don't know
- On average, how often do you use WhatsApp?
  (1) More than once a day (2) Daily (3) More than once a week (4) Once a week (5) More than once a month (6) Once a month (7) More than once a year (8) Once a year (9) Never
- Auto-download feature in WhatsApp allow your media (e.g. images, audio, and video) to be downloaded automatically without the need to explicitly do it manually. This feature is automatically activated in WhatsApp which can be altered later by going to the WhatsApp settings. Did you disable the auto-download feature on WhatsApp?
  (1) Yes (2) Maybe (3) No (4) I do not know
- Blocked feature in WhatsApp allows its users to add any person to the blocked list to prevent them from contacting the user. Did you use the Blocked feature in WhatsApp to block any person from contacting with you?
  (1) Yes (2) Maybe (3) No (4) I do know about this feature
- Have you enabled WhatsApp to send you notifications when there is a new message?
  (1) Yes (2) No (3) I do not know
- Have you previously shared your location using WhatsApp?
  (1) Yes (2) No (3) I did not know about the feature
- I frequently use WhatsApp to send/share private or sensitive chats/media:
  (1) Strongly agree (2) Somewhat agree (3) Neither agree nor disagree (4) Somewhat disagree (5) Strongly disagree
- Are you concerned that anyone who has your phone number is able to contact you and see the activity shared publicly using WhatsApp?
  (1) Yes (2) Maybe (3) No (4) I do not care
- When adding me to a group chat, I would like the app to:
  (1) Definitely ask me before adding (2) Ask me before adding only to specific groups (3) Does not really need to ask me before adding (4) I don't care
- WhatsApp has some privacy features. It allows you to show your last seen, profile photo or/and status to everyone (default option), just the people on your contact list, selectively choose some people, or nobody. What is your setting in each of the following: [Everyone (1) My Contacts (2) Nobody (3) I do not know (4)] (1) Last Seen (2) Profile Photo (3) Status (4) Live Location (5) Read Receipts

Research Questions	Independent Variables	Dependent Variables
RQ1, RQ2 (MWW Test)	(1) Origin/ Nationality (2) Gender	(1) Sensitive Data (2) Professional Contact (3) Targeted Ads (4) Group Add Ask (5) Stranger Contact Concern
RQ3 (Exploratory Factor Analysis)	(1) Sensitive Data (2) Professional Contact (3) Targeted Ads (4) Group Add Ask (5) Stranger Contact Concern (6) Platform (7) Frequency (8) Length (9) Age (10) Gender (11) Education	**Saudi** (1) Sensitive Data (2) Usage Platform (3) Usage Frequency (4) Education and Group Permission (5) Age and Targeted Ads (6) Professional Contact **India** (1) Age and Education (2) Usage Platform (3) Gender (4) Information Sensitivity (5) Targeted Ads
RQ3 (Regression Analysis)	**Saudi** (1) Sensitive Data (2) Usage Platform (3) Usage Frequency (4) Education and Group Permission (5) Age and Targeted Ads (6) Professional Contact **India** (1) Age and Education (2) Usage Platform (3) Gender (4) Information Sensitivity (5) Targeted Ads	(1) Blocking (2) Auto Download (3) Location (4) Notification (5) Profile Photo (6) Last Seen (7) Status

Table 8: List of independent and dependent variables for each research questions.

- What is your primary country of citizenship?
- Which gender do you identify with the most? (1) Female (2) Male (3) Other (4) Do not wish to specify
- What is the highest level of education you have completed? (If currently enrolled, highest degree received.) (1) Less than high school (2) High school graduate (3) Diploma (4) Vocational training (5) Bachelors degree program (6) Masters degree program (7) Professional degree (8) Doctorate (9) Other. Specify.

Realizing Choice: Online Safeguards for Couples Adapting to Cognitive Challenges

Nora McDonald, *University of Maryland, Baltimore County (UMBC)* Alison Larsen, *UMBC*
Allison Battisti, *UMBC* Galina Madjaroff, *University of Maryland*. Aaron Massey, *UMBC*
Helena Mentis, *UMBC*

Abstract

This paper investigates qualitatively what happens when couples facing a spectrum of options must arrive at consensual choices together. We conducted an observational study of couples experiencing memory concerns (one or both) while the partners engaged in the process of reviewing and selecting "Safety Setting" options for online activities. Couples' choices tended to be influenced by a *desire to secure shared assets* through mutual surveillance and a *desire to preserve autonomy* by granting freedom in social and personal activities. The availability of choice suits the uneven and unpredictable process of memory loss and couples' acknowledged *uncertainty about its trajectory,* leading them to anticipate changing Safety Settings as one or both of them experience further cognitive decline. Reflecting these three decision drivers, we conclude with implications for a design system that offers flexibility and adaptability in variety of settings, accommodates the uncertainty of memory loss, preserves autonomy, and supports collaborative management of shared assets.

1. Introduction

While having choice may enable people to identify more avenues for securing their privacy and safety online, scholars worry that this could create a false sense of agency [2]. With the flood of Internet of Things (IoT) and mobile devices, our digital records are increasingly being mined in ways that expose us to unprecedented theft [31]. "Notice and choice" models are not only insufficient to our information age [11]; they may also promote a fallacy of individual control [20] while subverting the role of collaborative oversight [5].

This issue is even more pronounced for an aging population subject to cognitive challenges. As dementia slowly becomes a global epidemic, it is estimated that the condition will affect roughly 115 million by the year 2050 [13]. Alzheimer's dementia is a common cause of age-associated memory loss, though not the only one [30, 37, 44]. Although the prevalence of Alzheimer's in the United States may vary quite a bit [37], recent estimates suggest that over five million people in the United States suffer from the disease [14, 41].

In addition to producing serious deficits in quality of life for those who experience it, memory loss takes an enormous emotional and economic toll on over 16 million unpaid caregivers in the United States every year [9]. Those are people to whom much of the responsibility for the support of safe online activities on social networks, email, and banking and shopping sites falls [6, 25, 36]. Yet, informal caregivers may feel they cannot adequately regulate online practices, which in turn may lead them to restrict online activities for cognitively-challenged partners (or cognitively challenged family members) in a way that may do further harm to the individual [29, 36].

Tools are emerging that help people seemingly exert more control over their networked privacy and security settings, exist more ephemerally on their social networks [45], and delete browsing and other location data [4]. But managing one's own privacy-related stress and sense of helplessness is difficult [8], to say nothing of the stress and helplessness those struggling with memory impairment (and those struggling to safeguard them) may experience. Having an intermediary assist in negotiating this space, such as one's spousal partner, may be useful, but it also creates new and complex interpersonal and cooperative challenges. It may also require more data generation and storage and less privacy in order to allow partners to retrace the digital steps of those with memory impairment. Better understanding of how safeguards are negotiated with respect to privacy between partners will allow us to design better technology solutions.

Because we wanted to capture how Negotiation Partners (NPs) manage the sociotechnical challenges of choosing online safety settings in the face of cognitive challenges, we sought out individuals who had concerns about their memory, or the memory status of their loved one, and also had associated concerns about their safety online. We presented NPs with a "Safety Settings" web page that offered a choice of safety-enhancing browser extensions to help the partners manage online activities. Our findings are that NPs generally chose the security options that were less overbearing and created more agency for both of them, but in making these decisions, took account of context in a customized way, depending on a number of factors. First, NPs' choice of Safety Settings is influenced by both their desire to secure

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2020. August 9 -- 11, 2020, Virtual Conference.

shared assets and their individual and collective technology habits and preference (e.g., one has a system for passwords, the other doesn't). Second, their perceived concerns are rooted in a desire to preserve autonomy wherever possible. The importance of choice is heightened by a desire to develop a strategy that serves memory-challenged partners on their own terms, accommodating the degree of memory loss as well as their lifestyle and baseline technology aptitudes and needs. Third, NPs acknowledge the unpredictability of memory loss and the fluidity of their shifting memory-related roles. This sensitivity to the uncertain course shapes their approach to selecting safety settings: they tend to think of these settings as conferring benefits on both partners. Further memory loss motivates their desire for an adaptive system that poses a variety of choices.

Our paper makes the following contributions. First, we establish that key factors driving choice (securing shared assets, autonomy, and fluidity or instability of roles) are favorable to collaborative approaches to cybersafety. Second, we demonstrate how uncertainty in the context of memory loss is well aligned with choice. Third, our work highlights the need for designs that balance autonomy with collaborative protection of shared assets and risk. In the related work section that follows, we consider individual choice and service provider models as well as studies of memory concern and technology to situate our study. We then describe our study design and findings, and we conclude with recommendations for future work and design.

2. Related Work

2.1. Framing of user choice and service provider paradoxes

Our current internet service provider model is built on the idea that, as long as users are given a choice about whether or not they want to provide data, they are sufficiently protected. Even while scholars have long since acknowledged that "notice and choice" is woefully insufficient [10, 42], there is also recognition that, from a regulatory perspective, the model is here to stay [19]. Technology designers have begun incorporating "notice and choice" style privacy into the engineering process [38, 40], often with the encouragement of regulators [12]. For the foreseeable future, users of information systems must be able to make meaningful choices to manage their privacy.

Many companies, however, are choosing to bury privacy-focused options in so-called dark patterns that users are unlikely to be aware of, or able to defend against [18]. Others simply count on framing privacy choices to encourage well-known paradoxical user behavior in making privacy choices [1]. The privacy community may be unanimous in believing that users cannot, nor should not, be required to read privacy policies. The result is that "choice" has become a dubious concept [46]. The "choice" envisioned 50 years ago by regulators is now rather illusory.

This environment does not bode well for older, cognitively impaired individuals who may be dependent on their partners for privacy and security. Some scholars have observed that couples are making privacy and security decisions jointly, rather than, as most systems assume, individually. Recognition of actual information practices must result in reevaluation of the designs and assumptions that led to them. For example, password sharing practices have been studied in some detail [23, 39], leading to meaningful changes in policy [17].

Shared oversight may be a solution to joint management, but it too has problems. Acquisti et al. [2] identify three themes in surveying the privacy literature to address this question, notably taking up the issue of choice as a third constellation in which users have merely the *illusion of choice* due to subtle machinations that enhance or stifle privacy concerns. Those overarching vulnerabilities raise interesting questions. When a partner is making privacy decisions on behalf of a dyad, are they any better-equipped to see realistic options? Is it realistic to think of them as a bulwark that further fortifies the other user (or both of them) against manipulations?

There is an inherent paradox introduced by security. Measures taken to enhance it sometimes, though not always, have the effect of limiting privacy. Typically, a service provider who affords the platform, tools, and policies, organizes and monitors our data to prevent its misuse. They provide, for example, protection against scams or data theft. But the notion of couples exercising surveillance over one another is unique, even if there are corollaries, because these attempts to exercise protection fall outside normative models of privacy. Analogs like the monitoring of children and survivors of intimate partner surveillance/violence (IPS/V) usefully apply and are discussed in the next section. The caregiver who is theoretically tasked with taking on some of the burden of oversight of privacy and security must also oversee the couple's mutual security, which is to say, they are functioning in some ways like the service provider.

2.2. Finding a balance: safety versus surveillance

Theories like boundary regulation [3], the privacy paradox [24], and contextual integrity [32] have usefully described how individuals and communities manage privacy. Yet, what of couples trying to manage their privacy and security collaboratively? While privacy is an important value (though arguably not fully addressed by these theories), it is complicated by the desire to provide autonomy to individuals who are losing their memory. Boundary regulation [3, 35] assumes that the individual has the power and privilege to regulate access to the self, which individuals suffering from memory loss may not. Moreover, it assumes that regulating access to the self is the path to autonomy, when in fact, giving up some unedited activity (or sharing it with a partner or caregiver) may be what is required to gain autonomy.

Sometimes guardrails that introduce some type of oversight or surveillance may be the only approach that is both safe and empowering.

These checks and balances exist for other unique vulnerable groups, like children experiencing parental monitoring, who may find it restrictive and harmful [16]. There are also concerns, in this realm, that networked surveillance tools, some of which are used in parental monitoring [33], are also being adopted by those who perpetrate IPV and IPS [15, 43]. Parental monitoring and IPV/S are not the same as safety for elderly NPs with memory concerns, but arguably developers should be responsible for the intended uses *and actual uses* of technologies [33]. Our research attempts to understand the extent to which NPs are bound by these systems, how decisions are being made and by whom, what terms are acceptable, and whether or not NPs have concerns about surveillance. Unlike parental monitoring or IPS/IPV, the need to regulate a partner's activities is theoretically premised on a need to safeguard shared data and data linkage while maintaining autonomy. But like child monitoring, and maybe IPS/IPV, these monitoring activities nevertheless invite non-normative models for thinking through the problem, precisely because of dual use risks and vulnerability.

Couples where one or both has undiagnosed memory loss or MCI exemplify a unique cybersecurity and privacy problem. They may be even a discursive element in some areas of privacy studies because, for the sake of their cybersecurity, certain aspects of their privacy may be compromised in the service of their individual autonomy and mutual security. As with children and IPV and IPS survivors, what makes these mechanisms for security and cybersecurity valuable (e.g., find me or my phone) or convenient (e.g., storage backup) may, in fact, be what makes them prone to abuse by family/parents or intimate partners. It is this paradox that we enter, as researchers, looking to give agency and security to older adult NPs—both those experiencing memory loss and those charged with overseeing them—without inviting other harms. This problem grows ever more urgent with cybersecurity threats on the rise [21].

2.3. Sociotechnical issues and support for memory loss

This section highlights what we know from the literature about memory concern and technology. Individuals with (un)diagnosed memory loss may be uniquely susceptible to threats because of diminished cognitive abilities that leave them less likely to detect scams and less able to regulate their financial or social activities [26]. Ultimately, oversight may be left to partners who must protect the other from phishing scams and unwise or duplicative purchases [29, 36]. This can create burden and worry for family members and caregivers who feel they are responsible for maintaining the agency of vulnerable partners with agency and their own piece of mind.

While technology could offer a means for NPs to potentially extend support to their partners with memory impairment while living at home, Mahoney et al. pointed out important ethical issues that arise from home monitoring [27]. Their work emphasizes the need for researchers to focus on respect and autonomy for the individual with memory loss, as well as quality of life, but also respect for family caregivers and relationships with caregivers. In so doing, Mahoney et al. usefully highlight the way in which end-users also include family and patient collaborators [21].

Mentis et al. found that in addition to not being able to discern scams and misleading emails, MCI can contribute to embarrassing episodes that cause tension in the family [29]. The solutions that couples formulate in response to online threats vary widely and are not apparently connected to cognitive decline, suggesting that perhaps solutions have more to do with the relationship dynamics than memory loss, or suspected memory loss itself. Couples described a wide range of strategies for managing MCI technology use: never leaving the others side (or "hovering"), to limiting access, to checking their activities once they were done (thus providing autonomy with a "checker"), to taking over for the individual with memory loss when they were no longer able to interact with the system. These strategies reflect not only the cognitive abilities of individuals with memory loss, but also partner dynamics, co-location, caregiver comfort, technological savvy, and tolerance for risk. They reflect a cooperative and constitutive approach that is *sui generis*—too specialized in its character to sit comfortably within a normative frame.

Mentis et al. found that couples are sometimes planning ahead to a point when the individual with MCI is unable to carry on as they once were online, but don't engage in concrete discussions around cybersecurity and access. Although couples express the desire to make shared decisions, in practice, things may happen differently. Mentis et al. report that "shared-decision making was not feasible as there was a lack of suitable options along a spectrum of care" from which the couple could choose, necessitating additional options—not merely "an illusion of choice" [28]. As a consequence, couples were often caught off-guard and prone to taking extreme, disempowering measures when it became clear that one of the partners could no longer manage online [29]. Depending on the couple, however, measures varied substantially from an emphasis on autonomy (wait and see) to an emphasis on safety (abstinence) [28]. The authors describe two ends of the spectrum of safeguarding approaches—on one end, complete oversight and on the other, no intervention at all [29]. Empowerment of those with cognitive challenges means that couples have to find a middle ground that leaves them latitude to plan ahead and gradually transfer knowledge, access, and responsibility. Although flexibility inherently requires work, couples nonetheless do

not want to be limited to a binary option of preserving or removing *complete* access and control. They prefer a degree of nuance that aligns with their relationship dynamics, their experience, and proficiency.

3. Methods

3.1. 'Choice" technology probe for online Safety Settings

We designed a technology probe that embedded hypotheses about choice that were built on prior findings, with the goal of empowering NPs with choice. The probe is realized as a Safety Settings web page, one that provides a spectrum of safeguard features for various online situations that NPs can discuss and choose together. Our ultimate goal is to develop browser plugins that map to these Safety Setting choices to engage NPs in their online activities in a way that ideally safeguards them on terms that are manageable, and which they completely understand. We would not interfere with the collection of data by these services, though.

Informed by the issues already identified by [29, 36] the situations that were presented were for email, Facebook, online banking or money transfer, online shopping, password management, and online browsing. NPs could select settings from any or all of these categories. For each situation, NPs were presented with the option to select a safety setting for two to three actions specified in Table 1.

Table 1. Online Context and Actions one could Perform that Entail Safety/Security Risk

Application/ Situation Category	Online Actions that Entail Risk
Email	• Clicking on a link in an email message. • Opening attachments in an email message.
Facebook	• Liking a Facebook post. • Commenting on a Facebook post. • Accepting/rejecting a Facebook friend request.
Online Banking	• Viewing bank/financial account online. • Transferring money online.
Online Shopping	• Visiting a site to purchase a product. • Purchasing a product online.
Password (PW) Management	• Setting/changing password to an online site. • Setting/changing password to computer.
Online Browsing	• Searching for information on the internet. • Clicking on a link to download a file off the internet.

For each action, there was a spectrum of choices provided for the Safety System to enact (see Table 2). This spectrum ranged from no interference—i.e., the Safety System would take no action when the person with cognitive challenges performed the action—to what we deem "full interference"—i.e., the Safety System ensures that the action cannot be completed. What is important to note is that between these two ends of the spectrum were three to four additional "levels" to choose from. As the choices moved from no interference to full interference, the choices generally added more security, and in turn one's privacy and autonomy was diminished.

Table 2. Safety Setting Option Levels and Coding for Analysis

Safety Setting Option Level	How Coded in Analysis
Not interfere	1
Record for partner to see later	2
Notify partner	3
Partner review before continuing	4
Review prior to posting (FB only)	5
Deactivate/not allow	6
Couple did not select the situation or unknown	N/A

3.2. Participants

People with some cognitive challenges may ultimately be diagnosed with MCI, but MCI is difficult to diagnosis formally, in part because it is easily confused with natural signs of aging. Our long-term interests are predicated on this uncertainty about cognitive challenges in an older population and thus, we approached recruitment as if cognitive challenges were a concern, not a diagnosis. Specifically, we used self-reported memory concern as a proxy for MCI in determining eligibility. Because deterioration in memory is a pervasive age-related experience and is not necessarily accompanied by a confirmed diagnosis of any kind, our goal, for this research, was to study people in partnerships who *perceive* memory loss, or have *concerns* about memory performance, rather than to study people with formally diagnosed memory loss. It ultimately became apparent in the interviews that both partners in all dyads had some memory insecurity or anxiety associated with aging or disease

We recruited a total of 14 individuals (seven NP dyads) to participate in this study. For a dyad to qualify, both partners had to be 65 and older; one or both had to have memory-related concerns (it was not relevant for us to document which person in the screening process); and one or both had to have security concerns online. While one couple was interviewed in-person in their home, the remaining six couples were interviewed using an online meeting tool, GoToMeeting.

The participants were recruited from a marketing panel and, in one case, a continuing care facility with which we have an established relationship. One individual in each dyad was the screened respondent who spoke on behalf of the pair. Assisted living centers were difficult populations from which to recruit couples. We thus turned to a panel to find older adult couples living at home with memory loss. We did not

gather demographic information, as is our practice whenever conducting qualitative research on sensitive topics. Our priority is to ensure that not only will participants remain unidentifiable, that they will have trouble identifying themselves.

In some cases, respondents did not want their memory issues emphasized or even discussed with the other partner while in the interview, so we have taken steps to further conceal identity.

3.3. Study design

The consent form was sent to participants ahead of time for each of them to read and sign. The sessions started with an introduction to the study. NPs were then asked to walk through the system. We used remote access given through GoToMeeting to allow participants to make selections themselves, on the Safety Settings web page, but this only worked once. In all other cases, participants told the investigator what options to select, and the investigator did not speak in order to allow for naturalistic observation.

As NPs made their selection, we asked them to share aloud their thought processes and speak freely with their NP as they decided what settings were most appropriate given their current situation. Given that in previous work, we know that some of the NPs meticulously plan and discuss privacy related issues and settings, this approach was natural.

We followed up each walk-through of the settings mock-up with a brief, semi-structured interview designed to probe usefulness of settings, how NPs might elevate privacy concerns, and how Settings might evolve with the disease progression.

Table 3. Safety Settings Options Chosen for each Situation

	All	Email	FB	Bank	Shop	PW	Browse
Not interfere	31%	21%	57%	29%	29%	7%	29%
Record for partner see later	35%	43%	24%	36%	43%	36%	36%
Notify partner	13%	0%	0%	36%	14%	29%	7%
Partner review	8%	29%	0%	0%	0%	0%	21%
Review prior to posting (FB)	1%	-	5%	-	-	-	-
Deactivate/ not allow	2%	7%	0%	0%	0%	0%	7%
Couple did not select	10%	0%	14%	0%	14%	29%	0%

3.4. Data collection

The walk-through selection of the safety mechanisms and post-interviews were conducted in one couple's home and, for the remainder of our participants, using online GoToMeeting at their convenience. These set ups were audio/video recorded and became part of our study, providing thick descriptions of the sociotechnical dynamic within the NPs. The interviews were completed with one couple at a time. Overall sessions lasted anywhere from one to three hours, depending on how much socializing, technology setup, and logistics (e.g., a NP that was not yet home) were involved. The sessions themselves where NPs engaged with the technology probe only lasted for roughly 20 minutes. These observations were captured on video, audio, and screen captures.

Table 4. Prevalence of Safety Settings Options Chosen for each NP (when NPs made a selection)

	Not interfere	Record for partner	Notify partner	Partner review	Review prior to posting (FB)	Deact./ not allow
NP0	0%	17%	33%	33%	0%	17%
NP1	77%	8%	0%	15%	0%	0%
NP2	0%	75%	25%	0%	0%	0%
NP3	8%	92%	0%	0%	0%	0%
NP4	38%	62%	0%	0%	0%	0%
NP5	23%	0%	54%	8%	8%	8%
NP6	82%	0%	0%	18%	0%	0%

3.5. Data analysis part I

The analysis of the resulting audio/video and screen captures was conducted in two stages. The first stage of analysis was to describe the types of options that were chosen when couples were presented with a full spectrum of choice. We did this in two ways: first, we counted the occurrence of a selection for each option in each context; second, we counted the occurrence of each selection for each option for each NP.

Couples had a total of 13 choices to make, as detailed in Table 1—two per setting (email, banking, shopping, password management, and browsing) with the exception of Facebook, which had three—for a total of 91 potential choices, including no choice at all or N/A. We report on choices NPs made as an aggregate number or percentage of these choices across total situations (out of 91 options) and within situations (out of 14 options, or 21 options for Facebook) (Table 3) and by couples (out of 6-13 choices per couple—

depending on how many settings options they provided an answer for) across situations (Table 4). All of our couples used Facebook, but some do note differences in use among them; and one couple, NP0, did not choose to select Safety Settings for Facebook. NP0 also did not choose to select Safety Setting for online shopping or password management, resulting in them only making 6 choices. NP6 did not choose to have Safety Settings for password management, resulting in only 11 choices. This resulted in 82 total choices being made by couples if N/As are not included. These results merely represent a description of our participants' choices and are not meant to invite statistical inference.

3.6. Data analysis part II

The second stage of analysis was to explain *why* the participants made the choices they made during the study. To answer this question, we used a qualitative approach and first transcribed the discussion the NPs had during the study as well as the post-study interviews. We used a thematic approach to analyze the transcribed data [7]. This approach provided us with the ability to move beyond surface level similarities to salient themes. The analysis focused on the way in which NPs interpreted and made choices around cybersecurity Safety Settings and the way in which sociotechnical roles, concern for autonomy (and its fluidity), and context and experience shaped choice. Our findings organize these themes around the concept of choice, specifically as it relates to protecting shared assets (more surveillance), social activities (less surveillance), fluidity of roles which might mean that they need mutual oversite of sensitive areas or just the option to adjust settings if things change. The analysis was primarily conducted by the first author, who wrote memos from audio/video recordings of the sessions and sorted these findings and transcripts into themes. These were continually presented to the other authors for review and discussion.

Our presence in this process from recruitment through interviewing takes the form of both silent observer and disrupter, but neither role can be deemed unobtrusive, as we will show. We did not seek out generalizability so much as an encounter with choices and how NPs understand and negotiate them, taking into account highly idiosyncratic and personal/private matters related to memory loss and broader uncertainty around aging.

4. Findings

4.1. Choices made

NPs most frequently choose Safety Settings where they could record their activities for their NP to see later, 32 out of 91 choices, followed by no interference, 28 out of 91 choices (see Table 3). Variations in their choice of Safety Setting stringency often reflect the context, whether they had more or less concerns about safety or were (sometimes regardless of memory loss) worried about being the target for scams in ways that required mutual oversite.

Email elicits the most stringent settings (5 out 14 selections required partner review or deactivate/not allow) followed by banking (5 out of 14 choices were notify partner) password management (4 out of 14 choices were notify partner), browsing (5 out of 14 with 1 choosing notify partner, 3 choosing partner review, and 1 choosing to deactivate altogether). The other categories (shopping and Facebook) tend to elicit less strict Safety Setting selections, erring (slightly) more often on the side of no interference or record for later. These were situations that were more social (more individual), and thus, more interlaced with autonomy, whereas the other settings tended to involve more shared assets and thus, shared safety.

Table 3 shows the safety selection by feature category within each situation. There are only a few exceptions where NPs make selections that were more stringent for a certain feature: opening email attachments (as opposed to opening links), commenting on Facebook posts or accepting/rejecting friend requests (as opposed to liking a post), transferring money (as opposed to viewing one's accounts), clicking on links while browsing (as opposed to searching). Password management Safety Settings choices are the least stringent when it came to changing passwords on the computer (as opposed to on sites).

Most often NPs choose to record activities for them self or the other partner to see later, but many still choose the setting "no interference." When NPs make these selections, it is frequently described as providing a log for *both* them and their partner.

4.2 Seeing from the other's perspective and memory loss uncertainty

Couples' choices tend to reflect awareness that memory loss could affect either partner, and this awareness inspires a dual perspective on Safety Settings, also enabling them to take account of the needs and the styles of each, and the challenges of co-managing the ramifications of memory loss.

NP2-2 imagines the breadth of those ramifications:

> NP2-2: It could be your own memory, you know you rely on one another and I think apps or things that could help you with managing things like passwords and certainly money we are handling fine but I can see a time where somebody might click on the wrong thing very innocently or maybe not so innocently and it could cause a big problem … and it could go fast maybe me more than [partner] because every indication I could have a problem. I do so I don't mind, you know it's less onerous for him as long as we just agree we're going to have some checks and balances.

Couples appreciate that they have the flexibility to choose settings as the memory of *either* one declines. They emphasize that collaborative oversight gives them the

opportunity to mutually manage risk but also that, for now, autonomy is essential ("the ball is in her court").

> NP1-1: If we were feeling that one had more memory loss than what we are initial thinking, which she will get, and I will too … She still can remember … We still have to have the safety because we are among each other to do it.
> NP1-2: I forgot today.
> NP1-1: But she will forget … The ball is in her court until she starts to really forget … Right now, we are still in early stages. But I think those questions are good because they hit all bases. In later years, in later times, it may occur.

Couples also express uncertainty about progression, and speculate that settings might need to change in a year, or five, or ten years:

> NP3-2: Yes. it could be 5 years it could be 10 years and we…
> NP3-1: It could be tomorrow we don't know…

> NP4-1: Yeah, because like say, I think that our circumstances are different right now. I think we would answer them differently maybe in a year so.

This way of thinking of memory loss as possibly affecting both of them and having an uncertain timeline supports, as we will show, a collaborative way of thinking about their cybersecurity as a shared challenge.

In the sections that follow, we present the qualitative analysis from our observations and interviews, describing how our participants viewed each level of choice.

4.3. Not interfere

The choice to "not interfere" was the second most frequently made (28 out of 91 choices) after 'record for partner to see later' (32 out of 91 choices). This was particularly the case in the context of social activities (like Facebook 12 out of 21 choices) and when the partner perceived their choice would encroach on the other's autonomy—e.g., with regards to shopping or browsing as well as banking (4 out of 14 choices in all cases).

In the context of social interaction, a common response to this setting is that only one of the NPs actually uses Facebook, even if they both have it. Those that do choose to have Safety Settings for Facebook, largely choose "no interference," agreeing that social activities is the other's private business.

Facebook was the only application where NPs were given the additional choice to have one partner review the activities prior to posting. This was not a top choice because couples do not want to infringe on the others social autonomy:

> NP1-1: Like what you like.
> NP1-2: Do not interfere.

This couple communicates a sentiment widely shared: that these settings were meant to provide security but not "clip their wings." There they draw the line.

Browsing is not an application NPs necessarily want any Safety Settings for, though they indicate it as an application or situation for which they want to set Safety Settings, perhaps because they assume it is something that they always do and thus seems obvious or necessary to discuss.

Couples want to ensure autonomy where there no shared assets. For example, for NP1, having accounts they did not share makes them comfortable with no interference:

> NP1-1. Separate accounts. Not interfere.

This couple, like others, is sensitive about preserving autonomy where it already exists. Along those lines, one couple considers only that "review" is necessary and otherwise "not interfere."

> NP3-1 The only way I would want this to work is if I need her to review it, otherwise "not interfere." Does that make sense? I don't know if that's an option.
> Interviewer: Did any of the options look to you as if the partner would have the opportunity to review it in a way that you would be happy with? For instance [reads options …].
> NP3-1: What does "deactivate all links" mean. [inaudible] Thank you, I want to know what they're looking for?
> Interviewer: This would make it so a person could not click on links in email.
> NP3-1: Okay. I don't think I'd want that. I guess, "Not interfere."

NPs frequently toggle between "not interfere" and "review for later," but often side with the choice that gave more autonomy if possible.

4.4. Record for partner to see later

Most often NPs choose to have their activities "recorded for their partner to see later" (32 out of 91 choices), and this applied across situations. NP6 communicates that they are not worried about memory loss, but rather malicious links that could be inadvertently selected for reasons having to do with the other's technical knowledge and past experience with scams in email and Facebook. Recall that NP3 wanted only to "review activities" when it made sense in context. Later, they refine their selection saying that what they wanted is to "record their activities for their partner to see later" (for both links and attachments) with the expectation that it will serve *their* memory (not necessarily for their partner's oversight, though some NPs suggest that is the ultimate expectation). For instance, in the following example, NP3 moves onto the next option setting and in the course of making this selection goes back to the prior selection to change it from "not interfere" to "record":

> NP3-1: Okay, okay. Now I understand the concept. If that is the case, then I would want to have a log of everything that I did so go back to the previous one [previous option setting question]. Okay. Yeah, I'd like to "record all the links I click on for your partner to see later." I'd say that one. In other words, I would have a log to refresh my memory because that's what I need.

This same couple selects the setting to record what their partner posted on Facebook for the other to see, later saying that they did not put personal things on Facebook but wanted it just in case.

> NP3-1: Well she does Facebook. I don't do Facebook. So.
> NP3-2: I guess record all Facebook comments for your partner to see later, is the only thing. I don't use it where I put anything personal on it, but just in case.

For this couple, concern that either of them could lose their memory counts as a reason to have some record for either or both of them—a decision that serves their sociotechnical habits well because each partner has different methods of organization:

> NP3-2: Well, should my husband and I lose our memory more, I think he understands most of my things, but I find whatever he does extremely complicated. We are not organized in the same way and to me he's all over the place. So, I would want a fixed place to know what he's on, what he needs to know, or what I need to know. It needs to be straightforward, not 14 different paths to get there.

Online banking and shopping are frequently recorded for one's partner to review later because, for most, it is not of great concern. NP2 approaches it with a mixture of humor and seriousness, allowing for the possibility to need a stricter setting later:

> NP2-2: I would say "record all the places that your credit card is going." Number 2. Otherwise or "immediately notify of the sites where you can enter your credit card number." But that's gonna slow you down.
> NP2-1: Yeah because you're gonna wanna buy stuff without having to talk.
> NP2-2: Well, I'm just concerned you're gonna reach a point where you are spending on what you want necessarily not that you need it. Okay?
> NP2-1: This means I can't surprise you with any presents.
> NP2-2: Oh boy I need to rethink that.
> NP2-1: Which one?
> NP2-2: How about "immediately notify your partner of purchase amount or record all purchase amounts for your partner to see later."
> NP2-1: Which one?
> NP2-2: You can do two if you want, maybe…
> NP2-1: Why don't you just do 1?
> NP2-2: You can do choice number 2 ….
> NP2-1: Okay.

At one point, NP2-2 again expresses concerns that their partner could buy things they wanted but did not need. Pointing out that their partner has bought a car once online, NP2 still selects "record" and not the more stringent option:

> NP2-2: I worry about… my worry is you might go buy a car at some point, and yes, he has bought a car on the Internet just once.

This couple's worries were clearly linked to memory loss and an impending sense of changing roles. Perhaps, as a result, they participated in a lot of back-and-forth in which they debated the option that they thought was most fitting, reluctant to give up autonomy. In these cases, options provided along a spectrum allowed for a negotiation space and outright discussion of what some of the potential incidents might be on the horizon.

4.5. Notify partner

The choice to keep NPs aware of what the other was doing in real time is much less often selected (12 out of 91 choices) but tends to come up where there were concerns about "shared assets," which included both banking and passwords. The function of these notifications was to be aware of activity for security and potential intervention (in the case of banking) and to stay abreast of changes, as well as for their own recall (in the case of passwords). In that sense, notification choice served different purposes, one being more about security from cognitive challenges and malicious activity, and the other more about memory management, respectively.

Several NPs choose to have their partner notified of online banking activity. Notably, NP3 remarks that these settings might become more stringent, in one case, citing worrisome incidences with other members of the family who have also experienced memory loss:

> NP3-1: Yeah. Banking account. Same thing: keep a log of what I am doing to help me remember for later. I guess it depends on the extensiveness of the mental disease that you are having as far as memory. If you want to be notified immediately or later. Right now, I would need it later.
> NP3-2: But this would be something that could be put in place if things changed. His mother suffered from a lot of memory issues and she denied she had a lot of it. And I would want this.

Again, we see a lot of discussion addressing concern about cognitive challenges that could change dynamics. This sentiment was illustrated by NP3 and also echoed by NP2; the only difference is the stringency of settings they finally settled on.

Password management also frequently prompts selection of "notifying" one's partner, mostly because NPs relate that they often forget their passwords and are constantly resetting them; and they consider passwords to be a shared asset. Not only did they want their partner to be able to see what they chose as their new password, they also want to be reminded of the password themselves. NP2-2 wants to notify their partner of a change because they feel that they would want that for them self, even if they did not have memory issues:

> NP2-2: Do you think you need notification when they change because you change your passwords frequently. Well you do when you can't remember you changed…
> NP2-1: Well this is for you to know
> NP2-2: OK
> NP2-1: Not for me to know
> NP2-2: Right, or that I can help you with passwords…

NP2-1: Focus on your own
[Both Laughing]
NP2-1: I might want notifications on mine
NP2-2: Do you wanna go back [to choosing option notify partner from record for partner to see later]
NP2-1: Yeah. We can click the third one

This exchange between NP1 and NP2 shows how this tool was not only about safeguards for memory loss but everyday memory issues associated with life online—and possibly, though not certainly, aging. It highlights how assets are a kind of shared concern that can overlap with autonomy.

Couples feel that choice of notifying partner was particularly important around issues of shared assets. They wanted to be able to enforce a kind of mild surveillance to ensure security—not just from forgetfulness but also bad actors, who, according to one couple, were sometimes in their own family. Another couple points out that with anything related to money or passwords, they are more leery:

NP3-1: If it's, if it's related to money I would say yes, depending upon, you know, how bad we are … We keep passwords the same thing. It's like giving the key to your house. you know somebody gets a gift card to a website they can do whatever they want.

To demonstrate the importance of the choice to notify partner as a way of keeping an eye on shared assets and providing protection from outside actors, one couple would like notification to alert them when their grandson is browsing and downloading a file:

NP5-1: My grandson is five. He gets on the computer. I would want [Safety Settings] to immediately notify you are downloading a file ... He uses it for school, to do homework stuff.

The idea of protecting yourself against family is, for NP5, made more salient by things they had heard from friends, as well as experiences they had had with family.

In general, we found that the choice to notify partner is critical in the context of shared assets, where the risks make surveillance much more acceptable.

4.6. Partner review

On occasion, NPs choose to have their partner review at the moment of action and approve or deny (7 out of 91). Desire for intervention is most pronounced with email, where there is a sense of being targeted and an accompanying concern about clicking on malicious links. As NP6-1 describes, she is worried, even now, about that vulnerability, and her partner's ability to assess what is malign, independent of memory issues.

NP6-1: I trust him but I don't trust other people on the computer and the different things they may do. If I send him something through an email … and I'll put some kind of little note where he'll know it's something. Where if I didn't, I'm wondering, "would he just click it?" So, I don't know all the scenarios, so that's where I would say that.

This couple mentions that they have talked about not clicking on links sent by their family members out of fear that they are malicious and have developed a practice of mutually alerting so that the person with more computer knowledge can assess the link. The ability to imagine that things could get worse for either member of the pair leads couples to appreciate the option of having a partner review. Still, some expressed concerns about the potential for such an option to become invasive and also burdensome.

4.7. Deactivate

Deactivation or allowance of activity was presented as disabling those links not on a preapproved safe list or disabling the activity altogether, depending on the setting. Only two couples chose to have settings deactivated, for email attachments and for browsing, out of concern that by the time their partner clicks on it, it will be too late. This was simply not a popular choice and not even one that couples discussed using as they imagined more stringent settings down the road.

NP5-1 chooses "deactivate links" in search on a list that they were able to curate because they feel that this safety measure protects *both of them* from malicious attack, not because of memory loss.

NP5-1: I would only put like places that … I normally browse.

NP5 was concerned about the need to adjust settings to accommodate memory loss but, these decisions tended to be between "no interference" and "review" or "notification," and not "deactivation."

The choice of deactivation is a last resort, one that couples consider only where they fear they may become helpless—not necessarily as a result of memory loss but rather, due to the activities of bad actors. We posit that, given all the choices couples do have, the prospect of deactivation seems remote.

5. Discussion

NPs like the option of being able to notify their partners, particularly in the realm of shared assets. Because they imagined themselves potentially in the same role, and because the course of memory loss is recognized as so uncertain, the concept of shared privacy has some appeal.

This runs in contrast to the idea that couples are managing their privacy settings individually as most systems assume. The choices couples make reflect joint ownership of the problem as well as respect for autonomy by (paradoxically) embracing uncertainty, a "see-as-we-go" attitude expressed by all of our couples.

5.1. Choice reinforces autonomy

Having flexibility of choice fits NPs well in that it allows them to begin with a light touch and then introduce more safety as they sense decline. Even measures like keeping a

record are quickly recognized as offering mutual benefit—the idea of providing a history for their own convenience and later, an oversight resource that is available to their partner. NPs often think about an indeterminate, future time when either of them might handle their finances in an unsafe way. This led them to want the ability to be aware of what the other was doing, and also provide access at a later date (maybe five to ten years, they could not be sure) to a family member who might need to supervise both of them.

Overall, preserving autonomy was paramount. NPs consider what these choices might mean for the person with oversight, as well as for the person in need of oversight. Their ability to pivot in these ways, between present and future, self and partner, reflected their grasp of mutual vulnerabilities.

We plan to test the adaptive nature of this design by allowing couples to adopt these Safety Settings over an extended study. Our future study design will prompt couples to reflect on Safety Setting changes to capture whether they are motivated to adjust their settings over time in response to experiences of risk [22] or memory decline [34].

Future designs will iterate on ways to make the option to adjust Safety Settings apparent. We will be interested to see if those design changes influence Safety Setting choices, and how the pace and rate of adjustment relates to breaches in security, changes in cognitive status, and even to changes in relationship dynamic as couples adapt to progression. Because participants themselves could not reasonably project the future, or even imagine themselves capable of doing so, the triggers for adjustment remain unclear.

5.2. Choice supports *social autonomy*

Choice allowed couples to extend continuing autonomy to their partners in social realms, where they deem latitude important. Facebook tended to be designated for nonintervention based on what couples explained was a desire to extend freedom to socialize. These decisions could be the result of couples' failure to fully appreciate all the ways in which Facebook invites risks. Note that we did not provide an extensive list of Facebook activities which might be considered more risky (e.g., posting or clicking on a link). Future design iterations should include a more concrete explication of these activities and risks.

5.3. Choice supports *shared assets* and sociotechnical idiosyncrasies

We found the logs and more overt forms of review and notification surveillance provided a way to personally retrace steps or intervene around shared assets. For both NPs, these more stringent settings provided insight into what was done that solved current struggles with maintaining shared assets. In other words, they served the current dynamic and provided a buffer for all parties.

5.4. Choice that embraces *uncertainty* supports autonomy for partner and self ("It could be me")

Simply by introducing choices, the couples were able to customize each safety setting in a way that preserves more autonomy for both the partner with greater memory concern and the one with less. Those roles were acknowledged to be *uncertain* at the beginning. Thus, the "record" option, in particular, was seen as allowing a person with memory loss to access their own logs (enacting a sort of personal surveillance) and also permitting the person with less memory concern to eventually review them. The fact that these roles could potentially be reversed in the event that health circumstances change (e.g., if one suddenly declined faster) made them more sensitive to the need for a system that was adaptive and, and sensitive to each other's feelings and requirements. For this reason, the wording of the system could be oriented towards more cooperative oversight, rather than for later review by just one partner.

We contend that this embrace of *uncertainty* shapes choice and broader, long-term thinking about the utility and place of this system. Even NPs who had identified one partner as suffering from more decline acknowledged that they could suddenly be the ones to require more assistance. We interpreted this admission as both acknowledgment of the fragility and uncertainty and unpredictability of memory loss, and also maybe a feeling that the individual at greater risk might skirt the supposed prognosis. Nothing is certain, which is why collaborative and adaptive approaches seem all the more appropriate.

Because NPs are open to the idea that memory loss is part of aging (even if it may overtake one of them more quickly, or dramatically) they are quick to offer that they would like to include a family member (or even a caregiver) in this system. Although the potential for caregivers or even family to take advantage of this access does come up, it is not a major concern. At the same time, as NP5 pointed out, family can be the source of security threats.

Future design iterations will explore ways to foster self-surveillance and make record-keeping less obtrusive and burdensome to the partner. These choices might still include latent monitoring and alerts that allow the other partner to retain oversight over those records. Because partners are open to the possibility that their roles might change, we will need to carefully consider how we articulate or impose them.

5.5. Choice means more risk

Even if they worry about shared assets, NPs are prone to accept more risk out of respect for partner autonomy and in deference to changing circumstances and roles. This tendency also coincided with a desire not to disclose memory concern or to accept that "it could be me." Because we did not seek out couples with a diagnosis, we had to be

comfortable with ambiguity in our approach. Future design iterations must be attuned to this ambiguity; the sensitive nature of disclosure; and the evolving nature of cognitive decline in relationship to risk.

When designing future iterations, we will look to collaborative service provider models for inspiration and to help frame, in particular, our understanding of shared risk in relationship to autonomy.

6. Conclusions

NPs facing memory loss with cybersecurity concerns think things through as a unit facing very certain health-related ambiguity. They confront the opacity of their situation as a team (one said "as a game"); while they have collaborated in life and in partnership, they are entering a new phase of sociotechnical collaboration around the others or one another's memory decline. We have looked at how NPs work through these issues, finding that relationship dynamics, technological habits, idiosyncrasies, and shared concern, or ability to imagine their own memory decline shapes decisions around cybersecurity Safety Settings. Our findings suggest that NPs need a dynamic system that adapts to their memory concerns (or progression) and anticipates fluidity of roles and the realization that they are not only collaborating in shared preservation of their safety but in a dynamic system that could change. The key component of negotiation was empathy—belief that they are a unit with shared stake and that the roles could be reversed at any time.

NPs are worried about cybersecurity independent of memory issues, like links in email, identity theft and impersonation on social media, social engineering in email, and family members without impulse control. It can therefore be difficult to parse concerns related to memory loss from those inspired by their own experience of risk and threat or media and advocacy group exposure (e.g., AARP).

6.1 Limitations

Our experimental design is limited in several notable ways. First, although we engaged in naturalistic observation, we nevertheless required that couples engage in negotiations out loud with us. Future research will involve diary studies over a longer period to allow participants to negotiate and adjust settings in their natural environment, at their own pace, and as circumstances change. Second, the scenarios we provide, particularly for Facebook, were limited. There are other activities on Facebook that one could engage in that may, in fact, be riskier. Third, despite intensive recruiting efforts, our study involved a limited sample drawn from an online panel, and thus technological adept enough to participate in online surveys, although the technological bar for online panel participation is relatively low. Finally, our study design looked exclusively at couples, and while these findings lend support to the view that cybersecurity is a joint (rather than individual) burden, we will need to conduct complementary research that engages partners as individuals, outside a dyadic context, for a different sightline.

ACKNOWLEDGMENTS

The work is supported by the National Science Foundation grant CNS-1714514.

REFERENCES

[1] Acquisti, A. 2009. Nudging Privacy: The Behavioral Economics of Personal Information. *IEEE Security Privacy*. 7, 6 (Nov. 2009), 82–85.

[2] Acquisti, A., Brandimarte, L. and Loewenstein, G. 2015. Privacy and human behavior in the age of information. *Science*. 347, (Jan. 2015), 509–514.

[3] Altman, I. 1975. *The Environment and Social Behavior: Privacy, Personal Space, Territory, Crowding*. Brooks/Cole.

[4] Apple and Google's tough new location privacy controls are working: 2020. *https://www.fastcompany.com/90454921/apple-and-googles-tough-new-location-privacy-controls-are-working*. Accessed: 2020-02-13.

[5] Baruh, L. and Popescu, M. 2015. Big data analytics and the limits of privacy self-management: *New Media & Society*. (Nov. 2015).

[6] Batchelor, R., Bobrowicz, A., Mackenzie, R. and Milne, A. 2012. Challenges of ethical and legal responsibilities when technologies' uses and users change: social networking sites, decision-making capacity and dementia. *Ethics and Information Technology*. 14, 2 (Jun. 2012), 99–108.

[7] Braun, V. and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*. 3, 2 (2006), 77–101.

[8] Brooke, A. and Raine, L. 2019. Key takeaways on Americans' views about privacy, surveillance and data-sharing. *Pew Research Center*.

[9] Caregiving for Person with Alzheimer's Disease or a related Dementia | Alzheimer's Disease and Healthy Aging | CDC: 2019. *https://www.cdc.gov/aging/caregiving/alzheimer.htm*. Accessed: 2020-05-20.

[10] Cate, F.H. 2006. *The Failure of Fair Information Practice Principles*. Technical Report #ID 1156972. Social Science Research Network.

[11] Cate, F.H. 2010. The Limits of Notice and Choice. *IEEE Security Privacy*. 8, 2 (Mar. 2010), 59–62.

[12] Cavoukian, A. 2009. *Privacy by Design - The 7 Foundational Principles*. Information and privacy commissioner of Ontario, Canada, 5.

[13] Dementia cases set to triple by 2050 but still largely ignored: *https://www.who.int/mediacentre/news/releases/2012/dementia_20120411/en/*. Accessed: 2019-10-04.

[14] Facts and Figures: *https://alz.org/alzheimers-dementia/facts-figures*. Accessed: 2019-10-03.

[15] Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T. and Dell, N. 2018. "A Stalker's Paradise": How Intimate Partner Abusers Exploit Technology. (2018), 1–13.

[16] Ghosh, A.K., Badillo-Urquiola, K., Guha, S., LaViola Jr, J.J. and Wisniewski, P.J. 2018. Safety vs. Surveillance: What Children Have to Say About Mobile Apps for Parental Control. *Proceedings of the 2018 CHI Conference on Human*

Factors in Computing Systems (New York, NY, USA, 2018), 124:1–124:14.
[17] Grassi, P.A., Garcia, M.E. and Fenton, J.L. 2017. Digital Identity Guidelines. (Jun. 2017).
[18] Gray, C.M., Kou, Y., Battles, B., Hoggatt, J. and Toombs, A.L. 2018. The Dark (Patterns) Side of UX Design. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada, Apr. 2018), 1–14.
[19] Hartzog, W. 2017. The Inadequate, Invaluable Fair Information Practices. *76 Maryland Law Review 952; Northeastern University School of Law Research Paper No. 301-2017*. (2017).
[20] Hull, G. 2015. Successful Failure: What Foucault Can Teach Us About Privacy Self-Management in a World of Facebook and Big Data. *Ethics and Information Technology*. 17, 2 (2015), 89–101.
[21] Identity Theft Soared to a Record High in 2017: 2018. *http://www.aarp.org/money/scams-fraud/info-2018/id-theft-fraud-fd.html*. Accessed: 2019-12-24.
[22] Kang, R., Brown, S. and Kiesler, S. 2013. Why Do People Seek Anonymity on the Internet?: Informing Policy and Design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2013), 2657–2666.
[23] Kaye, J. "Jofish" 2011. Self-reported password sharing strategies. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada, May 2011), 2619–2622.
[24] Kokolakis, S. 2017. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security*. 64, (Jan. 2017), 122–134.
[25] Lazar, A., Edasis, C. and Piper, A.M. 2017. Supporting People with Dementia in Digital Social Sharing. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), 2149–2162.
[26] Lorenzen-Huber, L., Boutain, M., Camp, L.J., Shankar, K. and Connelly, K.H. 2011. Privacy, Technology, and Aging: A Proposed Framework. *Ageing International*. 36, 2 (Jun. 2011), 232–252.
[27] Mahoney, D., Purtilo, R., Webbe, F., Alwan, M., Bharucha, A., Adlam, T., Jimison, H., Turner, B. and Becker, S. 2007. In-home monitoring of persons with dementia: Ethical guidelines for technology research and development. *Alzheimer's & dementia : the journal of the Alzheimer's Association*. 3, (Aug. 2007), 217–26.
[28] Mentis, H.M., Madjaroff, G., Massey, A. and Trendafilova, Z. In submission. The Illusion of Choice in Discussing Cybersecurity Safeguards Between Older Adults with Mild Cognitive Impairment and Their Caregivers. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing* (In submission).
[29] Mentis, H.M., Madjaroff, G. and Massey, A.K. 2019. Upside and Downside Risk in Online Security for Older Adults with Mild Cognitive Impairment. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), 343:1–343:13.
[30] Mild cognitive impairment - Symptoms and causes: *https://www.mayoclinic.org/diseases-conditions/mild-cognitive-impairment/symptoms-causes/syc-20354578*.
[31] Morgan, S. 2018. Global Cybercrime Damages Predicted To Reach $6 Trillion Annually By 2021. *Cybercrime Magazine*.
[32] Nissenbaum, H. 2010. *Privacy in context: technology, policy, and the integrity of social life*. Stanford Law Books.
[33] Parental monitoring apps: How do they differ from stalkerware? 2019. *https://blog.malwarebytes.com/stalkerware/2019/07/parental-monitoring-apps-how-do-they-differ-from-stalkerware/*. Accessed: 2019-12-17.
[34] Patil, S. and Lai, J. 2005. Who Gets to Know What when: Configuring Privacy Permissions in an Awareness Application. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2005), 101–110.
[35] Petronio, S.S. 2002. *Boundaries of privacy: dialectics of disclosure*. State University of New York Press.
[36] Piper, A.M., Cornejo, R., Hurwitz, L. and Unumb, C. 2016. Technological Caregiving: Supporting Online Activity for Adults with Cognitive Impairments. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (Santa Clara, California, USA, 2016), 5311–5323.
[37] Sachdev, P.S. et al. 2015. The Prevalence of Mild Cognitive Impairment in Diverse Geographical and Ethnocultural Regions: The COSMIC Collaboration. *PLoS ONE*. 10, 11 (Nov. 2015).
[38] Schaar, P. 2010. Privacy by Design. *Identity in the Information Society*. 3, 2 (Aug. 2010), 267–274. DOI:https://doi.org/10.1007/s12394-010-0055-x.
[39] Singh, S., Cabraal, A., Demosthenous, C., Astbrink, G. and Furlong, M. 2007. Password sharing: implications for security design based on social practice. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA, Apr. 2007), 895–904.
[40] Spiekermann, S. and Cranor, L.F. 2009. Engineering Privacy. *IEEE Transactions on Software Engineering*. 35, 1 (Jan. 2009), 67–82. DOI:https://doi.org/10.1109/TSE.2008.88.
[41] Tejada-Vera, B. 2013. Mortality from Alzheimer's disease in the United States: data for 2000 and 2010. *NCHS data brief*. 116 (Mar. 2013), 1–8.
[42] Tene, O. 2013. Privacy Law's Midlife Crisis: A Critical Assessment of the Second Wave of Global Privacy Laws. *Ohio State Law Journal*. 74, 6 (2013), 1217–1262.
[43] The Predator in Your Pocket: A Multidisciplinary Assessment of the Stalkerware Application Industry: 2019. *https://citizenlab.ca/2019/06/the-predator-in-your-pocket-a-multidisciplinary-assessment-of-the-stalkerware-application-industry/*. Accessed: 2019-12-24.
[44] What Is Mild Cognitive Impairment? *https://www.nia.nih.gov/health/what-mild-cognitive-impairment*. Accessed: 2019-10-03.
[45] Xu, B., Chang, P., Welker, C.L., Bazarova, N.N. and Cosley, D. 2016. Automatic Archiving Versus Default Deletion: What Snapchat Tells Us About Ephemerality in Design. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (New York, NY, USA, 2016), 1662–1675.
[46] Zuboff, S. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

Blind and Human: Exploring More Usable Audio CAPTCHA Designs

Valerie Fanelle
Georgia Institute of Technology

Sepideh Karimi
Georgia Institute of Technology

Aditi Shah
Georgia Institute of Technology

Bharath Subramanian
Georgia Institute of Technology

Sauvik Das
Georgia Institute of Technology

Abstract

For people with visual impairments (PVIs), audio CAPTCHAs are accessible alternatives to standard visual CAPTCHAs. However, current audio CAPTCHA designs are slower to complete and less accurate than their visual counterparts. We designed and evaluated four novel audio CAPTCHAs that we hypothesized would increase accuracy and speed. To evaluate our designs along these measures, we ran a three-session, within-subjects experiment with 67 PVIs from around the world — the majority being from the U.S. and India. Thirty three participants completed all three sessions, each separated by one week. These participants completed a total of 39 distinct audio CAPTCHA challenges across our prototype designs and the control, all presented in random order. Most importantly, all four of our new designs were significantly more accurate and faster than the control condition, and were rated as preferable over the control. A post-hoc security evaluation suggested that our designs had different strengths and weaknesses vis-a-vis two adversaries: a random guessing adversary and a NLP adversary. Ultimately, our results suggest that the best design to use is dependent on use-context.

1 Introduction

Completely Automated Public Turing tests to tell Computers and Humans Apart (CAPTCHAs) are commonly used online to differentiate between human users and non-human bots [22]. In doing so, many CAPTCHAs ask users to engage in visual-processing tasks that are simple for humans, yet difficult for bots [20]. However, these visual-processing tasks are inaccessible to the 285 million people with visual impairments (PVIs) worldwide — 39 million of whom are totally blind, and 246 million who have low vision [35]. Instead, PVIs rely on audio CAPTCHAs, which aim to differentiate humans from bots using acoustic processing tasks.

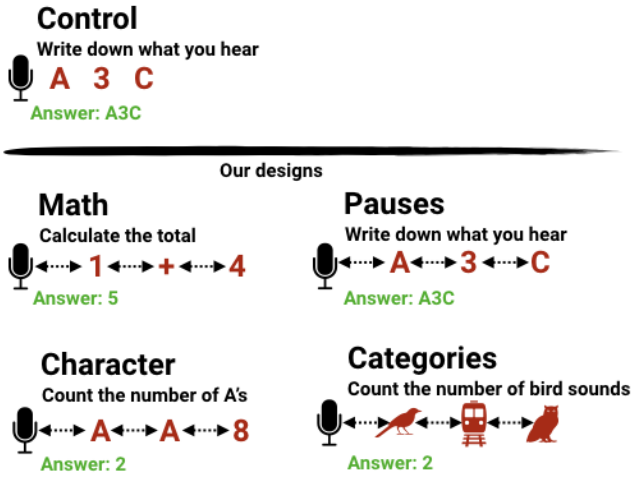


Figure 1: We designed, implemented and evaluated four novel audio CAPTCHAs. The Math prototype asked users to calculate a running total; the Character prototype asked users to count the occurrence of a character in an alphanumeric series; the Pauses prototype asked users to transcribe the alphanumeric characters they heard, but it incorporated longer pauses between characters; and, the Categories prototype, asked users to count the number of sounds, in a series, that belonged to a certain category.

In their current state, audio CAPTCHAs are significantly less usable than their visual counterparts [4, 10, 25, 37]. While visual CAPTCHAs take 9.8 seconds to solve with a 93% success rate, on average, audio CAPTCHAs take 51 seconds to solve with a 50% success rate [5, 10, 21, 33]. This

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

difference in speed and accuracy occurs because existing audio CAPTCHAs are modelled after their visual counterparts rather than using designs specific to the audio medium [5, 10, 11]. As such, existing audio CAPTCHAs require impractical levels of attention and memory-capacity from the users who depend on them [5]. This means that visual CAPTCHAs are not an equally challenging alternative to the audio CAPTCHA design; audio CAPTCHAs are more problematic for PVIs than visual designs are for fully-sighted people [17, 28, 32, 34].

Audio interference is one of the biggest issues that users face with existing audio CAPTCHAs [3]. For example, many PVIs rely on screen readers to help navigate user interfaces. When these users start typing the characters they hear in a CAPTCHA challenge, their screen reader software will read each typed letter out loud while they are simultaneously listening for the next character in the challenge. The audio conflict between the typed letter and the spoken letter thus creates unnecessary user frustration and errors. Owing to these frustrations, in a 2017 global study by WebAIM, of the 1,792 PVIs surveyed, 90% ranked audio CAPTCHAs as somewhat or very difficult [34]. These respondents also ranked CAPTCHAs as the second most problematic daily issue on the web, after Adobe Flash. The goal of our paper is to offer insights and designs that bridge the usability gap between audio and visual CAPTCHAs.

Informed by this prior work, as well as the personal experiences of one of the authors, who is blind, we followed an iterative design process to prototype and refine four new audio CAPTCHAs (see Figure 1). The Math prototype asked users to perform simple addition and subtraction; the Character prototype asked users to count the occurrence of a specific character within a string of alphanumeric characters; the Pauses prototype, which is a variation of existing alphanumeric audio CAPTCHA designs, asked users to transcribe the alphanumeric characters they heard but incorporated longer pauses between characters to minimize screen reader interference; and, the Categories prototype, asked users to count the number of sounds, in a series, that belonged to a certain category (e.g., bird chirps, baby cries).

To evaluate these designs, we were guided by three research questions — How do our novel audio CAPTCHAs compare to existing audio CAPCTHAs in terms of: **(RQ1)** task performance metrics such as accuracy and speed? **(RQ2)** security against common attacks (e.g., random guessing, machine-learning based audio classification)? and, **(RQ3)** self-reported and heuristic measures of usability?

To answer these research questions, we conducted a multi-session, within-subjects online experiment. We recruited 67 PVIs from around the world — 38 of whom live in the USA, 22 in India, 2 in Italy, 2 in Germany, 2 in the Czech Republic, and 1 in South Africa. Of the 67 PVIs, 33 participated in all three study sessions. In total, through three time-separated sessions, we asked participants to complete nine iterations of each of our four new prototypes. We recorded their accuracy and completion times with each challenge. Upon completing each challenge, we also had participants complete a brief questionnaire to gauge their in-the-moment reactions to our designs. Through a series of quantitative, qualitative and heuristic analyses on these data, we found that all of our prototypes showed statistically significant improvements in accuracy and completion time, were rated high on subjective and heuristic measures of usability and satisfaction, and were preferred over standard alphanumeric CAPTCHAs.

We also evaluated the security of our prototypes against two threat models: a random guessing adversary and an NLP adversary that leverages commercially available, state-of-the-art speech-to-text recognition and audio event classification. The control condition and our Pauses prototype offered the greatest security against random guessing attacks, but our Categories and Math prototypes offered the greatest resilience against the NLP adversary.

While all of our prototypes outperformed the control in most measures, no single design stood out as the best. The Math prototype was the most accurate, the second fastest, and provided reasonable security against both adversaries. The Character prototype was rated the most usable and satisfying, but was vulnerable against random guessing attacks. The Categories prototype was the most vulnerable against random guessing attacks, but was the fastest and most globally accessible — an important peripheral consideration, given that there are PVIs from various continents, countries, and cultural backgrounds [8, 9, 26, 35, 36]. Finally, the Pauses prototype was most preferred over the control condition, but was second lowest in accuracy and the slowest of our new designs.

2 Related Work

2.1 Challenges with Audio CAPTCHAs

In 2009, researchers at the University of Washington did a large-scale user study with 162 PVIs and found ten existing audio CAPTCHA designs to be difficult and time-consuming. They reported a 39% to 43% success rate for solving such designs on the first try and asserted that audio interfaces are often not direct translations of visual interfaces [5].

Prior work suggests that there are two types of audio CAPTCHAs: content-based and rule-based [19]. Content-based challenges require users to convert the speech of an audio file to text, an example of which is the existing alphanumeric standard. Alternatively, rule-based challenges ask users to interpret information they are hearing (e.g., 'count the number of times you hear the sound of an animal'). Rule-based CAPTCHAs can reduce the burden on short-term memory, because one only needs to remember a running total [19].

Sauer et al. studied the effects of content-based designs that closely resemble the current design norm for both visual and audio CAPTCHAs. In this study they played eight

numbers in distorted voices and asked users to input these numbers in sequence. However, they found that this technique disproportionately placed too high a cognitive load on PVIs, requiring them to either memorize the CAPTCHA series or use external tools to quickly note the entities they have heard. Due to a success rate of 46% and long average times of task completion (65.64 sec), these content-based designs exhibited low usability [13]. To address these concerns, we created low short term cognitive load CAPTCHAs that ask users to remember only one or two entities at a time. We accomplished this via rule-based designs and eliminating audio interference.

Furthermore, researchers have evaluated CAPTCHAs that employ text-based mathematical methods that ask questions such as, "What is the sum of two and four?" [17]. This text-based design is insecure due to the advancement of Natural Language Processing (NLP)-based bots [19]. Compounded with the open source tools available to adversaries online, there remains a need to create usable audio CAPTCHAs that are at least as strong as standard designs [7, 16, 30].

2.2 Improving Audio CAPTCHAs

Interesting innovation is occurring in this field. For instance, Soupionis et al. [31] implemented an audio CAPTCHA prototype for SIP-based Voice over IP (VoIP). However, their design was only tested with sighted users, raising concern about real-world outcomes for PVI's. [35].

Gao, Haichang, et al. [14] also designed a secure audio CAPTCHA that requires humans to read a sentence aloud in a natural voice and asked participants to differentiate the human voices from bots', but also only tested their design with sighted participants.

In another study [15], researchers proposed two alternative CAPTCHA designs: "auditory CAPTCHAs" and "non-nonsense CAPTCHAs," both of which were evaluated for both usability and security using Google's open-source reCAPTCHA technology. Their results showed that when comparing the accuracy levels of both humans and speech recognition algorithms, human success rates are 2.8 - 3.9 times higher. From their findings, Hendrick et al. concluded that all existing CAPTCHAs will eventually be broken, so future research should focus on incorporating human cognition as best as possible. Examples of this approach includes the use of deductive reasoning, sensory skills, and/or problem solving in order to answer correctly. This result motivated our pursuit of rule-based audio CAPTCHA designs.

Finally, Lazar et al. evaluated audio CAPTCHA designs that test sound category identification: e.g., identifying a sound clip as coming from a trumpet, a lion roaring, or a baby crying [18]. They achieved $\geq 90\%$ accuracy. However, they tested their designs with PVIs in a controlled environment with no baseline condition and with twenty participants all from the same location. This work inspired our Categories prototype, which we evaluate more broadly.

Prototype	Instructions	Sample Challenge	Correct Answer
Control (Content-Based)	Record each letter or number you hear.	8G6JVF	8G6JVF
Math (Rule-Based)	After you press play, please perform all of the calculations and provide one total at the end.	7+4-2-1	8
Character (Rule-Based)	Count the number of times '6' is spoken. Type the sum in the text box.	6R169Y6	3
Pauses (Content-Based)	Record each letter or number that you hear.	010J14	010J14
Categories (Rule-Based)	Count the number of times you hear sounds associated with those made by birds.	robin, train, motor, owl, rooster	3

Table 1: High-level summary of the prototype challenges we tested on our participants.

3 Our Audio CAPTCHA Designs

In exploring the design space for usable audio CAPTCHAs, our high-level design goals were to create challenges that minimize audio interference, reduce cognitive load (e.g., the amount of information in short-term memory), use no more than basic knowledge, and are robust against random guessing and off-the-shelf Natural Language Processing (NLP) tools.

Table 1 provides a high-level overview of each of our prototypes, along with example challenges and their corresponding correct answers. Of the four new CAPTCHAs we designed, three were rule-based in light of the aforementioned design goals. To ensure a baseline level of security in creating challenges for each of these prototypes, we followed advice from prior research to perturb the raw audio of the challenges [37]. The challenges we created contained variations in speed (very slow, slow, normal, fast, and very fast), pitch (male and female), and type of background noise. The background noises varied from public spaces (e.g., cafes), to the sounds of planes and wind. These challenges thus incorporated high randomness to complicate speech-to-text attacks.

3.1 Math Prototype

Our first prototype is a rule-based design that challenged users with basic addition and subtraction problems; some were mixed and others were exclusively focused on addition or subtraction. An example "Math" prototype challenge would be: "8 plus 4 minus 2 subtract 1 add 9" with two second gaps after each element. At each step, the user would only need to cognitively keep track of the running totals: 8, then 12, then 10, then 9, then 18. The need to do on-demand calculations might be challenging, but by keeping the operands within single digits and limiting the total number of operations, we hypothesized that the challenge would be easier and faster than the baseline control, owing to its reduced memory burden and single value entry.

3.2 Character Prototype

Our second prototype asked users to count the number of times they heard one specific character in a string of random letters and numbers. For instance, in one such challenge we asked people to identify the number of times they heard the character "s" within the string "3sjkS49sxo" — the answer being 3. Each character was read aloud with one second gaps in between. Similar to the Math prototype, we hypothesized that this design would result in greater accuracy and faster input completion speeds. This was due to the reduced cognitive demand of the need to keep track of just one running total and entering in only one input at the end of the recording.

3.3 Pauses Prototype

Our third prototype was a slight modification of the standard, content-based, alphanumeric CAPTCHAs that ask users to type, in sequence, all the characters heard in a random string. The key difference is that we included a two second pause in between characters to mitigate the interference between screen reader transcription and the challenge characters screen readers read aloud. This design should be relatively simple to deploy given its similarities to existing audio CAPTCHAs.

3.4 Categories Prototype

Our final prototype asked users to count the number of times they heard a certain "category" of sound (e.g., bird chirps, cars honks) embedded in a string of other sounds. Each sound was separated by a two second gap. For example, a user might have been asked to identify the number of times they heard birds chirping within a stream of sounds like trains and vehicular motors. The user answered with the total number of bird sounds detected throughout the CAPTCHA. Similar to the first two rule-based designs, due to reduced cognitive load, we expected this CAPTCHA to be completed with higher accuracy and speed than the control condition. A peripheral benefit of this design is that it is language-agnostic, though we note that there may sometimes be cultural differences in category membership — e.g., whether a rooster's crow should be counted in the bird chirp category.

4 Evaluation Methodology

We ran a controlled, within-subjects, online experiment with 67 blind and visually impaired users from around the world. Our study was IRB-approved.

4.1 Experiment and Procedure

Our experiment consisted of five conditions: four new designs and one baseline control condition that was used to emulate the industry-wide standard alphanumeric audio CAPTCHAs. To account for novelty and learning effects, we conducted three time-separated sessions, each spaced one week apart. In each session, participants completed three audio CAPTCHA challenges for each of our designs, and one challenge for the control. In total, participants were presented with the same 13 challenges per session in a randomized order.

We used Audacity, an open source audio-editing software, to create each CAPTCHA. Individual clips for each character and word were generated using a text-to-speech program that can synthesize audio in both male and female voices. These audio files included characters like 0-9, a-z, words for add, subtract, plus, and minus. We also accumulated open source audio clips of varying phenomena (i.e. birds chirping, instrument recordings, etc.) and background noises [1, 2]. These clips (apart from the categorical sounds used in the Categories prototype) were then distorted by applying audio effects that changed each clip's pitch, speed, and amplification. We used the same set and number of audio clips to create the 39 challenges and all CAPTCHAs were merged with distorted background noise at the same decibel level. These distortions were used to improve both the security and ecological validity of our designs [1, 2, 19].

The resulting audio CAPTCHAs were 16 to 18 seconds long, with a one second pause at the beginning to enable users to navigate to the edit text box to record their answers. In order to mimic existing audio CAPTCHA designs, the control challenge did not have an initial one second pause.

The web platform we designed to administer our CAPTCHA designs was tested in a preliminary pilot study in early 2019. According to feedback from the pilot, we then adjusted three of our designs, replaced another one entirely, and conducted a 3 week long within-subjects experiment in the summer of 2019.

We developed the online experimental test-bed using jQuery and HTML5 for the front-end, PHP for the back-end, and Heroku for hosting. In consultation with one of our research team members, who is visually impaired, we kept the user interface simple and accessible for PVIs who would need to navigate the interface with screen readers.

Participants first encountered a landing page in which they could see details about the study and provide informed consent. Next, participants were asked to complete a challenge under the standard design (control), followed by batches of three challenges each for our four custom prototypes (treatment). We randomized the order in which the treatment prototypes were presented to each user. An example challenge is illustrated in Appendix 4.

For each participant's first session, we conducted a video conference call on Zoom [38] in order to ensure that participants could use the experiment test-bed and complete the subsequent two sessions independently. We asked participants to share their screens to confirm their use of a screen reader to complete the study. Throughout the duration of the session we guided them between pages and answered their questions. One week after the first and second sessions were completed, we then emailed participants a subsequent link to complete the second and third sessions, respectively.

After completing the batch of 3 challenges for each prototype in each session, participants filled out a questionnaire in which we asked them to rate, on a Likert scale from 1 - 5, the usability of and their satisfaction with each prototype — "1" was coded as very low and "5" was coded as very high. We then asked participants if they preferred the prototype in comparison to the control. We also asked open-ended questions to solicit participants' thoughts and feedback on our designs. For the first session, we asked participants these open-ended questions in real time via Zoom. For the latter two, participants wrote-in their responses manually. Finally, at the end of the study, we collected each participant's age.

The data streams that informed our findings include quantitative and qualitative data, as well as our own facilitator observations, from both this study and the pilot in 2019.

4.2 Recruitment and Compensation

We reached out to a number of global organizations, including the American Foundation for the Blind, the National Federation of the Blind, Braille Works, the American Printing House for the Blind, the Blind Graduate's Forum of India, and VisionAid. We also leveraged blind social and support groups on social networking platforms (Facebook, WhatsApp, etc.) and mailing lists (Access India, Program-L and Voice Vision).

In total, we received 225 responses as a result of this outreach. Due to time and resource constraints we scheduled sessions with 150 participants over the course of six weeks, choosing participants in the order that we received their information. Accounting for those who dropped out or never responded, we interviewed 67 participants for at least one session. Each person was compensated 10 USD per completed session, for a total of 30 USD for completing all three sessions. Compensation was distributed in the form of regional Amazon gift certificates.

Our primary criteria for determining participant eligibility was their use of low-vision assistive technologies (e.g., braille displays, screen readers, and screen magnifiers) to navigate computer screens. Only one of our participants relied on screen magnification software. Although he had some vision, it was not clear enough for him to be able to solve visual CAPTCHAs. All other participants used screen readers and none used braille displays. Thus, we use the term "people with visual impairments" (PVI) to describe all of our participants.

	Accuracy Model (Logistic)	**Time Model (Linear)**
Fixed effect coefficients		
Session Number	0.35*	−0.17***
Math v. Control	2.78***	−0.73***
Character v. Control	2.50***	−0.70***
Pauses v. Control	1.77**	−0.61***
Categories v. Control	1.53*	−0.76***
Character v. Math	−0.27	0.03
Pauses v. Math	−1.00	0.12
Categories v. Math	−1.24*	−0.03
Pauses v. Character	−0.73	0.09
Categories v. Character	−0.97	−0.06
Categories v. Pauses	−0.24	−0.15
Age	0.005	0.002
Intercept	−0.87	−0.89***
Random intercepts variance		
Participant (N=67)	0.50	0.19
Challenge (N=39)	0.61	0.02

p <= 0.05, ** p <= 0.01, *** p <= 0.001

Table 2: Mixed-effects regression modeling both accuracy and completion time against prototype, session number, and age, with each participant and each challenge having its own random intercept. For the accuracy model, we ran a logistic regression and for the completion time model, we ran a linear regression. Broadly, the highlighted rows on the top indicate that all of our prototypes were significantly more accurate and faster than the control, and that participants grew more accurate and faster in subsequent sessions. The variance in random intercepts suggest significant variation across participants and challenges in success but not in completion time.

5 Results

5.1 Participant Demographics

Sixty-seven PVIs participated in the first zoom session; 34 of these continued on to remotely complete the two remaining sessions of our study. All participants were at least 18 years old and were, on average, 33.1 ($\sigma = 15.3$) years old. We did not collect gender data. All participants were able to speak, read, and write proficient English, which was verified in the first session via a face-to-face screen-sharing video chat.

5.2 Data Pre-Processing

Across all participants, we collected data on 2,259 CAPTCHA attempts. Of these attempts, we dropped 11 data points that were corrupted through data collection errors (i.e. with infeasible completion times of over 50 years, which we suspect is due to improperly set browser clocks). We also dropped one extreme outlier with a completion time of 93 minutes, or 46 standard deviations away from the dataset's mean completion time ($\mu = 36.9$ seconds, $\sigma = 120.6$ seconds). We suspect this participant left their browser window open while being away. Thus, we dropped 12 data points in total (0.5%). Our final dataset consisted of 2,247 CAPTCHA attempts from 67 PVIs.

5.3 RQ1: Task Performance Evaluation

We first evaluated how our novel designs compared to the control condition across two important task performance metrics: accuracy and completion time.

5.3.1 Accuracy

Across all our participants, in decreasing order, the accuracy rates for each prototype were: 89.2% for Math, 86.9% for Character, 76.2% for Pauses, 70.3% for Categories, and 42.9% for the control.

To test if these differences in accuracy were statistically significant, we modeled the accuracy of our designs with a random-intercepts logistic regression using the lme4 package in R. Our input data were the 2,247 individual attempts at solving a CAPTCHA challenge. Our dependent variable was a binary measure of whether or not a participant successfully completed the challenge. Our IV was the prototype used in the challenge — a categorical variable encompassing our four treatment designs and the control. As covariates, we included the session number and participant age. We also included a random intercept term for the 67 distinct participants and the 39 distinct challenges to account for and model the effects of repeated observations. We used R's multcomp package to conduct pairwise comparisons between each of the ${}_5C_2 = 10$ combinations of our novel prototype designs and the control. The results are shown in the first column of Table 2, with p-values adjusted using Bonferroni correction.

Most importantly, **we found that each of our prototypes — Math ($b = +2.78$), Character ($b = +2.50$), Pauses ($b = +1.85$), and Categories ($b = +1.50$) — were completed with significantly higher accuracy than the control.** We also found evidence of a learning effect: participants were significantly more accurate in later sessions ($b = +0.35$). After correcting for multiple testing, we did not find many statistically significant differences in accuracy between our four designs, with one exception: the Categories prototype was significantly less accurate than the Math prototype ($b = -1.24$).

The variance in random intercepts across distinct participants ($\sigma^2 = 0.50$) and challenges ($\sigma^2 = 0.61$) suggests that

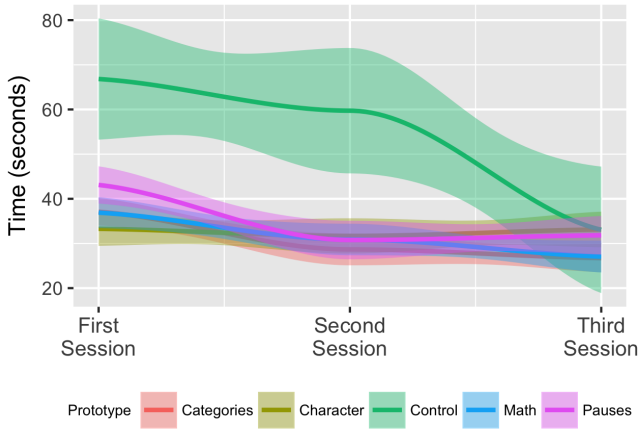


Figure 2: Distribution of average completion times, with 95% confidence intervals, for each prototype across sessions. There is evidence of a significant learning effect in which completion times drop with increased accuracy and repeated exposure. Thus, participants learned and adapted to our novel designs.

performance could vary in non-trivial ways between individual participants and across different challenges. To better illustrate this point, our most successful participant (a 49 year old from the USA) got 100% of their challenges correct, while our least successful participant (a 41 year old from India) got 46% of their challenges correct. Our most successful challenge was solved with 97.5% overall accuracy (the third challenge of the Categories prototype in the second session), while our least successful challenge was solved with 17.2% accuracy (the control challenge in the second session).

5.3.2 Completion Time

We next investigated how our designs varied by completion time. Broadly, the average completion time for a challenge was lowest for the Categories prototype (31.1 s), followed by the Math prototype (31.7s), Character (32.7 s), Pauses (35.4 s) and, finally, the control (53.6 s).

To test if these differences were statistically significant, we ran a second random-intercepts regression. The model parameters were the same as the aforementioned accuracy model, although with two exceptions: the DV was scaled and centered for time taken to complete each design. Because the DV was continuous instead of binary, we employed a linear regression. Once again, we used R's lme4 package to estimate the model, and R's multcomp package to do pairwise comparisons across the prototype designs, with p-values adjusted using Bonferroni correction.

The results can be seen in the second column of Table 2, labeled "Time Model." For a numeric predictor (i.e., Age, Session Number), the model suggests that a positive coef-

ficient of $b = +1.0$, as the predictor increases by one unit, means the estimated completion time will *increase* by one standard deviation. A negative coefficient implies that the estimated completion time would *decrease* by one standard deviation. For a categorical predictor (i.e, Prototype design), a positive coefficient of $b = +1.0$ would suggest that the estimated difference in completion time between two levels of the predictor — a treatment and a control (e.g., Math vs. Control) — is one standard deviation, with the comparison condition taking longer than the control. A negative coefficient implies the opposite — that the control takes one standard deviation longer than the comparison. All of these examples assume that all other predictors (IVs + covariates) are held constant.

We found that all of our novel prototype designs — Math ($b = -0.73$), Character ($b = -0.7$), Pauses ($b = -0.61$), and Categories ($b = -0.76$) — were significantly faster than the control. We did not find a significant difference in completion time between any two of our custom prototypes. We again found a learning effect: participants were significantly faster in later sessions ($b = -0.17$). Figure 2 illustrates the distribution of completion times across all four of our prototypes and the control condition, and shows how those distributions vary across sessions. On average, completion times decreased over the course of all three sessions for every prototype design, most significantly for the control.

The variance in random intercepts across distinct participants ($\sigma^2 = 0.19$) and challenges ($\sigma^2 = 0.02$) was fairly small, suggesting that, accounting for the fixed effects in the model, completion times did not dramatically vary between participants and challenges.

5.4 RQ2: Security Evaluation

While we consider our primary contribution to be a usability assessment of alternative audio CAPTCHA designs, we also evaluated the security of our prototypes relative to the control condition. We considered two threat models.

The first is a *random-guessing adversary*. For content-based CAPTCHAs like the control and Pauses prototypes, this adversary is trivially defeated. Assuming a 32-character alphabet (all English letters along with 0 - 9 digits), a random string of 6 characters would yield a search space of 32^6 possibilities, which would be impractical to randomly guess. For our rule-based audio CAPTCHAs, however, random guessing is more potent. For the Character and Categories prototypes, the space of possible outputs given a 10-character long string is 0 - 10 — i.e, the random guessing adversary would have a $1/11 \approx 9\%$ success rate. For the Math prototype, we assume five single-digit operands connected through either plus or minus operators, while there are $10^5 + 2^4 = 100,016$ possible inputs, the range of possible outputs varies from $[0-9-9-9-9 = -36, 9+9+9+9+9 = 45]$. Thus, there are 72 possible outputs, so a random guessing adversary would have a $1/72 \approx 1\%$ chance at breaking the Math prototype. A smarter adversary might notice that the distribution of outputs is not uniform but a normal distribution centered at 5. Thus, by guessing "5" on every attempt they would increase their chances of defeating the Math prototype to $\approx 3\%$.

The second adversary we considered is one who uses state-of-the-art NLP — either commercially available or easily trainable using public-domain knowledge and data — to deconstruct the audio file and solve the challenge. Motivated by Polakis et al. [27], for the Math, Character, and Pauses prototypes, we tested the robustness of our designs using Google's automated, off-the-shelf speech recognition software. We considered a clip broken if all the entities were successfully parsed from audio to text, because once the content is parsed, the application of rules to get the correct answer is trivial. For the control prototype, 0 out of 3 (0%) designs were successfully parsed. The percentages of challenges within each design that were broken by this threat model include: 2 out of 9 (22%) for Math; 1 out of 9 (11%) for Character; and 6 out of 9 (67%) for Pauses.

The Categories prototype could not be tested using off-the-shelf parsing services because an appropriate parser for real-world sound classification does not exist. Thus, we created our own parser using deep learning. We implemented this parser on Tensorflow and trained it on Google Research's Audioset data, a collection of 632 audio event classes and over 2 million 10 second sound clips labeled by humans [12]. We considered a clip broken if the parser was able to predict 'true positives' and 'true negatives.' For instance, if the CAPTCHA challenge is to count the total number of animal sounds, the 'true positives' include sounds that are identified as animal-related and 'true negatives' are successfully identified as non-animal sounds. Similarly, 'false positives' are identified as sounds that are incorrectly labeled as animal sounds and 'false negatives' are animal sounds that are incorrectly labeled non-animal sounds. We found that of all predictions made for 48 sub-clips across all categories, challenges, and sessions, 13 were true positives and 16 were true negatives. There were 8 false positives and 11 false negatives. The average error of 2.1% per clip contributed to either false positive or false negative predictions, so we can deduce that none of the clips belonging to the Categories prototype were fully parsed.

In summary, the random-guessing adversary would be trivially defeated by the Pauses prototype, but could have a small chance at defeating the Math, Character, and Categories prototypes. This could be made harder by increasing the length of the CAPTCHA, though likely at the cost of usability, speed, and accuracy. The NLP adversary would have good success at breaking the Pauses prototype, and a slightly better chance at breaking the Math and Character prototypes than the random-guessing adversary, but would struggle with Categories.

Next, we considered how these results compared to the control condition. Prior work has shown that standard audio CAPTCHAs are largely insecure against state-of-the-art machine learning. For instance, Bursztein et al. (2011)

showed that 45% of Yahoo, 49% of Microsoft, and 83% of eBay CAPTCHAs can be broken. Also, Sano et al. (2013) and Meutzner et al. (2014) successfully broke Google's reCAPTCHA with a success rate of 52% and 63% while Tam et al. (2009) achieved an accuracy of up to 71% using machine learning techniques like AdaBoost, SVM, and k-NN. With the deep learning evolution, audio CAPTCHA attacks continue to succeed. In 2017, Saumya et al. [27] developed a low cost 'AudioBreaker' system using off-the-shelf speech recognition services. It successfully broke seven popular audio CAPTCHA schemes along with 98% accuracy in breaking Google's reCaptcha. In 2019, Heemany et al. [29] even demonstrated 85% accuracy in breaking designs with higher background noise levels.

In short, all of the CAPTCHAs we considered, including the control, could be broken by motivated adversaries. Thus, we must consider the use-context. While our designs should not be used for security-critical applications, they should provide sufficient security for low-risk contexts in day-to-day web browsing (e.g., comment form submission).

5.5 RQ3: Usability Evaluation

Finally, we conducted a series of quantitative, qualitative and heuristic usability evaluations based on our observations of participants in the initial Zoom session as well as participants' survey responses.

5.5.1 Usability, Satisfaction, and Preference

After each challenge, participants rated the usability and satisfaction of our designs on a 5-point Likert scale and also answered whether or not they preferred our design over the control. This gave us 2,247 usability, satisfaction and preference data points. Due to the Hawthorne effect, the absolute values of these ratings were not as important as their relative ordering, which helped illuminate participant preferences.

The distributions for usability and satisfaction were highly skewed, with participants rating 1,916 challenges a "5" on usability and 1,865 a "5" on satisfaction. To simplify analysis, we converted these scales into binary values: "5" or not "5." We then conducted three random intercepts logistic regressions, using R's lme4 package, correlating usability, satisfaction and preference-over-control to prototype design. We included a random-intercepts term for participant and challenge to control for repeated observations. We ran pairwise comparisons using R's multcomp package, adjusting p-values with Bonferroni correction. The results are shown in Table 3.

Usability. The regression results in Table 3 suggest that, controlling for the effects of individual preference, challenge variance, and the session number, the Character and Categories prototypes were rated significantly more usable than Math and Pauses; the Math prototype was rated more usable than Pauses; and the Pauses prototype was rated less usable than the other three. We also found a significant positive effect of Session number, suggesting that participants found all prototypes more usable in later sessions.

	Satisfaction	Usability	Pref. Over Control
Character vs. Math	1.31***	0.96**	1.11***
Pauses vs. Math	−0.47	−0.57*	0.36
Categories vs. Math	0.32*	0.74***	0.34***
Pauses vs. Character	−1.79***	−1.52***	0.76***
Categories vs. Character	−0.99**	−0.22	−0.77***
Categories vs. Pauses	0.79**	1.31***	−0.02
Session	0.32**	0.74***	0.34***
Intercept	3.05***	2.65***	0.42

* p <= 0.05, ** p <= 0.01, *** p <= 0.001

Table 3: Random-intercepts logistic regression results modeling satisfaction, usability and preference as a function of prototype design and session number. The Character prototype was rated the most usable and satisfying. The Pauses prototype was most preferred over the control.

Satisfaction. Table 3 also shows that the Character prototype had a significantly higher satisfaction rating than all other prototypes; the Categories prototype out-performed Math and Pauses; and there was no significant difference found between Pauses and Math. We also found a significant positive effect of Session number, again suggesting that participants' satisfaction increased in later sessions.

Preference. Overall, participants reported preference for our prototypes over the control: 73% preferred Pauses to control, 67% preferred Character, 61% preferred Categories and 52% preferred Math. Controlling for repeated observations, challenge exposure, and session number, Table 3 shows which pairwise differences are statistically significant. In short, Pauses was preferred more than Character; Character was preferred more often than Categories and Math; and Categories was preferred more often than Math. Participants' overall preference for our prototypes over control also increased in later sessions.

In sum, the Character prototype had the highest usability and satisfaction ratings and was the second most preferred over control after Pauses. The Math prototype was generally rated the least usable, least satisfying, and least preferred. This result presents an unfortunate dilemma — the prototype that provided the highest accuracy and second highest speed was also the least subjectively "usable."

5.5.2 Heuristic Analysis: Quantitative

Beyond individual perceptions of usability, we next performed a quantitative heuristic analysis to assess the usability of our prototypes. Jacob Nielson's heuristics for designing usable systems span five core components to ensure design quality:

learnability, or the ease of correctly completing a CAPTCHA upon initial exposure; *efficiency*, or the rate at which users can learn how to complete new designs; *memorability*, or the ability to re-learn the correct use of a CAPTCHA design after a period of inactivity; *errors*, or the extent and nature of errors users make; and finally, *satisfaction*, or users' level of enjoyment in completing the CAPTCHA [24].

In order to *quantitatively* measure these criteria, we used the following data to address Nielson's heuristics:

1. *Learnability*: the session 1 accuracy rates for our prototypes against the control design.
2. *Efficiency*: whether the average number of replays decreased across sessions 1 through 3.
3. *Memorability*: whether the accuracy rates for our prototypes increased between sessions 1 through 3.
4. *Errors*: the average time it took for users to complete each challenge accurately.
5. *Satisfaction*: the self-reported user satisfaction scores (1 - 5) for each prototype.

Table 4 illustrates that the initial *learnability* of our designs, as measured by average accuracy in session 1, is comparatively higher than that of the control CAPTCHA. These numbers indicate that our designs were more learnable than the control, despite the fact that users had the most exposure to the control from day-to-day web browsing. Additionally, the significantly higher initial accuracy of the Pauses prototype, which was identical to the control apart from small time gaps in-between characters, signal that users need slower-paced CAPTCHAs to answer them correctly.

In terms of *efficiency*, two of our designs showed a steadily decreasing number of replays in subsequent sessions. The Math and Categories prototypes displayed clear improvements — users required fewer replays in session 3 than in session 1. The control also required fewer replays in subsequent sessions, but still had the highest average number of replays per session compared to all other designs.

Similarly, in terms of *memorability*, only the Math and Categories prototypes had improved accuracy scores in subsequent sessions. Users' subjective ratings of these prototypes — usability, satisfaction, and preference over the control — also improved over time (see Figure 3).

In practice, all of our designs and control, if answered correctly the first time, should take about the same amount of time to complete. However, users spent the longest time completing the control CAPTCHA correctly, suggesting that it was the most prone to *errors*. Generally, the Categories prototype was fastest: it took 9.4 fewer seconds to accurately complete than the control. The Math prototype was second fastest, followed by Character and then Pauses.

As we did not collect subjective feedback for the control, we are unable to contrast the *satisfaction* scores of our novel designs vis-a-vis the control. However, as we saw in the previous section, Character and Pauses had the highest satisfaction scores of 4.93 and 4.85, respectively.

5.5.3 Heuristic Analysis: Qualitative

Next, we used a two-dimensional subset of Yan et al.'s [37] qualitative usability assessment framework to analyze participants' open-ended feedback, in order to better understand their perceptions and difficulties with each of our new designs. Specifically, the two dimensions we qualitatively assessed were:

1. *Distortion*: level and type of distortion, use of confusing characters, and design difficulty for native and non-native speakers.
2. *Content*: language specificity of the characters used, the length of each CAPTCHA challenge, length of answers, and predictability of the design.

We start with a broad assessment of these dimensions, and then discuss individual participant feedback pertaining to these dimensions for each of our prototypes.

In terms of *distortion*, the audio files we used in the control, Math, Character, and Pauses prototypes were drawn from the same set of letters, numbers, and operators that varied in terms of voice, speed and pitch. As a compromise between usability and security, we picked sounds that were dynamic and not overwhelmingly loud. Because of the distortion of background noises and characters, participants noted that certain letter groupings like "2," "q," and "u" were hard to distinguish. Other participants with hearing problems had difficulty understanding some of the deeper voices that resulted from very slow audio speeds.

Additionally, all prototypes except Categories relied on the user's ability to understand letters, numbers, and mathematical operators spoken in English. However, knowledge of these fundamentals is certainly attainable for non-native speakers since challenge instructions can be translated by web pages into nearly any language. The Categories prototype, in its use of more universal audio events, had the fewest language-specific constraints [6] [37].

In terms of *content*, as portrayed in Table 1, the six-to-eight-character length strings of our CAPTCHA challenges were comparable to existing designs. While challenge length and instruction sets for all our prototypes were predictable, their content varied in predictability. Of our novel designs, Pauses was the most familiar, while the other designs were based on less predictable rule-based methods. However, the higher average accuracy of Math and Character prototypes over the Pauses prototype and control suggests that content predictability is not necessary for success in usage.

	Learnability	Efficiency	Memorability	Errors	Satisfaction
	Avg. session 1 accuracy	**Num. replays decrease across sessions?**	**Improved accuracy across sessions?**	**Avg. time to answer correctly (seconds)**	**Avg. satisfaction score (1-5)**
Control	32%	Y	N	39.0	N/A
Math	83%	Y	Y	30.2	4.6
Character	89%	N	N	30.1	4.8
Pauses	72%	N	N	34.6	4.8
Categories	56%	Y	Y	29.6	4.5

Table 4: This table illustrates the quantitative usability assessment based on Nielson's criteria. We found that all of our designs exemplified higher usability than the control, with the exception of the efficiency and satisfaction categories, which are discussed in more detail below. Notably, Categories scored well in the Efficiency, Memorability, Errors, and Satisfaction heuristics.

Math Prototype:

Front-heavy errors. Many of the mistakes participants made with this prototype occurred in earlier sessions, owing at least partially due to confusion with the *content* of instructions. For instance, a few users who had misunderstood the challenge submitted a string of the entire equation rather than providing a single sum. This content-specific error, which we also observed for the Character prototype, was likely due to the conflation of instructions with existing content-based designs that require users to repeat exactly what they hear.

Since average accuracy rates improved over time, from 84.8% in session 1 to 92.8% in session 3, we suspect that accuracy may continue to increase as familiarity with the design grows. Participants reported feeling better prepared to use the Math prototype in later sessions. For example, one 38 year old participant from the Czech Republic said, *"It is simply usable if people remember the last result. I had problems in previous runs but in this I learned how to concentrate."*

Accessibility concerns. Some participants were concerned about the accessibility of a math-based design, namely regarding the *distortion* and *content*. For instance, a 46 year old from the USA said, *"It wasn't difficult for me but sighted individuals do not have to do math like this and I don't feel I should have to be challenged in a way that others are not. Having said this I found it easy to use. I am concerned if this model were used with persons who had cognitive challenges or if it were used with children the task could be too complex."* Another participant noted that we failed to consider users with multiple physical impairments, such as loss of vision *and* hearing. This tendency to speak for less cognitively-able members of the PVI community was quite common, as research shows there is a trade-off in that advanced cognitive abilities allow audio CAPTCHA users to complete challenges faster [23]. These users were hinting at a fairness divide between visual and audio CAPTCHAs that emerged from real usability differences between the two authentication methods.

Character Prototype:

Confusing instructions. Similar to the Math prototype, many of the errors that users made with the Character prototype were due to confusion with the *content* of instructions we provided. Recall an example of the provided instructions which were: "You must count the number of times '6' is spoken throughout the audio clip. Type the sum in the text box at the end." In evaluating that challenge, a 55 year old user from India said, *"The term 'sum' is confusing particularly when we are asked to count the number of times '6' is spoken throughout the clip and to write the sum in the box. If '6' is spoken three times then we should write '3' or '18'? Hence more clarity is needed on this type of CAPTCHA."* This signals that the exact wording of instructions must be carefully considered before implementation.

Similar sounding characters. Participants also pointed out areas of difficulty related to *distortion*, such as the length of the alphanumeric string being too long or that letters with similar phonics sounded too similar to each other. Examples of confusing letter groupings were: ("2", "q", "u") and ("b", "e", "z", "d", "v", "p", "c", "t", "g"). This issue is inherent to all alphanumeric audio CAPTCHA designs, further suggesting the need for exploring non-language based designs.

Ease. Despite the aforementioned challenges, users generally found this prototype easy to use, thus suggesting positive outcomes related to Yan et al.'s heuristic usability criteria [37]. Several participants reported that its difficulty level was comparable to their perceptions of visual CAPTCHAs, which was one of our design goals. A 49 year old from the USA stated, *"I was able to understand all the letters and numbers even with the distortion of the sounds. Counting letters and numbers is easier than trying to remember or type in the whole set which sometimes requires listening to it 2-3 times."*

Pauses Prototype:

Accounting for hearing loss. Participants noted the impor-

tance of accounting for hearing loss, particularly in content-based CAPTCHAs like the control and Pauses. For instance, a 34 year old from the USA stated, *"Qs are spoken rather deep and some people have trouble hearing very deep voices such as myself. I have slight hearing loss in my left ear that makes it almost impossible to hear deep [male] voices so I would raise the pitch on Qs."* Participants indicated that higher quality audio samples with a consistent volume level could have improved the accessibility and *distortion* of this design.

Longer pauses were helpful. Participants found the longer gap between characters helped improve accuracy and deal with interruptions. A 58 year old participant from India reported: *"The pauses are helpful to solve the CAPTCHA and hence need to be implemented."* Similarly, a 27 year old from the USA explained how the extra gaps between characters afford greater flexibility and ease of use: *"This time I realized that this layout is much better than I thought. I was interrupted by someone asking me a question and I was able to record the last few characters and play it again to get the first ones. The gaps are so long that I believe people will also be able to find where they left off and keep going."* Overall, this feedback suggests that the *content* of Pauses was usable.

Categories Prototype:

Ambiguous category membership. Category membership can be culturally-specific. When we asked users to identify the number of times a bird sound was played, a few questioned whether a rooster is a bird. In fact, two participants noted that they associate the sound of roosters with their alarm clock, which led them to disassociate the sound of a rooster with a bird. Participants also thought animal categories can be too culturally-dependent and thus should not be used. In other words, in order to overcome barriers related to both *content* and *distortion*, sound categories should be specific and tailored to certain locales or universally recognizable.

Instrumental sounds could also be ambiguous at times and so participants needed instructions to identify a specific type of instrument such as a guitar. One 20 year old participant from the USA said, *"I strongly disagree with this design because it's asking people to make associations. Almost any sound can be associated with a musical instrument."*

Non-linguistic CAPTCHAs may be more universally appropriate. Other participants appreciated that Categories did not require knowledge of the English language. For example, a 21 year old participant from India stated, *"I was thinking these CAPTCHAs might be excellent for people whose their main language is not English and would be a great help for them. For example in many websites the CAPTCHAs are in English and I've talked with some friends who don't speak English at all. They used to tell me that due to these types of CAPTCHAs they needed to find help from their families."*

Fun and ease. Several users expressed that this design was more "interesting", "fun", and easier than alphanumeric alternatives. For instance, a 41-year-old from Italy stated, *"It is far more interesting than typing in numbers or letters because it is language independent, more pleasant and less cognitive demanding."* Likewise, a 27-year-old participant from the USA said, *"It is very easy to use because people can easily identify these common noises. It also requires less brain power than math or memorizing a long string of characters."*.

	Avg Accuracy	Avg Speed	Pref. to control	Security Random	Security NLP
Control	43%	53.6s	2.7%	++	-
Math	89%	31.7s	52%	+	+
Character	87%	32.7s	67%	-	+
Pauses	76%	35.4s	73%	++	-
Categories	70%	31.1s	61%	-	+

Table 5: Summary of key results. The Math prototype had the highest accuracy; the Categories prototype had the highest speed; the Pauses protootype was most preferred; the control and Pauses prototype were most resilient to random guessing; and, the Math, Character and Categories prototypes were most resilient to NLP.

5.5.4 Ad-hoc Usability Observations

Through our observations, user feedback and trial-and-error, we uncovered a number of one-off design attributes for audio CAPTCHAs to improve usability and accessibility. First, participants found 'auto-play' features to be a nuisance that rushed them through the task before they were ready if they, e.g., accidentally skipped through text instructions. Additionally, we found that placing a one-second gap, between hitting the play button and hearing the first character, helped both usability and accessibility. Finally, from our experience with conducting both the pilot and the full study, we found 1.25 seconds to be the optimal time gap between audio clips, regardless of prototype.

6 Discussion

Table 5 summarizes our designs, relative to the control, across key dimensions of interest. Our high-level goal was to design audio CAPTCHAs that were faster, more accurate, and that provided reasonable security. Our key results suggest that all four designs were significantly more accurate and faster (RQ1) than the control. The Math and Character prototypes showed average accuracy rates of 89% and 87%, respectively, which are on par with traditional visual CAPTCHAs. The Categories prototype was the fastest to complete (31.1 s), with the Math prototype being a close second (31.7 s). In terms of security (RQ2), the Math prototype provided decent resilience against both of the adversaries we tested. The Categories and Character prototypes were more vulnerable to random

guessing, while the Control and Pauses prototype were more vulnerable to NLP. Finally, through a series of quantitative, qualitative and heuristic usability analyses (RQ3), we found that the Pauses prototype was most preferred; the Character prototype was most satisfying; the Categories prototype was most globally accessible; and, the Math prototype was least usable, satisfying and preferred.

Based on the diversity of these results, the best design to use is dependent on use-context. The Math prototype provides the best balance of task performance and security against both types of adversaries, but was perceived to be least usable. The Characters prototype was vulnerable to random guessing but was the highest rated in usability and satisfaction. The Pauses prototype was most preferred over the control condition and would be the easiest to deploy in its similarity to existing audio CAPTCHAs. Finally, the Categories prototype was the fastest, inspired the most positive qualitative feedback, and utilized language-agnostic challenges.

6.1 Practical Design Recommendations

Through a combination of trial-and-error, along with open-ended feedback from participants in our pilot study, we have distilled a number of practical recommendations for designers or researchers who might use or improve upon our prototypes:

- Provide ≥ 1 second of silence after the user presses play;
- Place 'play' button beside the answer box without 'auto-play' functionality;
- Place 1.25 second gaps between audio clips;
- Avoid language or cultural-based challenges in favor of ones with universal sounds (i.e. running water);
- Choose specific sound categories when asking users to count non-alphanumeric sounds;
- Consider the loss of various physical abilities in users;
- Code instructions as audio elements to prevent skippage;
- Use high quality audio samples;
- Maintain a consistent volume level for all audio clips.

6.2 Limitations and Future Work

Participant Retention. Participant retention is a common limitation in multi-session studies. Our study took place over three time-separated sessions and required participants to complete the final two sessions independently. While we wanted these to be completed one week apart, some participants procrastinated for a week or more. Additionally, 34 of the initial 67 participants did not complete the final two sessions at all.
One-second pause. We incorporated a one-second pause at the beginning of our new designs so that screen reader users could navigate to the answer box before the challenge began. However, we did not incorporate this pause to the beginning of the control because we wanted it to emulate a real-world baseline, and existing audio CAPTCHAs do not have an initial one-second pause. This discrepancy could have increased the performance of our designs relative to the control.

Ecological validity. Our experiment test-bed was uniquely accessible. In practice, our designs may be embedded within otherwise inaccessible websites, which could impact PVI's performance. While the relative results between the conditions we tested should hold, in practice, accuracy and speed may be different. A field study of our novel audio CAPTCHA designs may help address these concerns in future work.

Intersectional accessibility. Our designs were for PVIs, but we did not consider other disadvantages and impairments. For example, our designs assumed participants did not have hearing impairments. Our Math prototype assumed it was simple for people to do mental arithmetic. Our Categories prototype did not consider cultural influences on category membership. A fruitful area of inquiry for future work may be designing human-intelligence proofs that don't rely on the acuity of human senses, that don't pre-suppose educational background and cognitive abilities, and that are more culturally inclusive.

7 Conclusion

Motivated by the usability shortcomings of existing audio CAPTCHAs, we designed, implemented and evaluated four alternatives that we hypothesized would improve the audio CAPTCHA user experience for people with visual impairments. We experimentally tested this hypothesis in a controlled, randomized within-subjects experiment with 67 PVIs and found that all of our designs significantly outperformed the control condition in both performance measures (accuracy, completion time) and perceptions of usability. None of our designs stood out as a clear winner. Rather, each of them boasted complementary improvements to the user experience — the Math Prototype was the most accurate and second fastest, the Character prototype was rated the most usable and satisfying, the Pauses prototype was the most familiar and preferred over the control, and the Categories prototype was the fastest and most globally accessible. These improvements, however, came at the expense of increased vulnerability against random guessing attacks for three of our four designs. For use-cases where high security is not critical — e.g., form submissions in everyday web browsing — this trade-off may be worth considering to improve the day-to-day browsing experiences of PVIs. In short, our findings help extend the state-of-the-art in usable audio CAPTCHAs and should strengthen the foundation for researchers and practitioners to explore the design space of more usable audio CAPTCHAs.

Acknowledgments

We thank all participants who were generous with their time, plus all the organizational staff members for their support in distributing our recruitment message. We could not have reached so far around the world had it not been for this assistance. We are also indebted to Dr. Yang Wang at the University of Illinois at Urbana Champaign for his extensive critiques during the editing process, Dr. Elissa Redmiles for her thoughtful suggestions on refining our study design, and the support of Youngwook Do from the Georgia Tech SPUD Lab. This work was generously supported by seed funds from Georgia Tech's School of Interactive Computing.

References

[1] Big sound bank. https://bigsoundbank.com/, Accessed: 2019-09-19.

[2] Zapsplat. https://www.zapsplat.com/sound-effect-categories, Accessed: 2019-09-19.

[3] K. Aiswarya and K. S. Kuppusamy. A study of audio captcha and their limitations. 2015.

[4] Sacha Brostoff Angela Sasse and Dirk Weirich. Transforming the 'weakest link' — a human/computer interaction approach to usable and effective security. *BT Technology Journal*, 19:122–130, 08 2001.

[5] Jeffrey P. Bigham and Anna C. Cavender. Evaluating existing audio captchas and an interface optimized for non-visual use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1829–1838, New York, NY, USA, 2009. ACM.

[6] Stacey Burling. Captcha: The story behind those squiggly computer letters, 2012. https://phys.org/news/2012-06-captcha-story-squiggly-letters.html.

[7] Elie Bursztein and Steven Bethard. Decaptcha: breaking 75% of ebay audio captchas. 01 2009.

[8] Artemios Voyiatzis Christos Fidas. On users' preference on localized vs. latin-based captcha challenges. pages 358–365, 09 2013.

[9] HT Correspondent. Number of blind to come down by 4m as india set to change blindness definition, 2017. https://tinyurl.com/y8se7h4d.

[10] Celine Fabry John C. Mitchell Elie Bursztein, Steven Bethard and Dan Jurafsky. How good are humans at solving captchas? a large scale evaluation. In *2010 IEEE Symposium on Security and Privacy*, pages 399–413, New York, NY, USA, 2010. IEEE.

[11] Bowei Du Melissa Densmore Matthew Kam Sergiu Nedevschi Joyojeet Pal Rabin Patra Sonesh Surana Eric Brewer, Michael Demmer and Kevin Fall. The case for technology in developing regions. *Computer*, 38:25–38, 06 2005.

[12] Google. Audioset. https://research.google.com/audioset/, Accessed: 2019-09-21.

[13] Jinjuan Feng Graig Sauer, Harry Hochheiser and Jonathan Lazar. Towards a universally usable captcha, 2008. https://cups.cs.cmu.edu/soups/2008/SOAPS/sauer.pdf.

[14] Dan Yao Xiyang Liu Haichang Gao, Honggang Liu and Uwe Aickelin. An audio captcha to distinguish humans from computers. *SSRN Electronic Journal*, 01 2010.

[15] Dorothea Kolossa Hendrik Meutzner, Santosh Gupta. Constructing secure audio captchas by exploiting differences between humans and machines. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2335–2338, New York, NY, USA, 2015. ACM.

[16] Sean Hyde Jennifer Tam, Jirí Simsa and Luis Ahn. Breaking audio captchas. pages 1625–1632, 01 2008.

[17] Jinjuan Heidi Feng Jonathan Holman, Jonathan Lazar and John D'Arcy. Developing usable captchas for blind users. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pages 245–246, New York, NY, USA, 2007. ACM.

[18] Tim Brooks Genna Melamed Brian Wentz Jon Holman Abiodun Olalere Jonathan Lazar, Jinjuan Feng and Nnanna Ekedebe. The soundsright captcha: An improved approach to audio human interaction proofs for blind users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2267–2276, New York, NY, USA, 2012. ACM.

[19] George Hughey Dave Levin Kevin Bock, Daven Patel. uncaptcha: A low-resource defeat of recaptcha's audio challenge, 2017. https://uncaptcha.cs.umd.edu/papers/uncaptcha_woot17.pdf.

[20] Sunny Behal Kiranjot Kaur. Captcha and its techniques: A review. *International Journal of Computer Science and Information Technologies,*, 5, 01 2014.

[21] Patrice Simard Kumar Chellapilla, Kevin Larson and Mary Czerwinski. Designing human friendly human interaction proofs (hips). In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 711–720, New York, NY, USA, 2005. ACM.

[22] Manuel Blum Luis von Ahn and John Langford. Telling humans and computers apart (automatically). *Communications of the ACM*, 47, 05 2002.

[23] Panagiotis Germanakos Marios Belk, Christos Fidas and George Samaras. Do human cognitive differences in information processing affect preference and performance of captcha? *International Journal of Human-Computer Studies*, 84, 07 2015.

[24] Jakob Nielsen. Usability 101: Introduction to usability, 2012. https://www.nngroup.com/articles/usability-101-introduction-to-usability/.

[25] A. Bujari Ombretta Gaggi, Giacomo Quadrio. Accessibility for the visually impaired: State of the art and open issues. pages 1–6, 01 2019.

[26] T. Braithwaite M.V. Cicinelli A. Das J.B. Jonas Y. Zheng R.R. Bourne, S.R. Flaxman. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: A systematic review and meta-analysis, 2017. https://doi.org/10.1016/S2214-109X(17)30293-0.

[27] Varshini Sampath Saumya Solanki, Gautam Krishnan and Jason Polakis. In (cyber)space bots can hear you speak: Breaking audio captchas using ots speech recognition. pages 69–80, 11 2017.

[28] Andy Schlaikjer. A dual-use speech captcha: Aiding visually impaired web users while providing transcriptions of audio streams. 01 2007.

[29] Heemany Shekhar. Breaking audio captcha using machine learning/deep learning and related defense mechanism. page 741, 2019.

[30] Takuma Otsuka Shotaro Sano and Hiroshi Okuno. Solving google's continuous audio captcha with hmm-based automatic speech recognition. 11 2013.

[31] Yannis Soupionis and Dimitris Gritzalis. Audio captcha: Existing solutions assessment and a new implementation for voip telephony. *Computers Security*, 29:603–618, 07 2010.

[32] Noshina Tariq and Farrukh Khan. *Match-the-Sound CAPTCHA*, pages 803–808. 07 2018.

[33] Stephen Goglin Travis Schluessler and Erik Johnson. Is a bot at control? detecting input data attacks. In *Proceedings of the 6th ACM SIGCOMM workshop on Network and system support for games*, pages 1–6, New York, NY, USA, 2007. ACM.

[34] WebAIM. Screen reader user survey 7 results, 2017. https://webaim.org/projects/screenreadersurvey7.

[35] WHO. Global data on visual impairments 2010, 2010. https://www.who.int/blindness/GLOBALDATAFINALforweb.pdf.

[36] WHO. Visual impairment and blindness 2010, 2010. https://www.who.int/blindness/data_maps/VIFACTSHEETGLODAT2012_2.pdf.

[37] Jeff Yan and Ahmad Ahmad. Usability of captchas or usability issues in captcha design. pages 44–52, 01 2008.

[38] Zoom. Video conferencing, web conferencing, webinars, screen sharing. https://zoom.us/, Accessed: 2019-09-19.

A Appendix: Screenshot of Experimental Test-Bed and Questionnaires

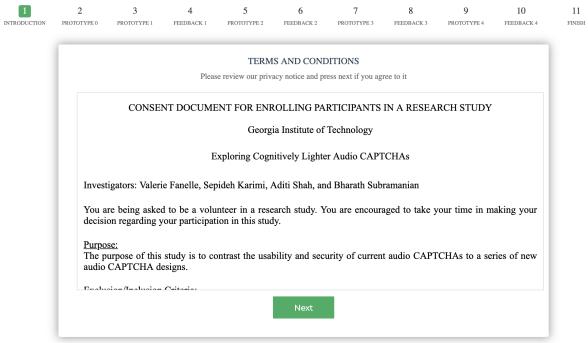


Figure 3: The landing page for the study website on which users clicked the 'Next' button if they decided to consent to the study.

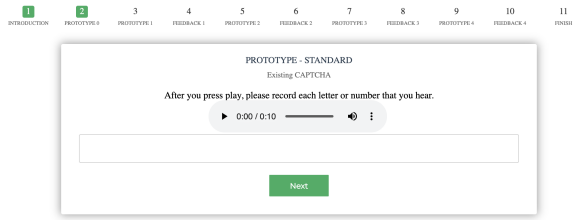


Figure 4: We used a webpage format as a test-bed for each of our CAPTCHA designs. Participants could replay the CAPTCHA as many times as they needed until they felt confident of their answers.

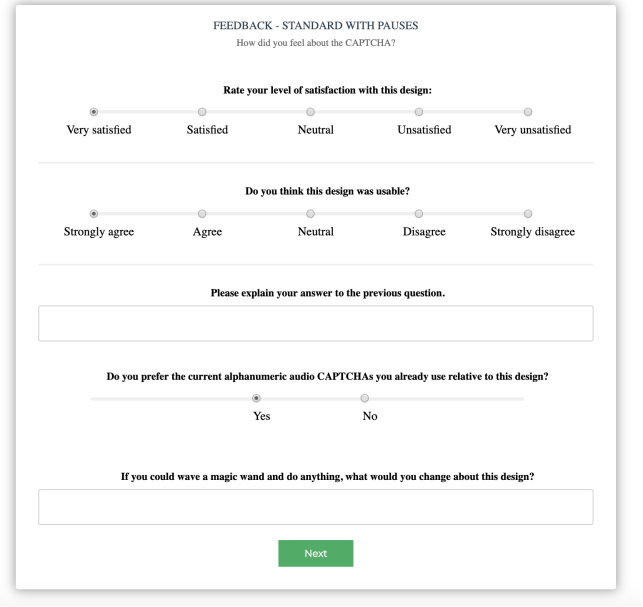


Figure 5: We remotely obtained participant feedback on the usability and satisfaction of each design via the above feedback form.

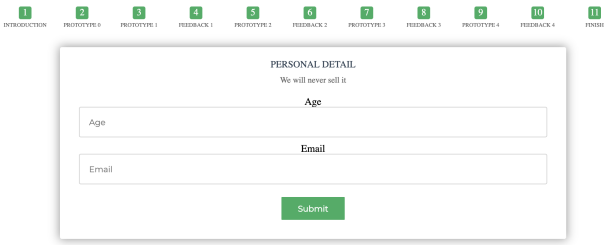


Figure 6: After all 13 CAPTCHAs for that session were presented, participants were asked to enter their email and age for demographics and transcription purposes.

Usable Sexurity: Studying People's Concerns and Strategies When Sexting

Christine Geeng
Paul G. Allen School of Computer Science & Engineering
University of Washington

Jevan Hutson
School of Law
University of Washington

Franziska Roesner
Paul G. Allen School of Computer Science & Engineering
University of Washington

Abstract

Sexting, while becoming commonplace in the age of mobile phones, is still not well-studied outside of the context of youth or risk prevention. Taking the perspective that sexting is a normal intimacy-building behavior rather than a deviant practice that should be prevented, this work studies the computer security and privacy mental models and technology use of adults who sext. We conducted an online survey of adults who have sexted ($N = 247$), asking participants about which platforms they use to sext, their general sexting behaviors, what security and privacy concerns they have around sexting, and how they mitigate those concerns. We find, for example, significant concerns around sexts participants send "getting out" and being misused, as well as concerns around receiving unsolicited sexts. We also find that while participants use some technical strategies (e.g., using platforms with disappearing messages), they commonly rely on non-technical strategies like trust in their partner to mitigate concerns. We ground our findings in Citron's legal framework of sexual privacy to support individual autonomy, intimacy, and equality, and we make design recommendations for communication platforms to support usable security and privacy for sexting.

1 Introduction

Sending or receiving nude or semi-nude photos and other media (often called "sexting") is a common sexual behavior for adults in the United States, as technology has allowed greater ease of sharing and accessing sexual media [34]. The rise of ubiquitous mobile devices has supported sexting practices — for example, the major social media platform Snapchat was founded based in part with the idea of making it less risky to sext [28], and many other apps exist that aim to provide privacy protections for interpersonal communications.

While sexting has become common, much of psychology work has characterized it as deviant behavior [64]. This perspective leaves out the opportunity to understand and design for sexting as a normal human behavior. Sexting *does* carry real risks — not due to an inherent immorality, but because sexts can be abused. For example, 1 in 25 Americans has been a victim of "revenge porn" (in which sexual imagery of someone is distributed without their permission) or threats of it. Women below 30 are more likely to be targeted than men or older women, and queer individuals are more likely to be targeted than heterosexual individuals [39]. But sexting also provides significant benefits, enabling self-expression and intimacy building in consensual relationships [15].

Thus, in this work, we engage with sexting as a normal adult practice and study people's technology-related concerns and practices related to sexting. Our work builds on and complements a call from the law and policy community to "conceptualize sexual privacy clearly and to commit to protecting it explicitly" [15]. We aim to understand how people navigate and conceptualize issues of privacy and security, and begin to articulate a framework for future research and development in usable security and privacy for sexting. We ask the following research questions:

1. *RQ1: Practices and Experiences.* What are people's general practices and experiences with technology-mediated sexting?
2. *RQ2: Concerns.* What are people's computer security concerns and threat models related to technology-mediated sexting?
3. *RQ3. Mitigations.* What are people's (technical or non-technical) mitigation strategies for managing these concerns?

We present results from an anonymous online survey of adults who have sexted at some point in their lives ($N = 247$). We asked questions about with whom, how often, and on what platform people sext, as well as about their concerns and mitigation practices. We find significant concerns around both

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

sending sexts (e.g., that they will "get out" somehow, be misused in specific ways, or be seen accidentally by the wrong person) as well as receiving them (e.g., receiving unsolicited sexts or shoulder-surfing of solicited sexts). We further find that people rely heavily on non-technical strategies for mitigating these risks, including conscientiously establishing trust and social contracts with their sexting partners — suggesting a potential role for sexting platforms in helping scaffold or support these social contracts.

We close by making design recommendations and identifying opportunities for future research, grounded in Citron's legal sexual privacy framework to support individual autonomy, intimacy, and equality around sexting. Our work lays a foundation for considering and supporting security and privacy for sexting as a normal behavior among technology-using adults.

2 Motivation and Related Work

We define "sexting" as the technology-mediated interpersonal exchange of sexual media, including flirtatious or sexually explicit text or emojis, and nude or semi-nude photos and videos. In this section, we survey prior scholarship on sexting and usable security, identifying gaps that motivate our work.

2.1 Scholarship on Sexting

Sexting has become a common practice: Herbenick et al. found that 27% of adult women and 24% of adult men in the United States sent nude or semi-nude photos of themselves to someone [34], and Madigan et al. found that 14.8% and 27.4% of teens send and receive sexts, respectively [42].

Academic Framing. Despite the pervasiveness of sexting, much of the academic work has focused on youth and young adults [22, 33, 64]. Furthermore, early literature on sexting treated it as a high-risk, deviant behavior, rather than an important part of adult social life that is just as normal as not sexting [21, 37]. Döring calls for an approach to sexting that acknowledges both "vulnerability and sexual agency" [21].

Research on youth has pointed out important concerns, such as adolescents feeling pressured to sext due to the erroneous belief that "everyone is doing it" [41, 77]. While youth and adult sexting both share some of the same risks and questions, it is important to also study the adult risk landscape. Our work seeks to better understand how adults (not just students) sext, from a perspective that views sexting as normal (and even important) intimate communication. Research scholarship on consensual sexting behavior has highlighted its positive role in relationship satisfaction [11, 20, 72], and that the affordances of sexting may lead to stronger sexual norms around explicit communication and consent [32]. Other work highlights potential issues that can arise with sexting, such as if the content is distributed without authorization, or if it occurs under pressure or as the result of coercion [5, 13, 73].

Mitigation Strategies. Our study expands on prior work on sexting concerns and mitigations. Sex education researchers have studied how to teach youth about safe sexting and navigating consent, coercion, and digital footprints [36, 61]. Renfrow et al. found that college students minimized perceived risks through strategies around controlling sexting content, including 'keeping it fun' (avoiding more vulgar terms), limiting explicitness, and creating plausible deniability [64]. Amundsen [5] conducted qualitative interviews with women about the role trust has as a mitigation strategy for non-consensual sext sharing, and how the responsibility of mitigating risk may disproportionately fall upon victims, which are themes reflected in other work [74]. These prior studies do not deeply consider the role of technology (which can both create new concerns and support new mitigations) in sexting. In this work, we take a computer security point of view.

Beyond academic research, there are numerous applications that aim to (or are commonly used to) support sexting, as well as online guides for how to sext "securely". For example, Vice [45] lists guidelines including: get consent and set expectations, check for identifying details in photos, turn off services that automatically backup photos, wipe photos of EXIF metadata, and choose a communication app based on one's concerns. In terms of applications, Snapchat is popular with disappearing messages (which disappear quickly from the user interface, and are deleted from Snapchat servers within 30 days [1]), among other features such as screenshot notifications and a password-protected photo album. Other sexting guides list encrypted messaging platforms such as Signal, Whatsapp, and Facebook Secret Messenger. Less well-known examples include Kaboom (which allows users to send a disappearing message through a link, so that the receiving party can see the message without installing Kaboom) and Confide (a messaging app that has disappearing encrypted messages and screenshot notifications).

2.2 Sexual Privacy Framework

Legal scholar Danielle Citron argues that sexual privacy — "the social norms (behaviors, expectations, and decisions) that govern access to, and information about, individuals' intimate lives" — is a privacy value of the highest order because it is central to sexual agency, intimacy and equality: "[w]e are free only insofar as we can manage the boundaries around our bodies and intimate activities" [15]. Citron outlines how sexual privacy is foundational to (1) securing autonomy, (2) enabling intimacy, and (3) protecting equality.

While other privacy frameworks exist, such as contextual integrity [56], we find sexual privacy to be most appropriate for framing our study, as it forefronts the existence of unequal vulnerabilities (something that norm-based privacy theory does not do [49]). We briefly summarize these properties here, helping to motivate why protecting sexual privacy is crucial. We then place our results and recommendations in terms of

this framework in Section 5.

Securing Autonomy. Citron and others argue that sexual privacy is fundamental to the exercise of human agency and autonomy [15]; it is what allows individuals to manage the boundaries of their bodies and their intimate lives [3,48]. This autonomy, in turn, is viewed as fundamental to individual self-development and identity formation (who we are and who we might be in the future) [8,48,52,62]

Enabling Intimacy. Scholars have also argued that sexual (and other) privacy is critical to cultivating interpersonal intimacy, affection, and trust [15,26,65]. Indeed, research demonstrates that sexual privacy is key to the formation, maintenance, and growth of intimate relationships [4, 10, 27, 63, 76]. Intimacy is associated with important consequences for individual personal welfare, including heath, well-being, community attachment and sexual sociality [30]; research has further established a positive relationship between sexual activity and such outcomes as lifespan [59] and overall happiness [9].

Protecting Equality. Sexual privacy also implicates issues of equality, justice, and power [18,66], as women, sexual minorities, and nonwhites continue to bear the disproportionate burden of sexual privacy harms, such as surveillance, harassment, and abuse [14, 15, 67, 70, 71]. More broadly, political theorists argue that intimacy is a matter of justice, as access to access to intimacy is critical to accessing primary social goods such as wealth and self-respect [6, 15]. Scholars also underscore how the intimate sphere, both digital and non, is inextricably tied to relations of power [7, 8, 14, 23, 35, 58] and has historically been a key determinant of social and economic welfare [6, 23, 31].

2.3 Scholarship in Usable Security & Privacy

Finally, our work is situated in the broader space of usable security and privacy research, particularly studies on how people navigate sharing information in interpersonal relationships, such as account and device sharing in relationships [46, 60], online dating [16], social media [43], and human trafficking [12]. Freed et. al and others have studied how technology and information shared during the beginning of a trusting relationship gets abused when that turns into intimate partner violence [24, 25, 47]. These various settings surface both overlapping lessons (e.g., how changes in relationships over time lead to different security or privacy vulnerabilities [40], such as a parent giving a child more privacy as they grow older [29]) as well as distinct challenges for different populations. At the highest level, these works reflect that threat modeling and design must follow a socio-technical approach, considering the properties of technology, how people use it, how people interact with each other, and societal expectations for such behavior.

3 Methods

We designed an anonymous online survey, using both close- and open-ended responses, to investigate the technology-related sexting behaviors and concerns of adults.

3.1 Ethical Considerations

Our study was reviewed and determined exempt by our institution's IRB. Given the potentially sensitive nature of our topic of study, we did not collect any identifying information from participants. Only participants who indicated that they were 18 years old or older were able to complete the survey. The opening paragraph to the survey emphasized that sexting is a common practice and that we as researchers are not taking a judgemental stance on it. The majority of questions were optional, including the choice "Prefer not to say". Participants could opt out of allowing their (anonymous) quotes from free-response answers to be used in this publication. The quotes we include in our results were chosen to illustrate patterns of behavior, rather than any individual's unique or potentially identifiable situation.

3.2 Recruitment

To recruit participants, we posted links to the survey on our personal Facebook, Twitter, and TikTok accounts. To widen recruitment beyond our personal networks, we distributed our survey via physical fliers in a major U.S. city and posted online to Reddit "subreddits" (e.g., /r/sex). Because prior literature often fails to capture the nuances of sexting among sexual and gender minority communities (e.g., [19]), we also recruited specifically from queer social media groups and apps; this may explain why we sample more non-straight identifying participants than reflected in the United States population [55]. Four $20 gift cards were provided to random participants.

Upon beginning the anonymous online survey and after indicating their informed consent, participants were directed to questions that established whether they were at least 18 years old and whether they have ever engaged in sexting. If participants were under the age of 18 or indicated never having participated in sexting, they were dismissed from further data collection and analysis. We excluded people who had not sexted before because our research questions focus on existing behaviors.

3.3 Procedures

We chose an online survey method to allow us to capture a broad population of people who sext [54]. We were especially interested in engaging with individuals over 18 because existing literature skews heavily towards youth and adolescent sexting practices [64]. To ensure that our research reflected

Gender		Sexual Orientation		Age Range		Intimacy Status	
Male	30.36%	Straight	46.96%	18-24	52.23%	Single	25.10%
Female	61.54%	Questioning	2.02%	25-34	37.25%	Dating	35.22%
Non-binary	9.72%	Gay	8.10%	35-44	6.07%	Engaged	3.24%
Prefer not to say	0.81%	Lesbian	7.29%	45-54	3.24%	Married	10.53%
Self-describe:	1.21%	Bi/Pan	23.08%	55-64	1.21%	Divorced	2.02%
		Queer	6.48%			Friend-with-benefits	16.60%
Trans	5.67%	Asexual	1.21%			Casual sex	11.74%
Cis	89.47%	Prefer not to say	1.62%			Casual dating	16.19%
Questioning	3.24%	Self-describe:	3.24%			Prefer not to say	1.62%

Table 1: Demographics of 247 survey participants included in our analysis. Gender and intimate status categories were not mutually exclusive, so participants could use multiple labels to describe themselves.

Relationship Practice	
Monogamous	77.33%
Polyamorous	16.19%
Prefer not to say	6.48%

Table 2: Participants' relationship practice ($N = 247$).

the full constellation of gender identity, we followed Klaus et al.'s HCI Guidelines for Gender Equity and Inclusivity [68] in designing our demographics questions, which were asked at the end of the survey.

We asked three classes of questions in our survey, corresponding to our three research questions: (1) questions about general technology-enabled sexting practices; (2) questions about sexting-related concerns; and (3) questions about mitigation strategies to manage those concerns. The full survey instrument can be found in Appendix A.

For questions related to concerns around sending or receiving sexts, we first asked an open-ended free response version, followed by a multiple-choice version. The goal was to first elicit concerns naturally, without priming participants about what they could or should be concerned about.The multiple choice options were based on our own hypotheses as well as informed by an existing survey related to concerns among university students around sexting [64].

For participants who completed the survey, they could opt-in to submitting their email to be entered into a raffle for a $20 gift card. Their email was not linked to their survey data.

3.4 Data Analysis

We collected a total of 330 finished surveys, 249 with respondents who selected that they were at least 18 years old and had sexted before. Three researchers went through the open-ended data to look for disingenuous (e.g., joke) answers; we removed 2 respondents and completed our data analysis based on the remaining 247 responses.

For each open-ended free-response question, three researchers independently inductively coded those questions before discussing and agreeing on a set of qualitative codebooks (a separate codebook for each open-ended question). For questions where we asked an open-ended question followed by a similar multiple choice question, we incorporated the multiple choice options into our codebooks where appropriate. With the finalized codebooks, two researchers independently recoded the open-ended responses. All open-ended responses could have been coded with multiple labels.

Following McDonald et al.'s guidelines on when to seek coding agreement [50], for open-ended questions with simple responses, we used only one coder. For open-ended questions with more complex responses that we discuss quantitatively (concerns about sending sexts, and managing sending concerns), two researchers double-coded all responses. We calculated Cohen's κ for inter-coder reliability, given that we had two coders and nominal data [51]. For concerns about sending sexts, we had a κ of "substantial" (0.61–0.80) to "almost perfect agreement" (0.81–1.00) for 91.3% of categories. For managing sending concerns, we had a κ of "substantial" to "almost perfect agreement" for 93.76% of categories. (More details on κ per category can be found in Appendix C). We discussed discrepancies between coders for all codes until we reached a near-consensus.

We compare answers to some multiple choice questions across genders for statistical differences. Since respondents were able to select multiple genders, we evaluate differences with a test of multiple marginal independence, which is calculated using a modified Pearson's Chi-squared statistic and a bootstrapping method to estimate the sampling distribution [38]. We use a significance level of $\alpha = 0.05$. We report Cramer's V for effect size on a scale of 0 to 1 (with associations > 0.10 indicating at least a small effect [17]). We discarded participants who selected "Prefer to self-describe" (2 respondents with different answers) or "Prefer not to say" for their gender, leaving us with three nominal variables (male, female, non-binary). Since we asked if participants were transgender in a separate question, we cannot distinguish that category with this analysis, and our results are limited in that respect.

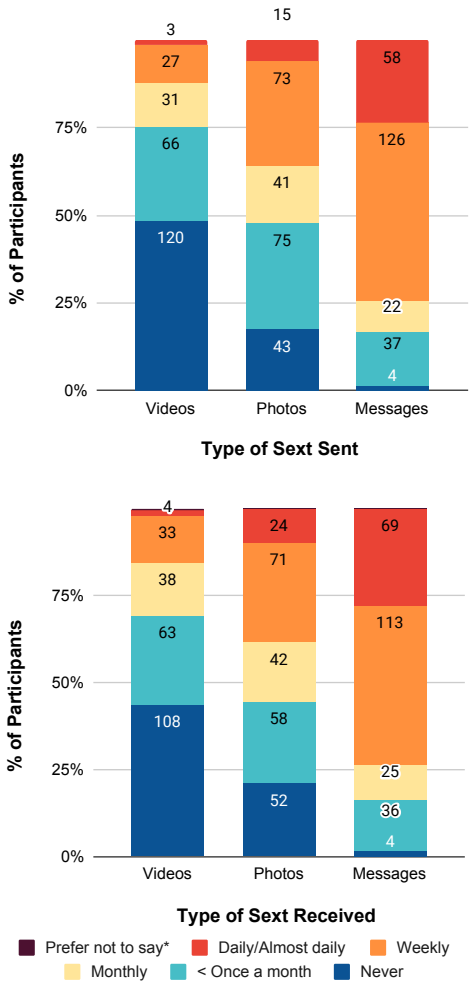


Figure 1: Participants reported how often they sent and received nude or semi-nude videos, photos, and sexual or intimate messages ($N = 247$). *1 person selected "Prefer Not To Say" for how often they received videos.

4 Results

We now turn to our results, based on the 247 valid survey responses from sexting adults, and organized around our three research questions (general practices in Section 4.1, concerns in Section 4.2, and mitigations in Section 4.3).

Table 1 summarizes the demographics of our participants, and Table 2 summarizes the types of relationships they considered while reporting on their sexting behaviors.

4.1 Sexting Practices and Experiences

We begin by considering general sexting practices and experiences, to help provide context for the concerns, mitigations, and design recommendations that follow.

Social Media Platform	**Direct Message**	**Private Post**	**Public Post**	**Story**	**Other**
Facebook/ Messenger	44	3	2	0	4
Twitter	13	6	2	0	1
Instagram	36	7	2	2	1
Snapchat	114	14*	3	1	1
Other	42	4	4	0	5

Table 3: Multiple-choice responses to what social media platforms and communication type that participants use to send/receive photos or videos ($N = 220$; this excludes participants who only send intimate text). *Snapchat has "private stories", not "private posts"; this wording error may have confused participants.

4.1.1 Sexting Frequency

We found that 58.6% (144) of our participants said they currently sext, 33.6% said they have sexted before and may again in the future, and 8.8% said they have sexted before but no longer plan to. Considering sexting medium (i.e., video, image, or text), we found that text-based messages and nude or semi-nude photos were the most common, compared to nude or semi-nude videos (which only half of our sexting participants reported sending or receiving). The results look similar for sending and receiving frequency, suggesting this behavior is reciprocal. Figure 1 breaks this down in detail.

For participants who reported not currently sexting (regardless of whether they plan to in the future), we asked them why they stopped sexting. This question was multiple choice and optional; 106 participants responded. Of those, 45 said they stopped because they were no longer in a relationship with the person they sexted, and 31 said it was because they were no longer in a long-distance relationship. 28 said they were not interested, and 10 said they had had a poor experience. In a free-response follow-up question, out of the 10 people who selected having had a poor experience, 3 said sexting felt awkward, and 1 person said they were scammed.

For people who said they would sext again in the future, the majority marked their reason for stopping as no longer being in a relationship. For the people who said they would not sext again in the future, the majority marked their reason as not interested, with their explanations including "I honestly wasn't super into it", "no longer feels private", and "it felt like I was forcing myself into it".

4.1.2 Device and Platform Usage

Our participants primarily sext on their smartphone devices: 244 use their phone, 73 use their computer, and 16 use their tablet (with some using multiple devices, i.e., these responses are not mutually exclusive).

When asked about how they use social media to send

Other Platforms		
SMS	137	62.27%
Whatsapp	38	17.27%
Grindr	19	8.64%
Tinder	11	5.00%
Telegram	10	4.55%
Signal	6	2.73%
OkCupid	4	1.82%
Hinge	3	1.36%
Kaboom	1	0.45%
Confide	1	0.45%
Other	28	12.73%

Table 4: Multiple-choice responses to what other platforms participants use to send/receive photos or videos ($N = 220$).

sexting-related photos/video, the majority of our participants answered that they use Snapchat direct message (114; see Table 3 for the full breakdown). For other platforms, SMS had the highest usage (for sending photos/video) at 137 participants, possibly because it is a phone's default messaging app. (2 "Other" responses explicitly mentioned iOS Message, which respondents may have also counted as SMS.) "Other" responses included: Tumblr, Whisper, Kik, Skype, Reddit, Discord, Wired, Burner, email, and other dating sites. We do not have data about whether participants used these exclusively for sexting or also for other purposes. Considering sexual and intimate text messages, the distribution of platforms used looks similar to what we see for photos/video in Table 3

Some participants reported using platforms that explicitly include security- and/or sexting- related functionality. For example, Snapchat has disappearing messages, screenshot notifications, and a password-protected photo album. Considering the more obscure platforms mentioned by participants: Burner provides a temporary new phone number that allows for communication while obscuring one's actual phone number, Wire is an end-to-end encrypted communication app, and Kik and Whisper tout themselves as anonymous social media services. We returns to people's uses (or non-uses) of these features in Section 4.3

4.1.3 Storing Sexts

We explicitly asked participants about their practices around storing their own and other people's nude or semi-nude images or videos (which we refer to by the shorthand "nudes" below). We note that these responses must be interpreted under the risk of social desirability bias: participants may have underreported socially undesirable behaviors, such as storing or sharing sexts without consent.

With respect to one's *own* nudes, we found that out of 247 responses, 130 (52.6%) said they stored nude or semi-nude photos or videos of themselves, 114 said they did not, and 3 said preferred not to answer. We find that storing nudes

Reasons for Saving Received Nudes	
To use/look at later	80
Insurance	5
Saving is app default	5
Partner is away	4
Asked by sender to save	4
To share with friends/others	3
To be able to find more easily	2
Mutual saving	2

Table 5: Coded counts of participant open-ended responses to why they saved sexts sent to them ($N = 145$).

received from *other* people is even more common: among 234 participants who reported having received a sext and answered this question, 145 (62.0%) stated they have saved nude photographs or videos they received, 85 said they had not saved any, and 4 preferred not to answer.

For participants who store others' nudes, we asked additional questions about why and how they are saved (which may help explain why more participants' save other people's nudes than their own). In response to an open-ended question about why, most participants mentioned saving others' nudes for later use (e.g., nostalgia, to masturbate). Several participants mentioned saving to share media with friends, but none of those responses explicitly mentioned getting consent from the sender to share. Two people noted they saved content with no intention of sharing it, and 9 people thought it was assumed they would save a partner's photos. We note that 24 people *did* mention that they saved nudes with permission from sender (and 4 people mentioned they were even asked by the sender to save). In this case, this number is a lower bound on how many participants received explicit consent to save nudes, since we did not specifically ask this in the question.

In the words of one participant:

> *"I saved them because my girlfriend took the time to take a nice photo, just for me, and she's given me the OK to save them. When I miss her, it helps to look through a medley of sexual and non-sexual photos of her."* – Male, straight, 18-24

In another question, we asked explicitly about whether senders knew that nudes had been saved by the participant. Out of 145 responses, 110 said that senders knew, 24 said some of the senders know, 7 said they do not know, and 4 preferred not to say. When we consider *how* nudes were saved, we note that only a small number of participants reported methods that explicitly aim to avoid knowledge by the recipient: 56% of participants directly stored to device, 37% took a screenshot, 5% (10) took a photo (presumably to circumvent screenshotting notifications, i.e., taking advantage of the "analog hole"), and 3% selected "Other": one respondent mentioned using screenshotting apps to prevent Snapchat screenshot notifications (e.g., Private Screenshots), and an-

Plans for Received Nudes that are Saved		
Save them until asked to delete them	56	38.62%
Save them indefinitely	48	33.10%
Save them for some amount of time	27	18.62%
Other	13	8.97%
Prefer not to say	1	0.69%

Table 6: Single-choice responses from participants who said they save nudes sent to them for what they plan to do with them ($N = 145$).

other mentioned saving to a private folder.

For participants who save other people's nudes, 75 said they store on their device's photo storage (e.g., "Camera Roll"), 49 said a separate on-device album, 32 said a specific secret-keeping app (e.g., Snapchat's "My Eyes Only" folder, Vault, or an encrypted folder on a desktop/laptop computer), and 17 said online (e.g., Google Photos, Dropbox; some apps mentioned in the prior option are also online storage). Only 5 participants noted that their reason for saving is because this is the app's default behavior, but we suspect that this is an under-count, since many participants mentioned using SMS and other apps that automatically save content by default.

Out of the 145 participants who said they save other people's nudes (see Figure 6), most said they would save until asked to delete. For respondents who selected "Other" and "Save them for some amount of time", many explained that they might save until the end of the relationship, or save for time periods ranging from a week to years. One person mentioned they assume the photos would be automatically deleted when iPhone's cache is full, and another person mentioned that they have not thought that far ahead.

Overall, our results suggest common practice involves saving received sexts — not typically for nefarious purposes, and often (reportedly) with the consent and knowledge of the sender. Given these legitimate uses, a recommendation or platform-enforced policy of not saving received nudes would often be impractical and overly restrictive.

4.1.4 Sharing Sexts

We also asked explicitly about whether participants shared received sexts with others. Out of 247 responses, 32 said yes, 213 said no, and 2 said they prefer not to say. For yes answers, many people mentioned showing to friends or partners, and some mentioned sharing unsolicited photos for support and mockery. Some participants took measures to protect the privacy of the sext's creator, including getting explicit permission to share or anonymizing or otherwise editing the sexts — for instance, cropping and censoring identifying information.

Concerns Sending Sexts		
Sexts get around to other people	93	75.00%
Sexts used as blackmail	83	66.94%
Receiver's devices will get hacked and the content will get out	78	62.90%
Receiver will intentionally share content with others	72	58.06%
Regret	58	46.77%
Sexting causes ridicule from others	40	32.26%
Not sure I sent it to the right person	38	30.65%
Bullying or harassment from others	29	23.39%
Unwanted attention	28	22.58%
Legal liability	25	20.16%
Sexting makes people feel led on, used, or misunderstood	24	19.35%
Damages relationships	19	15.32%
Conflicts at work	19	15.32%
Unwanted sexual contact	15	12.10%
Engagement with law enforcement	10	8.06%
Other	9	7.26%

Table 7: Multiple-choice responses to what concerns participants have when sending sexts ($N = 124$). This excludes 122 participants who send sexts but did not indicate that they have concerns around sending sexts.

4.2 Concerns around Sexting

We now turn to our participants' concerns around sending and receiving sexts. For both types of concerns, we first asked an open-ended question to elicit natural, non-primed concerns, followed by a multiple-choice question that allows us to evaluate the frequency of named concerns.

4.2.1 Concerns Around Sending Sexts

Table 7 shows participant concerns about sending sexts in response to our multiple choice question. The following concerns were most prevalent: "Sexts get around to other people" (93) and "Sexts used as blackmail" (83). Similarly, the most-used codes for the open-ended concern question (which, again, was asked before the multiple choice) was "Sexts will get out" (38) and "Shared/shown to others" (24). While some participants indicated only generic concern, others indicated a more specific threat model, specifying adversary (e.g., partner or platform) or consequence (e.g., impact on career or possible negative judgement). For example, one participant wrote:

> *"We live in a society of prudes — I worry that things will leak and get out there and people will judge me for what I have shared with someone under the pretext that it was going to be private."* – Non-binary, asexual, 25-34

Many fewer participants in the open-ended response mentioned "Sexts used as blackmail" (9), "Revenge porn" (8),

Concerned about sexts being used as blackmail?			
	No	**Yes**	**Total**
Male	57	17	74
Female	93	59	152
Non-Binary	15	9	24

Table 8: $N = 243$. N here and for Table 9 includes all participants who said they send sexts and who selected at least Male, Female, or Non-binary for their gender.

Concerned about sexts causing ridicule from others?			
	No	**Yes**	**Total**
Male	69	5	74
Female	123	29	152
Non-Binary	20	4	24

Table 9: $N = 243$.

or "Misuse" (11) as concerns, compared to the 83 responses to "Sexts used as blackmail" in the multiple-choice question; there is a similar discrepancy with the hacking concern. Participants may not have considered the multiple-choice concerns without being prompted to, or found the multiple-choice questions easier and/or less effort to answer.

Our open-ended question surfaced concerns we had not anticipated in our multiple choice options. The language "revenge porn", "blackmail", and "misuse" was used rather than "Bullying or harassment" (from the multiple choice question). Some responses also gave insights into participants' thoughts about potential adversaries: 10 respondents noted they were not concerned because they trusted their partner, and 3 noted they were not concerned because they trusted the app. On the flip-side, 6 respondents mentioned not trusting the platform companies, and 3 respondents were concerned about bugs or vulnerabilities in the app. Other concerns not mentioned in the multiple choice options included deanonymization (7), recipient will save sexts (6), insecure network or cloud (2), photos will be modified (1), and images will be used to impersonate sender (1). Referring to both saving sexts and deanonymization concerns, one participant wrote:

> *"People will save the photos. Specifically photos of my face and body together."* – Male, gay, 18-24

In the open-ended questions, many responses were vaguely worded and did not specify the person or platform that somehow distributes or leaks their content. Such responses could be a reflection of vague or broad threat models in the mind of the participant, or of survey fatigue and the limitation of not being able to follow up for elaboration. To the extent that these responses suggest genuinely adversary-less threat models, they reflect Venema et al.'s findings, in which the responsibility of people who share explicit photos without consent is invisible in how the risks are described (e.g., "they [i.e., the photos] spread" or using the passive voice) [74].

Concerns Receiving Sexts		
Receiving unsolicited content	43	66.15%
Shoulder surfing	34	52.31%
My device will get hacked and their content will get out	29	44.62%
Not really the person I think it is	10	15.38%
Other	6	9.23%

Table 10: Multiple-choice responses to what concerns participants have when receiving sexts ($N = 65$). This excludes 185 participants who receive sexts but did not indicate that they have concerns around receiving sexts.

Gender Differences for Sending Concerns. Men were significantly less likely to be concerned about blackmail ($p = 0.02$, $V = 0.15$), with 22.97% of men, 38.82% of women, and 37.50% of non-binary individuals selecting being concerned (Table 8). Men were also significantly less likely to be concerned about ridicule ($p = 0.04$, $V = 0.15$), with 6.76% of men, 19.08% of women, and 16.67% of non-binary individuals selecting being concerned (Table 9). We did not see significant gender differences for other sending-related concerns.

4.2.2 Concerns About Receiving Sexts

Table 10 breaks down participant concerns about receiving sexts in response to our multiple-choice question. The greatest concern was over unsolicited sexts — a concern that is well-founded, given that out of 247 responses, 56% of people (138) said they have previously received unsolicited sexts.

Another major concern was shoulder-surfing, a concern for both receiving and sending sexts. One participant wrote:

> *"I want a warning before. [I] do not want to open a snap with a nude and have my grandmother sitting next to me. [I] must have warning in advanced."* – Female, lesbian, 18-24

Again, our open-ended question surfaced additional concerns, including the sender may escalate behavior/harassment (3), receiving a sext in an inappropriate context (2), being triggered(1), future regrets (1), receiving illegal material (1), and false accusations (1). There were 10 multiple-choice responses to concern over sender authenticity (i.e., being sure about the identify of the sender), versus only 1 response in the open-ended question.

Another concern was feeling forced to reciprocate the sext (4), i.e., forced to send back a sext or engage in other related behavior. For example:

> *"I am concerned that by me receiving sexts, it gives off the impression that I am open to any sexual activity/interaction with the other party."* – Female, straight, 18-24

Have you received an unsolicited sext?			
	No	**Yes**	**Total**
Male	42	33	75
Female	55	94	149
Non-Binary	7	16	23

Table 11: $N = 243$. N includes all participants who selected at least Male, Female, or Non-binary for their gender.

> *"I often feel coerced into responding via reciprocation and if I don't then the person will be angry."* – Female, bi/pan, 24-34

Gender Differences for Unsolicited Sexts. Women and non-binary individuals were significantly more likely to receive unsolicited sexts ($p = 0.005$, $V = 0.19$) and be concerned about that ($p = 0.04$, $V = 0.15$). 63% of women, 69.5% of non-binary people, and 44% of men have received an unsolicited sext (Table 11), and 21.85% of 151 women, 20.83% of 24 non-binary people, and 9.33% of 75 men indicated that they were concerned about this.

Women were also significantly more likely to be concerned about shoulder-surfing ($p = 0.02$, $V = 0.15$), with 17.88% of 151 women, 6.67% of 75 men, and 8.33% of 24 non-binary individuals being concerned. These comparisons can be viewed in table form in Appendix B. This result echoes the earlier finding that women are more likely to be concerned about negative judgement (ridicule) as a consequence of sending sexts. We did not see significant gender differences for other receiving-related concerns.

4.3 Mitigation Strategies

Finally, this section reports on participants' mitigation strategies for the concerns mentioned above, again elicited via both open-ended and free-response questions. We observed that participants mentioned both technical as well as significant non-technical mitigations strategies.

4.3.1 Technical Strategies

In both the open-ended question (34) and in the multiple choice question (57), participants mentioned that they manage concerns by picking a platform with specific features they want. The most common featured mentioned (23 in open-ended) was disappearing messages. (The most-mentioned disappearing message app was Snapchat, consistent with responses about platforms used for sexting.) Another feature often mentioned (and also supported by Snapchat) is notifications when the recipient takes a screenshot of a sent message or image. While these UI-based features may be sufficient to enforce privacy in most circumstances, we note that this mitigation feature alone would not be sufficient if someone is concerned about a receiver sharing supposedly ephemeral

Concern Management Behavior		
Only sexting with people you trust	110	79.14%
Using disappearing messages (e.g., Snapchat, Instagram stories)	65	46.76%
Prior talks to set rules/boundaries	59	42.45%
Choose app with features you want	57	41.01%
Limiting how explicit the sext is	53	38.13%
Ensuring plausible deniability e.g. not including identifying marks in photo	51	36.69%
Password-protect or encrypt sexts	31	22.30%
Other	7	5.04%
No strategies to manage my concerns	5	3.60%

Table 12: Multiple-choice responses to what kind of actions people take to manage their concerns, both around sending and receiving sexts ($N = 140$).

sexts — we note that there exist screenshot apps to circumvent notifications, and recall that 10 of our participants said they take photos to save nudes rather than screenshot them.

Other technical strategies mentioned include: having a passcode that protects access to an image, app, or device (e.g., Snapchat's password-protected "For My Eyes Only" photo album), explicitly deleting messages or media, using encrypted platforms such as Telegram or Signal, and using app or platform settings to ensure that notifications do not make the sexts visible (e.g., to a shoulder-surfer). Some participants explicitly wrote about the threat model they considered when picking a platform. For example, the following participant specifically picks a platform with content deletion because they are concerned about shoulder-surfing:

> *"Telegram has message & chat history delete functionality (and I'm most concerned about messages being *seen* on my device, not on the other person's device - I trust them)."* – Non-binary, pansexual, 18-24

Many participants listed anonymizing sexts as a strategy (29 open-response, 51 in MC) — for example, cropping or blurring faces, or taking photos without identifying features within the frame. Only one respondent out of 20 mentioned being aware of potentially identifying locations in the photo:

> *"Using Signal, not showing face, no identifying marks/locations, no posting public photos that correspond in time/place to explicit photos, no full nudity, only send images that if they would get out I could claim they were art photography or not of me."* – Female, bisexual and queer, 25-34

No participant mentioned being concerned about EXIF data, image metadata that can compromise privacy and that some online safe-sexting guides recommend deleting [45]. (We note that some apps strip EXIF data automatically. For example, Signal strips EXIF data from photos taken within

the app, based on the authors' testing and some anecdotal online sources [2], though not official Signal documentation).

4.3.2 Non-Technical Strategies

The most common mitigation strategy in both the multiple choice (110) and free-response question (54) was only sexting with someone the person knows and trusts. Communicating rules and boundaries (which includes asking the receiver to delete the photos) was also common (25 in free-response), whether the receiver is a long-term partner:

> *"I sext with my partner whom I trust and we had several conversations about sexting before we started (when to delete photos, if we were at risk of revenge porning each other (we're not)), from there we talked about several different platforms and ultimately chose an encrypted platform. It's not completely safe but it's a calculated risk."* – Female, bi/pan, 25-34

or someone the participant does not know as well:

> *"I don't have extensive conversations with the people, but I'll say something like...'if I send this, don't show it to anyone else.' Usually it's a one time comment and when they agree to keep it to themselves, everything is on the table to share. I need to have a minimum level of trust with a person before I'll sext."* – Female, bi/pan, 35-44

Other non-technical strategies included limiting the explicitness of the photo (24 in the free-response) and only sending content the participant would be comfortable appeared in public (3). Three people listed not sexting as their mitigation strategy — i.e., potentially feeling forced to forgo opportunities for building intimacy, as we discuss further in Section 5.

One participant mentioned acquiring collateral as a strategy — i.e., ensuring that the other person sends a photo first that they can save as "insurance", to discourage the other person from ever misusing their images. For example:

> *"I save them because usually the person whom I have sent content to had saved mine in chat (Snapchat) or screenshotted them. So I save them as a precaution/insurance/leverage (if it comes to that)."* – Female, straight, 18-24

Often, participants mentioned using a mix of technical and non-technical strategies. The particularly high prevalence of interpersonal trust and norms as a mitigation strategy points to an opportunity for platform design that can help create and support such norms, which we discuss further in Section 5.

4.3.3 Mitigations for Unsolicited Sexts

The previously discussed mitigation strategies are ones that can be deployed proactively. By contrast, receiving an unsolicited sext requires a reactive strategy. Table 13 summarizes coded open-ended responses to a question about how participants manage unsolicited sexts. Participants used both platform-supported mitigations (such as reporting or blocking senders) as well as ad-hoc, conversation-based approaches. While most participants do not engage with unsolicited sexts, some reported reacting by "trolling" the sender in response:

Handling Unsolicited Sexts	
Block	82
Ignore	56
Ask them to stop/Confront	33
Delete message	15
Report to platform	8
Troll them	4
Change subject (keep talking)	2
Take screenshot	1
Stop using platform	1
Respond positively	1

Table 13: Coded counts of how participants respond to unsolicited sexts ($N = 138$).

> *"Don't pay it much attention. Sometimes mess with them a bit, tell them something like 'It won't let me open it (your pic) it keeps giving me an error message'. They spend ages checking their message settings, trying to resend it, and trying to explain to me how to open it. But usually just ignore and don't respond. Sometimes will have a short conversation and maybe a bit of a laugh about it."* – Female, straight, 35-44

Also, 8 people mentioned that their behavior depends on if the behavior is repeated, and 4 people said their behavior depends on if they know the sender. For example:

> *"It depends. When it someone I have never spoken to, I will usually screen grab it then delete it from the app – share it with someone who is complaining about their wonderful relationship. However, if it is someone I know and have met, it is more upsetting – and I will either message them about what I don't want, or stop talking to them altogether. Usually with a simple 'No.' and block it."* – Non-binary, asexual, 25-34

5 Discussion

Our study captures a rich account of adult privacy and security behaviors around sexting and expands existing knowledge of how individuals navigate sexual privacy in the digital age. Through this work we aim to conceptualize usable sexual privacy and security clearly and commit to protecting it explicitly (echoing Citron [15]). To that end, we adopt and embrace Citron's framework of sexual privacy from legal

scholarship [15] — particularly the values it embodies: security autonomy, enabling intimacy, and protecting equality — to ground both the discussion of our research's implications and directions for future research and development.

5.1 Securing Autonomy

Establishing Sexting Norms & Boundaries. Sexual privacy is, at its core, about the "social norms governing the management of boundaries around intimate life" [15]. Our results surfaced the ad hoc ways in which people articulate and establish norms, expectations, and boundaries around sexting in order to mitigate sexual privacy harms. Critically, our research implicates the role of platforms in scaffolding the articulation and establishment of these norms. Indeed, developing levers to articulate and establish one's preferences with respect to sexting is arguably key to individual sexual agency and autonomy as well as to establishing trust and ensuring accountability. Yet scaffolding sexting norms raises serious questions about both platform and individual responsibility. We consider two ways of scaffolding sexting norms that should be explored in future research and development: product/platform-level policy and user interface design.

First, platforms can establish product policies or community guidelines with respect to sexting. For example, community guidelines could contain language like "Make sure to ask others for consent before screenshotting images or messages in chat feeds" or "Receiving a sext does not obligate a response in kind". In such an approach, the platform plays a central role in articulating norms around sexting (which can in turn reduce user autonomy in some ways).

Second, platforms can leverage user interface design to better enable users to articulate their own preferences and expectations around sexting. For example, platforms could provide fixed disclosure options for users to express particular preferences. The gay dating and hookup platform Grindr already provides an "Accepts NSFW Pics" profile disclosure field where users can select "Never", "Not At First", and "Yes Please." Here, the platform plays a more co-constructive role with respect to sexting preferences and expectations.

On the other hand, platforms could provide open-ended disclosure options (similar to a free-form "About Me" field) for users to express more individualized preferences around sexting. Here, the platform allows norms to be driven by individual users rather than the platform itself. Having the latter free-form space would make more sense on a mixed-use messaging app (e.g., Snapchat) than having a specific disclosure field for sexts, as users may use their messaging profile to contact different people for non-intimate reasons.

Platform Management of Unsolicited Sexts. Our research aims to understand and support individual actions and mitigation strategies towards sexual privacy and safety, but our results must be viewed in a broader societal context where sexting can be both empowering or disempowering [5]. On the negative side, unsolicited sexts are a major concern (and a disproportionate burden for women). While norm-supporting mitigations can help reduce some unsolicited sexts as discussed above, they cannot prevent explicitly malicious behavior. This issue is particularly challenging, because while people can take some proactive actions to manage their concerns about sending intimate content (e.g., avoiding identifying features), they can only take reactive steps to manage unsolicited sexts (short of being forced to opt out of platforms entirely).

This threat model suggests that platforms may need to take a more proactive role in mitigating unsolicited and harassing sexual content, not only in response to user reports. This role could take the form of messaging affordance design: for example, not allowing photos to be sent unless both people in a conversation enable the feature. Platforms could also play a greater role in detecting and blocking certain types of content directly, as others have proposed and begun experimenting with [44]. However, this approach comes with significant challenges that future work must consider — e.g., how to integrate or balance content detection with end-to-end encryption, and how such a feature might interact with (in the U.S.) the First Amendment and a company's legal liabilities [57].

5.2 Enabling Intimacy

Existing Platform Affordances. Our study surfaced a number of extant design practices that worked to preserve sexual privacy and enable intimacy. Important affordances that surfaced in our study include screenshot notifications, disappearing messages, and password protection for files. While these features are common on platforms like Snapchat, they do not pervade a variety of the other platforms or mediums people use (but are not necessarily designed) for sexting (e.g., Facebook, SMS, Grindr, Tinder). Our results suggest that these affordances are an important starting point for other platforms looking to account for sexting. We note that the design of these affordances can and should be informed by the real threat models of users like our participants. For example, while Snapchat has been criticized in the past for its disappearing messages not being truly secure [69], we note that even less-than-perfect security may be sufficient against the more common threat models involving violations (sometime accidental) by communication partners rather than company access or sophisticated external "hackers". (Though supporting stronger threat models is also crucial for some users, especially those particularly vulnerable to targeted attacks).

More generally, the focus of our work has *not* been on unpacking the technical properties of different platforms used by our participants. Our work lays a foundation for important future work to study these technical properties and to identify — and bridge — any gaps between the threat models they securely support and the threat models important to individuals engaging in sexting.

Designing for Storage. One area of technical design that our results draw attention to is around storage. We found that large fractions of respondents save their own (52.6%) and others' (62.0%) nudes or sexual images — often for legitimate purposes and with consent — suggesting that platforms must design for this behavior as a norm rather than an indication of the intent to misuse the content. These findings shed light on the need to grapple with the digital footprint of sexting: who stores it, how and where is it stored, and how is it secured? Our findings suggest a variety of overlapping models, including storage on personal devices, storage in a file-sharing service (e.g., Google Drive), and platform-based storage (e.g., some dating and hook-up platforms, like Scruff, Growlr, and Jack'd, provide built-in "private photo" storage where images can be access-controlled and access-monitored). Participants sometimes used a storage strategy deliberately and sometime incidentally (e.g., storing one's own photos in the device "Camera Roll"). Different strategies support different threat models, and we recommend that both users and platform designers face these choices consciously.

Opting Out as a Last Resort. Finally, while only a few people indicated not sexting as a mitigation strategy, we highlight that this decision relates to (or rather, hinders) enabling intimacy. First, abstaining from sexting is a valid behavior to alleviate sexual privacy risk as well as a valid boundary one might establish with respect to their intimate life. Second, however, if sexual privacy concerns are causing people not to sext, but there are other behaviors or design practices that *could* mitigate these concerns, then it is not the optimal outcome for people to feel like they have to choose to opt out of sharing sexual media. A goal of sexting platform design or other interventions, then, should be to support positive sexting and not force people to opt out due to unmitigated risks.

5.3 Protecting Equality

Supporting prior work [14, 19, 39, 53, 67, 75], our results provide further evidence to suggest that women and sexual minorities are disproportionately burdened by certain sexual privacy risks — receiving more unsolicited sexts, feeling pressured to sext, worrying more about negative judgements (both for sending and receiving) and the potential misuse of their intimate content. It is crucial that future work in this space further study such disparate impacts and take them into account when designing to mitigate potential risks with sexting.

Our results also highlighted the potentially generic threat models of many participants when asked to consider sexual privacy concerns without prompting. Many participants expressed vague concerns, often in the passive voice, about their sexts "getting out". Though these responses could be due in part to the survey methodology (where we could not follow up to clarify the vague responses, and participants might have opted to answer the question quickly rather than exhaustively), these results echo findings from prior work [74] and raise concerns that participants have internalized "victim-blaming" perspectives, shifting the responsibility away from untrustworthy partners and other actors who take advantage of normal sexting behavior. We recommend that future work dig deeper into these questions.

6 Limitations

Our survey-based approach is subject to standard limitations of this methodology: for example, we could not ask follow-up questions to clarify or dig further into participants' responses. Sometimes participants gave generic responses about their concerns, mentioning only "privacy" without additional detail about their mental threat models (e.g., privacy of what, or against whom). Since we could not follow up, we cannot distinguish vague mental models from survey fatigue, and we report our results cautiously in such contexts.

Our sample of participants is not representative of the overall U.S. population — with higher rates of LGBTQ individuals and women in particular, and with only 10.43% participants over 34 — so our results do not support any census-representative claims. With our screening method, it is unclear if older adults were reached and chose not to sext, excluding them from from our participant pool, or if they were not reached through our recruitment methods. We also do not capture the sexting concerns of people who haven't sexted before. We also recognize that an online survey on sexting behaviors will elicit some degree of self-selection bias insofar as those who participate may have higher baseline comfort and knowledge around sexting.

7 Conclusion and Future Work

Via an online survey of 247 adults who sext, our findings contribute to the field of usable privacy and security by expanding our understanding of how adults navigate sexting, using both technical strategies such as disappearing messages and non-technical strategies such as relying on trust. We show (similar to prior work) that men were less likely than women and non-binary individuals to be concerned about certain potential sexting risks, and less likely to receive unsolicited sexts. Placing our results in the context of the sexual privacy framework, we suggest ways platforms can support autonomy, intimacy, and equality through platform affordances and policies.

Future work on usable sexual privacy and security should consider (1) how communication platforms can surface and scaffold individuals' norms, expectations, and boundaries around sexting, (2) how usable security can address the broader inequities in the experience of sexual privacy harms, and (3) how the technical properties of privacy-enabling affordances compare to user expectations and assumptions around the security and privacy implications of such features.

Acknowledgments

We are especially grateful to our survey participants. We also thank our reviewers and shepherd for their helpful feedback. We thank Ryan Calo, Catherine Holmes, Naveena Karusala, Shrirang Mare, Eric Zeng, Ben Zisk, and the UW Statistical Consulting Services for their guidance and input. And we thank the moderators of reddit.com/r/sex and of queer social groups for helping advertise our study. This research is supported in part by the National Science Foundation under Award CNS-1513584.

References

[1] When does Snapchat delete snaps and chats? https://support.snapchat.com/en-US/article/when-are-snaps-chats-deleted.

[2] Signal removes EXIF data #7862, 2018. https://github.com/signalapp/Signal-Android/issues/7862.

[3] Irwin Altman. *The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding*. Brooks/Cole Publishing Company, 1975.

[4] Irwin Altman and Dalmas A Taylor. *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston, 1973.

[5] Rikke Amundsen. 'The Price of Admission': On Notions of Risk and Responsibility in Women's Sexting Practices. In Karen Lumsden and Emily Harmer, editors, *Online Othering: Exploring Digital Violence and Discrimination on the Web*, Palgrave Studies in Cybercrime and Cybersecurity. Palgrave Macmillan, 2019.

[6] Sonu Bedi. Sexual racism: Intimacy as a matter of justice. *The Journal of Politics*, 77(4):998–1011, 2015.

[7] Lauren Berlant. Intimacy: A special issue. *Critical inquiry*, 24(2):281–288, 1998.

[8] Lauren Berlant and Michael Warner. Sex in public. *Critical inquiry*, 24(2):547–566, 1998.

[9] David G. Blanchflower and Andrew J. Oswald. Money, sex and happiness: An empirical study. *Scandinavian Journal of Economics*, 106(3):393–415, 2004.

[10] Edward J. Bloustein and Nathaniel J. Pallone. *Individual and group privacy*. Routledge, 2018.

[11] Melissa Burkett. Sex (t) talk: A qualitative analysis of young adults' negotiations of the pleasures and perils of sexting. *Sexuality & Culture*, 19(4):835–863, 2015.

[12] Christine Chen, Nicola Dell, and Franziska Roesner. Computer security and privacy in the interactions between victim service providers and human trafficking survivors. In *USENIX Security Symposium*, 2019.

[13] HyeJeong Choi, Joris Van Ouytsel, and Jeff R. Temple. Association between sexting and sexual coercion among female adolescents. *Journal of adolescence*, 53:164–168, 2016.

[14] Danielle Keats Citron. *Hate crimes in cyberspace*. Harvard University Press, 2014.

[15] Danielle Keats Citron. Sexual privacy. *128 Yale Law Journal 1870 (2019); U of Maryland Legal Studies Research Paper No. 2018-25*, 2019.

[16] Camille Cobb and Tadayoshi Kohno. How public is my private life? Privacy in online dating. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, 2017.

[17] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[18] Judith Wagner DeCew. *In pursuit of privacy: Law, ethics, and the rise of technology*. Cornell University Press, 1997.

[19] Allyson Dir, Ayca Coskunpinar, Jennifer Steiner, and Melissa Cyders. Understanding differences in sexting behaviors across gender, relationship status, and sexual identity, and the role of expectancies in sexting. *Cyberpsychology, behavior and social networking*, 16, 05 2013.

[20] Michelle Drouin, Manda Coupe, and Jeff R. Temple. Is sexting good for your relationship? It depends.... *Computers in Human Behavior*, 75:749–756, 2017.

[21] Nicola Döring. Consensual sexting among adolescents: Risk prevention through abstinence education or safer sexting? *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 8, 01 2014.

[22] Nicola Döring and M. Rohangis Mohseni. Are Online Sexual Activities and Sexting Good for Adults' Sexual Well-Being? Results From a National Online Survey. *International Journal of Sexual Health*, 30(3):250–263, July 2018.

[23] Elizabeth F Emens. Intimate discrimination: The state's role in the accidents of sex and love. *Harv. L. Rev.*, 122:1307, 2008.

[24] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. "A stalker's paradise": How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, New York, NY, USA, 2018. Association for Computing Machinery.

[25] Diana Freed, Jackeline Palmer, Diana Elizabeth Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. Digital technologies and intimate partner violence: A qualitative analysis with multiple stakeholders. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22, 2017.

[26] Charles Fried. *An anatomy of values*, volume 2. HeinOnline, 2013.

[27] Erich Fromm. *The art of loving: The centennial edition*. A&C Black, 2000.

[28] Billy Gallagher. How Reggie Brown invented Snapchat, February 2018. https://techcrunch.com/2018/02/10/the-birth-of-snapchat/.

[29] Christine Geeng and Franziska Roesner. Who's in control? Interactions in multi-user smart homes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery.

[30] Adam Isaiah Green. The social organization of desire: The sexual fields approach. *Sociological Theory*, 26(1):25–50, 2008.

[31] Adam Isaiah Green. Sexual capital and social inequality. *Introducing the New Sexuality Studies*, 272, 2016.

[32] A.A Hasinoff. How to have great sext: consent advice in online sexting tips. *Communication and Critical/Cultural Studies*, 13(1):58–74, 2016.

[33] Amy Adele Hasinoff. Sexting in Context: Privacy Norms and Expectations. page 24, 2014.

[34] Debby Herbenick, Jessamyn Bowling, Tsung-Chieh (Jane) Fu, Brian Dodge, Lucia Guerra-Reyes, and Stephanie Sanders. Sexual diversity in the United States: Results from a nationally representative probability sample of adult women and men. *PLOS ONE*, 12(7):e0181198, July 2017.

[35] Jevan A Hutson, Jessie G Taft, Solon Barocas, and Karen Levy. Debiasing desire: Addressing bias & discrimination on intimate platforms. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–18, 2018.

[36] Clara Rübner Jørgensen, Annalise Weckesser, Jerome Turner, and Alex Wade. Young people's views on sexting education and support needs: Findings and recommendations from a UK-based study. *Sex Education*, 19(1):25–40, 2019.

[37] Kami Kosenko, Geoffrey Luurs, and Andrew R Binder. Sexting and sexual behavior, 2011–2015: A critical review and meta-analysis of a growing literature. *Journal of computer-mediated communication*, 22(3):141–160, 2017.

[38] Natalie A. Koziol and Christopher R. Bilder. MRCV: A package for analyzing categorical variables with multiple response options. *R Journal*, 6(1):144–150, 2014.

[39] Amanda Lenhart, Michele Ybarra, and Myeshia Price-Feeney. Nonconsensual image sharing, 2016. https://datasociety.net/output/nonconsensual-image-sharing/.

[40] Karen Levy and Bruce Schneier. Privacy threats in intimate relationships. *Journal of Cybersecurity*, 6(1), 05 2020. tyaa006.

[41] Julia R. Lippman and Scott W. Campbell. Damned if you do, damned if you don't... if you're a girl: Relational and normative contexts of adolescent sexting in the United States. *Journal of Children and Media*, 8(4):371–386, 2014.

[42] Sheri Madigan, Anh Ly, Christina L. Rash, Joris Van Ouytsel, and Jeff R. Temple. Prevalence of Multiple Forms of Sexting Behavior Among Youth: A Systematic Review and Meta-analysis. *JAMA Pediatrics*, 172(4):327–335, 04 2018.

[43] Alice E Marwick and danah boyd. Networked privacy: How teenagers negotiate context in social media. *New Media & Society*, 16(7):1051–1067, 2014.

[44] Adrienne Matei. The war on (unwanted) dick pics has begun. The Guardian, September 2019. https://www.theguardian.com/lifeandstyle/2019/sep/19/its-a-violation-the-war-on-unwanted-dick-pics-has-begun.

[45] Louise Matsakis. The Motherboard guide to sexting securely, 2017. https://www.vice.com/en_us/article/mb3nd4/how-to-sext-securely-safely-what-apps-to-use-sexting.

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

[47] Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F. Churchill, and Sunny Consolvo. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[48] Linda C. McClain. Inviolability and privacy: The castle, the sanctuary, and the body. *7 YALE J.L. & HUMAN. 195, 241*, 1995.

[49] Nora McDonald and Andrea Forte. The politics of privacy theories: Moving from norms to vulnerabilities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[50] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[51] Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3), Oct. 2012.

[52] Maurice Merleau-Ponty. *Phenomenology of perception*. Routledge, 1982.

[53] Robert Meyer and Michel Cukier. Assessing the attack threat due to irc channels. pages 467–472, 02 2006.

[54] Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. Survey research in HCI. In *Ways of Knowing in HCI*. Springer-Verlag New York, New York, NY, USA, 2014.

[55] Frank Newport. In U.S., estimate of LGBT population rises to 4.5%, 2018. https:

//news.gallup.com/poll/234863/estimate-lgbt-population-rises.aspx.

[56] Helen Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

[57] Anna North. One state has banned unsolicited dick pics. Will it fix the problem? Vox, September 2019. https://www.vox.com/policy-and-politics/2019/9/3/20847447/unsolicited-dick-pics-texas-law-harassment.

[58] Natalie Oswin and Eric Olund. Governing intimacy. *Environment and Planning D: Society and Space*, 28(1):60–67, 2010.

[59] Erdman B Palmore. Predictors of the longevity difference: a 25-year follow-up. *The Gerontologist*, 22(6):513–518, 1982.

[60] Cheul Young Park, Cori Faklaris, Siyan Zhao, Alex Sciuto, Laura Dabbish, and Jason Hong. Share and share alike? An exploration of secure behaviors in romantic relationships. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 83–102, Baltimore, MD, August 2018. USENIX Association.

[61] Justin W Patchin and Sameer Hinduja. It is time to teach safe sexting. *Journal of Adolescent Health*, 66(2):140–143, 2020.

[62] Jeffrey H Reiman. Privacy, intimacy, and personhood. *Philosophy & Public Affairs*, pages 26–44, 1976.

[63] John K Rempel, John G Holmes, and Mark P Zanna. Trust in close relationships. *Journal of Personality and Social Psychology*, 49(1):95, 1985.

[64] Daniel G. Renfrow and Elisabeth A. Rollo. Sexting on Campus: Minimizing Perceived Risks and Neutralizing Behaviors. *Deviant Behavior*, 35(11):903–920, November 2014.

[65] Neil Richards and Woodrow Hartzog. Taking trust seriously in privacy law. *Stan. Tech. L. Rev.*, 19:431, 2015.

[66] Dorothy E Roberts. *Killing the black body: Race, reproduction, and the meaning of liberty*. Vintage, 1999.

[67] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. "They don't leave us alone anywhere we go": Gender and digital abuse in South Asia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery.

[68] Morgan Klaus Scheuerman, Katta Spiel, Oliver L. Haimson, Foad Hamidi, and Stacy M. Branham. HCI guidelines for gender equity and inclusivity, 2010. https://www.morgan-klaus.com/sigchi-gender-guidelines.

[69] Alyson Shontell. Actually, snapchat doesn't delete your private pictures and someone found a way to resurface them. Business Insider, 2013. https://www.businessinsider.com/snapchat-doesnt-delete-your-private-pictures-2013-5.

[70] Scott Skinner-Thompson. Performative privacy. *UCDL Rev.*, 50:1673, 2016.

[71] Scott Skinner-Thompson. Privacy's double standards. *Wash. L. Rev.*, 93:2051, 2018.

[72] Emily C. Stasko and Pamela A. Geller. Reframing sexting as a positive relationship behavior. Drexel University, 2015. https://www.apa.org/news/press/releases/2015/08/reframing-sexting.pdf.

[73] Jeff R. Temple. A primer on teen sexting. *JAACAP Connect*, 2(4):6–8, 2015.

[74] Rebecca Venema and Katharina Lobinger. "And somehow it ends up on the internet." Agency, trust and risks in photo-sharing among friends and romantic partners. *First Monday*, 22(7), Jul. 2017.

[75] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1231–1245, New York, NY, USA, 2017. Association for Computing Machinery.

[76] Ari Ezra Waldman. Privacy as trust: Sharing personal information in a networked world. *U. Miami L. Rev.*, 69:559, 2014.

[77] L. Monique Ward. Understanding the role of entertainment media in the sexual socialization of american youth: A review of empirical research. *Developmental review*, 23(3):347–388, 2003.

A Survey Instrument

1. Are you 18 years old or over?
   ○ Yes ○ No
2. Have you ever sexted? That is, do you create, send, or receive sexually suggestive messages, or nude or partially-nude photos, through digital communications?
   ○ Yes, I currently sext. ○ No, I have never sexted. ○ Yes, I have sexted before and may again in the future. ○ Yes, I have sexted before but no longer plan to.
3. Why did you stop sexting? (Select all that apply.)
   □ No longer in a relationship with the person who I sexted with (please elaborate) □ No longer in a long-distance relationship □ Not interested (please elaborate) □ Poor experience (please elaborate) □ Other: □ Prefer not to say
4. How often do you send:
   (Options: Never, Less than once a month), Once a month, Once a week, A few times a week, Almost everyday, Multiple times per day, Almost hourly, Prefer not to say)
   Nude or semi-nude photos ○ ○ ○ ○ ○ ○ ○ ○ ○

Nude or semi-nude videos ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Sexual or intimate messages (such as words or emojis) ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

5. How often do you receive:
(Options: Never, Less than once a month), Once a month, Once a week, A few times a week, Almost everyday, Multiple times per day, Almost hourly, Prefer not to say)
Nude or semi-nude photos ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Nude or semi-nude videos ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
Sexual or intimate messages (such as words or emojis) ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
6. What social media platforms do you use to send/receive nude or semi-nude photos or videos? Select all that apply (or select nothing if you do not send this type of content).
(Options: Direct Message, Private Post, Public Post, Story, Other)
Facebook/FB Messenger, Twitter, Instagram, Snapchat, Other social media app:
7. What other platforms do you use to send/receive nude or semi-nude photos or videos? Select all that apply, if any.
☐ SMS ☐ Whatsapp ☐ Tinder ☐ Grindr ☐ Hinge ☐ OkCupid ☐ Signal ☐ Telegram ☐ Confide ☐ Kaboom ☐ Dust ☐ Other platforms (separated by comma):
8. What social media platforms do you use to send/receive sexual or intimate messages (such as words or emojis)? Select all that apply (or select nothing if you do not send this type of content).
(Options: Direct Message, Private Post, Public Post, Story, Other)
Facebook/FB Messenger, Twitter, Instagram, Snapchat, Other social media app: ☐ ☐ ☐ ☐ ☐
9. What other platforms do you use to send/receive sexual or intimate messages (such as words or emojis)? Select all that apply, if any.
☐ SMS ☐ Whatsapp ☐ Tinder ☐ Grindr ☐ Hinge ☐ OkCupid ☐ Signal ☐ Telegram ☐ Confide ☐ Kaboom ☐ Dust ☐ Other platforms (separated by comma):
10. What devices do you use to send/receive sexts? Select all that apply.
☐ Phone ☐ Tablet ☐ Computer ☐ Other
11. With whom do you send nude or semi-nude photos? Select all that apply.
☐ Partner ☐ Regular sexual hookup (purely sexual relationship) ☐ Casual date or one-time hookup ☐ Friend-with-benefits ☐ Friend ☐ Acquaintance ☐ New person online ☐ Other: ☐ Prefer not to say
12. With whom do you send nude or semi-nude videos? Select all that apply.
☐ Partner ☐ Regular sexual hookup (purely sexual relationship) ☐ Casual date or one-time hookup ☐ Friend-with-benefits ☐ Friend ☐ Acquaintance ☐ New person online ☐ Other: ☐ Prefer not to say
13. With whom do you send sexual or intimate messages (such as words or emojis)? Select all that apply.
☐ Partner ☐ Regular sexual hookup (purely sexual relationship) ☐ Casual date or one-time hookup ☐ Friend-with-benefits ☐ Friend ☐ Acquaintance ☐ New person online ☐ Other: ☐ Prefer not to say
14. With whom do you receive nude or semi-nude photos? Select all that apply.
☐ Partner ☐ Regular sexual hookup (purely sexual relationship) ☐ Casual date or one-time hookup ☐ Friend-with-benefits ☐ Friend ☐ Acquaintance ☐ New person online ☐ Other: ☐ Prefer not to say
15. With whom do you receive nude or semi-nude videos? Select all that apply.
☐ Partner ☐ Regular sexual hookup (purely sexual relationship) ☐ Casual date or one-time hookup ☐ Friend-with-benefits ☐ Friend ☐ Acquaintance ☐ New person online ☐ Other: ☐ Prefer not to say
16. With whom do you receive sexual or intimate messages (such as words or emojis)? Select all that apply.
☐ Partner ☐ Regular sexual hookup (purely sexual relationship) ☐ Casual date or one-time hookup ☐ Friend-with-benefits ☐ Friend ☐ Acquaintance ☐ New person online ☐ Other ☐ Prefer not to say
17. Do you or have you ever saved any of the nude photographs or videos you have received?
○ Yes ○ No ○ Prefer not to say ○ Not applicable
18. Why did/do you save them?
19. How do you save them?
☐ Directly store to device ☐ Screenshot ☐ Take a photo ☐ Other:
20. What do you plan to do with them?
○ Save them indefinitely ○ Save them for some amount of time (please indicate an estimate): ○ Save them until asked to delete them ○ Other: ○ Prefer not to say
21. Approximately how many nudes of other people do you have saved on your device?
○ None ○ 1 to 10 ○ 11 to 100 ○ 101+ ○ Prefer not to say
22. How do you store other people's nudes? Select all that apply.
☐ Device's photo storage: Camera Roll ☐ Device's photo storage: separate album ☐ Online (for example: Google Photos, Dropbox): ☐ Specific secret-keeping app (for example: Private Photo Vault): ☐ Other:
23. Does the person(s) who sent you the nudes know you've saved them?
○ Yes ○ Some of the senders know ○ No ○ Prefer not to say
24. Do you store nude or semi nude photos or videos of yourself?
○ Yes ○ No ○ Prefer not to say
25. Approximately how many nudes of yourself do you have saved?

○ None ○ 1 to 10 ○ 11 to 100 ○ 101+ ○ Prefer not to answer

26. How do you store your nudes? Select all that apply.
    □ Device's photo storage: Camera Roll □ Device's photo storage: separate album □ Online (for example: Google Photos, Dropbox): □ Specific secret-keeping app (for example: Private Photo Vault): □ Other:
27. Do you share received sexts with people other than the sender?
    ○ Yes (please elaborate how you share): ○ No ○ Prefer not to say
28. If you share other people's sexts digitally, do you digitally edit these sexts?
    ○ Yes (please elaborate): ○ No ○ I do not share digitally ○ Prefer not to say
29. Have you received sexts or nudes from people you did not want to receive them from?
    ○ Yes ○ No ○ Prefer not to say
30. How do you manage receiving sexts or nudes from people you do not want to receive them from?
31. Do you have any concerns related to sending sexts?
    ○ Yes (please elaborate below) ○ No ○ Prefer not to say
32. If you answered "yes" to the previous question, please elaborate here: What are your concerns related to sending sexts? Or if not, why not?
33. What concerns do you have about sending sexts?
    □ Sexts get around to other people □ Damages relationships □ Conflicts at work □ Legal liability □ Engagement with law enforcement (e.g police) □ Sexting causes ridicule from others □ Unwanted attention □ Unwanted sexual contact □ Sexts used as blackmail □ Bullying or harassment from others □ Regret □ Sexting makes people feel "led on", "used", or "misunderstood □ Not sure I sent it to the right person □ Receiver's devices will get hacked and the content will get out □ Receiver will intentionally share the content with others □ Other
34. Do you have any concerns related to receiving sexts?
    ○ Yes (please elaborate below) ○ No ○ Prefer not to say
35. If you answered "yes" to the previous question, please elaborate here: What are your concerns about receiving sexts? Or if not, why not?
36. What concerns do you have about receiving sexts?
    □ Not sure if it's really the person I think it is □ My device will get hacked and their content will get out □ Shoulder surfing □ Receiving unsolicited/non-consensual content □ Other:
37. If you don't have concerns about sexting, why is that?
    □ I've done something to manage my concerns □ I trust the people I sext with □ I trust the platform I use to sext □ I don't care about how people react to my nudity and sexual expression □ My sexts are already public □ I'm just not really worried about it □ Other:
38. Describe your level of concerns related to sending certain types of sexts.
    (Options: Not at all concerned, Slightly concerned, Somewhat concerned, Moderately concerned, Extremely concerned, N/A)
    Photo, Video, Text Based
39. Describe your level of concerns related to receiving different types of sexts.
    (Options: Not at all concerned, Slightly concerned, Somewhat concerned, Moderately concerned, Extremely concerned, N/A)
    Photo, Video, Text Based
40. You selected that you were concerned about the following when sending sexts: [input]. Do your concerns depend upon the type of person with whom you sext, the type of platform you use, or other considerations?
41. You selected that you were concerned about the following when receiving sexts: [input]. Do your concerns depend upon the type of person with whom you sext, the type of platform you use, or other considerations?
42. Do you do any of the following to manage your sexting concerns?
    □ Choose a platform with specific features you want □ Using disappearing messages e.g. Snapchat, Instagram stories □ Password-protect or encrypt sexts □ Prior conversations to establish rules and boundaries □ Ensuring plausible deniability e.g. not including identifying marks in photo □ Limiting how explicit the sext is □ Only sexting with people you trust □ I do not have any concerns □ I do not have any strategies to manage my concerns □ Other:
43. You selected that you use the following strategies to manage your sexting: [input]. Could you please elaborate?
44. Can we use anonymized quotes from your free-response answers in future research publications?
    ○ Yes ○ No
45. What gender(s) do you identify as?
    □ Male □ Female □ Non-binary □ Prefer not to say □ Prefer to self-describe:
46. Do you consider yourself transgender?
    ○ Yes ○ No ○ Questioning ○ Prefer not to say
47. What is your sexual orientation?
    ○ Straight ○ Questioning ○ Gay ○ Lesbian ○ Bi/Pan ○ Queer ○ Asexual ○ Prefer not to say ○ Prefer to self-describe:
48. Do you consider yourself polyamorous or monogamous (regardless of current relationship status)?
    ○ Polyamorous ○ Monogamous ○ Prefer not to say
49. Which racial background(s) do you identify as?
    □ Asian □ Black □ Latino □ Native American □ Pacific Islander □ White □ Prefer not to say □ Prefer to self-describe

50. What is your current intimate status? (Select all that apply)
    ☐ Single ☐ Dating ☐ Engaged ☐ Married ☐ Divorced ☐ Widowed ☐ Friends-With-Benefits ☐ Casual sex ☐ Casual dating ☐ Monogamous relationship ☐ Polyamorous relationship ☐ Prefer not to say
51. What is your age?
    ○ 18-24 years old ○ 25-34 years old ○ 35-44 years old ○ 45-54 years old ○ 55-64 years old ○ 65-74 years old ○ 75 years or older
52. What kind of location do you live in?
    ○ Urban ○ Suburban ○ Rural ○ Prefer not to say
53. Have you ever been in an IT/technology related job?
    ○ Yes ○ No ○ Prefer not to say
54. I understand how to control or protect my personal data online.
    ○ Strongly agree ○ Agree ○ Somewhat agree ○ Neither agree nor disagree ○ Somewhat disagree ○ Disagree ○ Strongly disagree
55. What is your education level?
    ○ GED ○ Some high school ○ Some college/technical training ○ Some graduate school ○ Prefer not to say

B Additional Data on Gender Comparisons

Tables 14 and 15 show the data on gender comparisons for two receiving-related concerns, discussed in Section 4.2.2. The *N* value for both tables includes all participants who said they receive sexts and who selected at least Male, Female, or Non-binary for their gender.

Concerned about receiving unsolicited sexts?

	No	**Yes**	**Total**
Male	68	7	75
Female	118	33	151
Non-Binary	19	5	24

Table 14: $N = 242$

Concerned about shoulder-surfing?

	No	**Yes**	**Total**
Male	70	5	75
Female	124	27	151
Non-Binary	22	2	24

Table 15: $N = 242$

C Inter-Coder Reliability

Tables 16 and 17 show the breakdown of Cohen's κ for inter-coder reliability per code, discussed in Section 3.4.

Managing Sending Concerns	κ
Only sext with trusted/known person	0.94
Don't sext	0.66
Communicating/establishing rules and expectations	0.90
Limit explicitness	0.91
Anonymize sext (no face, tattoos, names, or location)	0.93
Ask person to delete	1.00
Acquire collateral	1.00
Passcode protect image/app/device	1.00
Disappearing messages	1.00
Explicitly deleting messages/chat/media	0.66
Encrypted platforms	0.62
Screenshot notifications	1.00
Only send stuff willing to go public	0.49
Platform choice/Platform affordances	0.88
Making sure notifications don't make sexts visible	1.00
Other	0.82

Table 16: Cohen's Kappa for codes for elaboration on management strategies for sexting concerns.

Concerns About Sending Sexts	κ
Impact career	1.00
Sexts will get out (Vague)	0.69
Sexts used as blackmail	0.87
Deanonymization	1.00
Sending to wrong person	0.94
Bug or vulnerability in app	0.50
Revenge porn	0.87
Hacking or stealing	0.92
Recipient will save sexts	0.83
End up online	0.73
Shared/shown to others	0.81
Seen accidentally by non-recipient	0.77
Recipient will misuse (Generic, Other)	0.73
Not concerned because trust partner	0.79
Not concerned because trust app	0.80
Not trusting companies	0.72
Access by government	1.00
Images will be used to impersonate sender	0.66
Accidentally posting publicly	1.00
Judgement from others (also: embarrassing)	0.76
Insecure network or cloud	1.00
Photos will be modified	1.00
Other	0.32

Table 17: Cohen's Kappa for codes for open-ended question, "What are your concerns related to sending sexts?"

Towards Understanding Privacy and Trust in Online Reporting of Sexual Assault

Borke Obada-Obieh
University of British Columbia
borke@ece.ubc.ca

Lucrezia Spagnolo
Vesta Social Innovation Technologies
lucrezia@vestasit.com

Konstantin Beznosov
University of British Columbia
beznosov@ece.ubc.ca

Abstract

According to the United States Department of Justice, every 73 seconds, an American is sexually assaulted. However, sexual assault is under-reported. Globally, 95% of sexual assault cases are unreported, and at most, 5 out of every 1,000 perpetrators end up in prison. Online anonymous third-party reporting systems (O-TPRSs) are being developed to encourage reporting of sexual assaults and to apprehend serial offenders. This paper reports survivors' concerns with trusting and using an O-TPRS. We conducted focus groups and interviews with 35 participants who are sexual assault survivors, support workers, or both. We asked questions related to participants' concerns with trusting an O-TPRS. Our results suggest that participants had technological and emotional concerns that are related to survivors' security and privacy. We provide insights into the challenges of designing O-TPRSs to increase the reporting of sexual assault.

1 Introduction

The goal of our research is that interdisciplinary innovations in human-computer interaction, privacy, and security can be used to empower survivors of sexual assault to encounter healing and justice. Our investigation into designing safe spaces online for anonymous third-party reporting (TPR) is a response to the clear need for a confidential and accessible platform that survivors of sexual assault can use to communicate their experiences in the hope of holding perpetrators accountable.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

The stark reality is that 1 in 3 Canadian women will experience sexual assault in their adult life [50]. Further, 1 in 14 American men and 1 in 5 American women have been victims of completed or attempted sexual assault during their lifetime [60]. Sexual assault has no single impact but affects multiple areas of the survivor's life, including but not limited to the survivor's somatic and psychological health [14, 19]. One in four survivors reported that they had difficulty carrying out everyday activities because of the incident [43]. Further, one in six survivors reported experiencing three or more longer-term emotional consequences, such as post-traumatic stress disorder, substance abuse, depression, and suicidal thoughts [20, 42, 43].

However, statistics alone fail to capture the significant repercussions of sexual assault on survivors, not only because the effects of such trauma are unquantifiable [14] but also because sexual assault is greatly underreported [48, 55]. Only 5% of cases are reported to the police [51], and only 11% of those reported cases eventually lead to the conviction of the perpetrator [56]. The reluctance of survivors to report the crime to the police has mainly been attributed to the cumbersome reporting process and to the grueling interview procedure involved in filing a formal police report, which can be adversarial and emotionally very unpleasant for survivors [39, 43, 44, 53, 54].

To expand the reporting options for survivors, third-party reporting centers have been put in place. Third-party reporting is when someone else reports the crime to the police on behalf of the survivor [11], who remains anonymous. Third-party reporting systems (TPRSs) allow survivors to anonymously report sexual assault to the police through a community-based support center [11, 40]. TPRS is an option used when a survivor does not want to visit a police station to make a formal police report. This option is useful for two main reasons. First, it allows survivors to record details of a perpetrator anonymously [40]. Second, when multiple survivors indicate the same perpetrator, a serial offender is identified. In this case, the police contacts the community-based support center to ask the survivor if they would consent to make a formal

police report so that the police can begin a formal investigation [11]. Many of the survivors who file a third-party report and are then approached by the third party and told that the police are interested in investigating their report follow up and file a formal report with the police [10]. The resulting filing of formal police reports has led to an increase in arrests of serial offenders [10].

Third-party reporting is, however, very limited in scope. It is currently administered on paper (P-TPRS), and there are no online systems to facilitate the reporting process, which makes the process cumbersome (for instance, survivors have to locate and visit a third-party reporting center) [11, 40]. Further, third-party reporting is also not available in all sexual assault support centers but only in a few select jurisdictions [11, 13], which defeats its purpose of increasing sexual assault reporting [40, 57]. Online third-party reporting systems (O-TPRSs) are being developed to increase the reporting choices for survivors. With an O-TPRS, survivors can, at their convenience, document their experience and offender information before submitting the report to the police. An O-TPRS could decrease barriers for vulnerable populations who do not currently have access to reporting options, and whose reporting rates are even lower than the estimated averages already cited.

Since an O-TPRS will hold sensitive information, we must address the privacy and security concerns of survivors. A considerable amount of research has been conducted on sexual assault and sexual assault survivors [7, 9, 20, 42, 58]. Some research also investigates the reporting experiences of survivors [7], including sexual assaults within the armed forces [17] and police-reported sexual assaults against youths and children [18]. However, no research has focused on survivors' concerns regarding trusting O-TPRSs. To this aim, the objective of this research is to answer these research questions:

1. RQ1: What are survivors' privacy and security concerns (if any) regarding trusting O-TPRSs?
2. RQ2: What could help participants trust O-TPRSs?

"Trust is the degree to which people believe in the veracity or effectiveness of a tool or system to do what it was created for and is purported to do [31]." The act of measuring trust is used to predict whether survivors would make use of O-TPRS technology [32]. Answering these research questions, therefore, will lead to understanding what it would take for users to make use of an O-TPRS. These answers could lead to an increase in the reporting of sexual assaults.

We addressed our research questions by conducting six focus groups and eight individual semi-structured interviews with a total of 35 participants. They were survivors, sexual assault support workers, or both. We asked questions relating to participants' concerns with trusting an O-TPRS and analyzed the results using thematic analysis.

Our study has two major contributions. First, we performed the first empirical study on sexual assault survivors to discover their privacy and security concerns regarding trusting an O-TPRS. We group our findings into technological and emotional concerns, and we show how technological concerns can lead to emotional issues for survivors. For example, the technological concern about the *insecurity of technology* can lead to the emotional issue of *anxiety* about making an online report, the *fear of perpetrators having access to the sexual assault report*, and the *re-victimization of survivors*. Second, we discovered concerns that technologists need to consider in developing O-TPRSs. For instance, on the one hand, survivors did not trust that an O-TPRS could protect their anonymity and privacy. On the other hand, the police did not trust that the anonymous reports sent from an O-TPRS were linked to real survivors. Technologists would, therefore, need to find a balance in how an O-TPRS can ensure both parties can trust the system.

Our contributions provide insights into concerns that survivors and support workers have about using online systems to report sexual assault. We are optimistic that when O-TPRSs are designed with careful attention to users' feedback and research, such systems could increase reporting.

2 Background and Related Work

In its current format, a TPRS is a process or protocol to make an anonymous report of a sexual assault by a community-based support center. A TPRS is not a substitute for an emergency call, nor is it a formal police report. It is not to be used when the survivor or others are at risk of further violence. A TPRS is intended to be used when the survivor does not want to make a formal police report but prefers to report anonymously. A TPRS is useful for the identification of offenders, especially repeat offenders.

2.1 P-TPRS

2.1.1 The P-TPR form

The current TPRS is in paper form. We describe a P-TPRS currently in use in a jurisdiction in Ontario, Canada. Page one of the P-TPRS is a cover sheet where survivors write their personal information. On pages two and three, survivors describe the offender and the offense (see Appendix C for the questions asked on a sample P-TPR form.)

2.1.2 The P-TPR process

The survivor goes to a community-based center to carry out the P-TPRS process. The community-based center, which is usually a hospital or a sexual assault support center, is the third party. The survivor meets with a representative, either a nurse or a social worker, at the third-party reporting center. If the survivor is not willing to make a formal police report at this time, the representative at the center can provide the option

of filling out a third-party report form. The survivor has to fill out the form at the center and return it to the representative before leaving the center. If the survivor doesn't feel capable of filling out the form by themselves, the representative can listen to the survivor's story and fill out the form with the survivor's consent. Afterward, the representative de-identifies the form by removing the cover sheet. The representative sends the de-identified P-TPR form to the police. However, the hospital or the sexual assault support center, which is the third party, maintains the identity of the survivor. The police receive the content of the form and enter it into a database, making it easier to identify serial offenders [11].

A serial offender is identified if at least three people accuse the same person of sexual assault. If a serial offender or a trend is identified, or if the police believe the survivor is in imminent danger, the police can contact the community-based center. The center can reach out to the survivor to see if the survivor is willing to take further part in the investigation or even if they might consider changing their report from an anonymous report to a formal police report [11]. Figure 1 shows the P-TPR process.

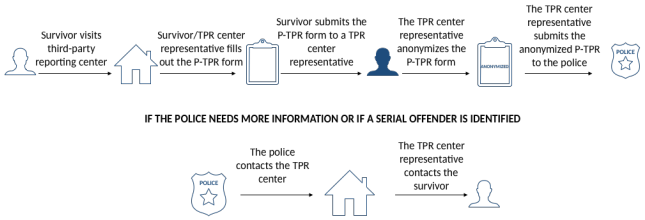


Figure 1: P-TPR process

2.2 O-TPRS

The O-TPRS supports the goal of reducing barriers to reporting by providing survivors with a new way to report that is anonymous and does not require visiting a community-based center. It also streamlines the third-party reporting process by removing the human involved in the P-TPRS.

2.2.1 The O-TPR form

The O-TPR form works similarly to the P-TPR form. We provide the description of an O-TPRS being developed by VESTA Social Innovation Technologies (Vesta) [62]. The O-TPRS includes a cover page and pages to type out information about the survivor, offender, and the offense (see Appendix D for a sample of an O-TPRS prototype).

2.2.2 The O-TPRS process

The survivor fills out the TPR form online. The O-TPRS, which could be an app or a website, is the third party. The survivor can download the O-TPRS app from the app store or can use the website version. Unlike the P-TPR form, the O-TPR provides unlimited space for the survivor to type out their experience. The survivor fills out their information, and they can save and review the information before submitting it. Before the form gets sent to the police, the O-TPRS automatically de-identifies the form. The O-TPRS, which is the third party, maintains the identity of the survivor. The police enter the content of the de-identified form into a database, making it easier to identify serial offenders. If a serial offender or a trend is identified, or if the police believe the survivor is in imminent danger, the police can contact the O-TPRS. The O-TPRS then reaches out to the survivor to see if the survivor is willing to take further part in the investigation or even if they might consider changing their report from an anonymous report to a formal police report.

O-TPRSs are not widely available. However, several organizations are looking into deploying O-TPRSs. For instance, Vesta has developed an experimental version of an O-TPRS, which is being deployed to various sexual assault centers to pilot the program. Figure 2 shows the O-TPRS process.

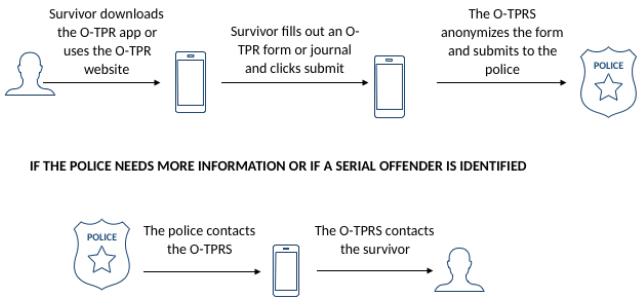


Figure 2: O-TPR process

2.3 Trust and technology

Research has been done on the concept of trust and technology usage. McKnight et al. define trust in technology as "belief that a specific technology has the attributes necessary to perform as expected in a given situation in which negative consequences are possible [45]." Prior work shows that heightened levels of trust are associated with heightened levels of intended use [27]. Trust in technology is used to predict the intended or actual adoption of technology [66]. It is also connected to appropriate and inappropriate use of technology [46] and technology over- and under-reliance [5].

Many works on technology and trust exist. Hardre, for instance, studied when, how, and why people trust technology too much [31]. Hardre analyzed various scenarios of everyday technology use where users tend to trust technology. Some of these scenarios include massive breaches of banking systems,

even though people believed that these systems would keep their financial information safe [31].

Minimal research has been done on how survivors build trust in sexual assault technology. Work by Liu is closest to ours [41]. Liu discussed issues that sexual assault prevention (such as the Circle of 6 app) and reporting technologies (such as the I've-Been-Violated app) may have in the future. The author evaluated these apps using the US Federal Trade Commission's fair information practice principles (FIPPs). Based on these principles, the author predicted that the following concerns could arise with using the apps: false allegations, security issues with the internet, fears of lack of anonymity, insensitivity to survivors' experience, lack of clarity on collected information, and lack of user-friendliness.

Our contributions are as follows: 1. We performed the first empirical study with survivors and sexual assault support workers to identify issues related to trusting O-TPRSs. 2. In addition to corroborating concerns of Liu [41] that technology could be used to make false allegations, we identify additional concerns with trusting O-TPRSs, such as the dual use of technology in not only reporting but also aiding sexual assault. 3. Further, we uncover the relationships between these concerns and discuss the issues related with designing an O-TPRS.

3 Methodology

3.1 Data Collection

We recruited participants using three methods and specific eligibility criteria. First, we used word of mouth in the professional network of one of the authors, who had extensive contacts with the workers and administration of sexual assault centers. Second, after we presented our study to an association of sexual assault centers in the Province of Ontario, its members distributed our recruitment notice to their clients, some of whom were in support groups. Third, we used snowballing with the help of already recruited participants. To be eligible to take part in the study, participants had to be 19 years old or above. Further, participants had to be survivors of sexual assault, support workers, or both. We defined support workers as those who supported survivors throughout the process of reporting sexual assault. Support workers included volunteers and staff of sexual assault report centers and the police. We recruited both survivors and support workers because both parties are involved in the TPRS process. None of the recruited participants had prior knowledge of TPRS. We recruited participants who had no prior knowledge of TPRS to get an unbiased view of both the paper and the online version of TPRS.

We piloted our study procedure with three participants—one participant for an interview session and two participants for a focus group session. In the interview pilot study, we asked the participant about her thoughts regarding O-TPRS. We realized that it was difficult for the participant to imagine how an O-TPRS would look and function. Based on this result, we made a video showing an O-TPRS prototype (see Appendix D for pictures of the prototype). We showed participants this video to illustrate an O-TPRS and to help participants understand how an O-TPRS would function. We chose to use a video for three reasons. First, for interview and focus group sessions facilitated through online video calls, we found a video more effective than a verbal explanation. Second, using a video provided a consistent explanation of the user interface across all sessions. Finally, the use of a video helped to fit each session into one hour. We piloted this approach in the pilot focus group, and we discovered that the participants could understand the O-TPRS better. We therefore used this approach for the main study. Apart from this change, all other procedures in the pilot interview and focus group were the same as those used in the main study. After adjusting the study design based on the outcomes of the pilots, we recruited participants for the main study.

We used multiple qualitative research methods [47, 65]. As suggested by Hammarberg et al. [30] and illustrated by Willis [65], using various data collection methods helps to provide better insights for sensitive research topics. We conducted semi-structured individual interviews and focus groups with participants [47]. Because of the sensitivity of the research, we gave participants the option to decide whether they were more comfortable having a semi-structured interview or participating in a focus group. For our interviews, we chose a semi-structured style to allow participants to express their thoughts in their own way and add information as they saw fit, without the restriction of a structured interview [16]. We also offered focus groups because focus groups allow participants to discuss sensitive or controversial topics in a group setting [47]. Due to participants' shared experience, sometimes focus groups "reveal aspects of experiences and perspectives that would not be as accessible without group interaction [47]," which leads to a better quality of data on sensitive topics [47].

We conducted in-person or video interviews and focus groups, based on the participants' preference, at the participants' preferred location. Some of these locations included the participants' home or a sexual assault support center. We conducted video calls via Skype or Zoom. To protect participants' privacy, online sessions were audio recorded not using Skype or Zoom but locally on a laptop. Collected data is stored on a disk encrypted with 256-bit AES seeded with a 22-character random password. Participants were compensated with $20, paid in person or sent via e-transfer. For in-person interviews, sexual assault social workers were present to provide support to participants if needed. We sent online support materials that were created by sexual assault centers to the participants that we interviewed via video call. All focus groups were held at sexual assault support centers, either by using existing support groups or by forming focus groups for interested support workers at the centers. Participants in both

online and in-person focus groups were physically present in the support centers, and sexual assault social workers were available to provide support. The social workers were compensated by their support centers, as focus groups took place during their regular work hours. We conducted seven interview sessions and five focus groups via video calls, with the rest (one interview and focus group) in person. Our institution's Research Ethics Board approved the research before any data collection took place.

We wanted to conduct separate focus groups for survivors and support workers. However, during the focus groups for support workers, some support workers self-identified as survivors. Further, when we collected participant demographics for the survivors' focus groups, we discovered that some survivors were also support workers. During data analysis, we realized that the responses from survivors and support workers were similar; therefore, distinguishing between the two groups was unnecessary. Appendix A shows participants who self-identified as survivors.

3.2 Interview and Focus Group Procedure

We proceeded with the interviews and focus groups after the participants gave informed consent to participate in the study. We assigned pseudonyms to participants and asked for their demographic information. Though we asked participants about sensitive issues, we did not ask them to disclose any sensitive information that they did not feel comfortable sharing. We reminded participants that they could skip questions they did not feel comfortable answering. During each session, we explained the meaning of P-TPRS, showed participants a copy of the P-TPR form described in Section 2.1, and asked participants their thoughts on using the P-TPRS to report sexual assault. Afterward, we played a video that explained the O-TPRS (see Section 2.2 for an explanation of the O-TPRS that was shown to participants). We then asked participants their thoughts on using the O-TPRS to report sexual assault.

To avoid priming participants, we asked participants their thoughts on using both systems rather than asking just about O-TPRS. We also asked participants what would make them comfortable using each system. We assured participants that there were no right or wrong answers, and participants could skip questions they did not feel comfortable answering.

We conducted online focus groups and interviews via Skype or Zoom based on participants' preference. For online interviews, participants chose a quiet and private location convenient for them. For the online focus group, the participants gathered at their preferred sexual assault center meeting room, and the researcher called in to conduct the focus group. We chose this arrangement because it allowed participants to get support from social workers present at the center if needed. We used focus groups and interviews because literature suggests that vulnerable populations participate better in data collection when they are given multiple choices [23]. Further, online focus groups have been found to be useful for reaching members of hard-to-reach populations [24]. Underhill and Olmsted [61] showed that there was no difference between the quality and quantity of data obtained in face-to-face and online focus groups.

Afterward, we compensated the participants. One researcher took part in each interview session. All interview sessions were audio recorded.

3.3 Data Analysis

We transcribed and coded more than 12 hours of recorded interviews and focus group sessions, each an average of 55 minutes long. We analyzed interviews using thematic analysis [29], a "set of procedures designed to identify and examine themes from textual data in a way that is transparent and credible [28]." We followed the data analysis steps outlined by Guest et al. [28].

One researcher segmented and coded the transcribed interviews into categories and types. Two researchers discussed the relationships that developed from the codebook. Afterward, two researchers identified the themes that emerged from the data. We conducted data analysis concurrently with the data collection and reached theoretical saturation after 34 interviews and focus group sessions, as no new codes emerged from the last data collection session. Appendix B shows the saturation graph depicting the total number of codes after each interview.

3.4 Participants

We recruited 35 participants (33 women and 2 men), aged 19 to 80 years (the mean age was 40 and median was 36). Appendix A provides the demographics of the participants. Participants' occupations included counselor, police officer, daycare worker, cook, barista, event planner, social worker, baker, frontline worker, stay-at-home mother, and student. All participants were survivors, support workers, or both.

4 Results

To better understand survivors' concerns regarding trusting an O-TPRS, we grouped our findings into **technological** and **emotional** concerns. We define *technological* concerns as the issues participants had with using an O-TPRS to report sexual assault. We define *emotional* concerns as the psychological issues participants had with using O-TPRS. Most of the *emotional* concerns are related to issues with the *technology* of the O-TPRS. In the next sections, we illustrate these concerns and explain how the concerns are related. To provide more context, in the rest of the paper, we use SW, SR, and SWSR along with participants' ID to indicate if participants are support workers, survivors, or both respectively.

4.1 Technological Concerns

4.1.1 The insecurity of technology

The insecurity of technology was a concern. Participants found it challenging to trust that the technology would be safe to use in reporting sexual assault incidents. P8-SR, for instance, remarked: *"I wouldn't feel comfortable at all [using an O-TPRS]. I have zero confidence in online. Although I [use the] computer [and], I know the computer, ... I don't know it like hackers do. So, therefore, I would not put any of my information [into an O-TPRS]."*

When comparing the submission of a TPR form to a human versus online, participants trusted humans more. P5-SR, for example, commented: *"I still see [the] human factor is [a] dominant form of communication rather than technology, which can be twisted and broken and is not secured ... Technology to me is not safe because there are so many ways to hack it."*

Because of news of past data breaches, participants assumed that a breach would also happen with an O-TPRS. P6-SR, for instance, remarked on past data breaches: *"[Technology is not] safe. I don't care who says it is; it isn't. [You] just have to listen to the news. The banks have been hacked ... the government's been hacked ... Everybody else [has been hacked]."*

The lack of trust in the internet's security also led to the fear of survivors' losing their confidentiality and privacy. Because of this fear, participants limited the amount of personal information that they shared online. P20-SWSR explained: *"I personally don't put or do anything on the internet that I'm going to be upset about anyone knowing. If I don't want people to see pictures of me with less clothes on, I probably just should not post those. ... So I don't know how I would trust [an O-TPRS] with something that I would be upset about someone seeing."* In their research on trust in e-commerce technology, Araujo and Araujo [3] note that the fear of lack of information privacy is associated with a distrust of technology.

The insecurity of technology led to **anxiety** about using technology to report sexual assault. P7-SWSR, for instance, explained how the use of technology could lead to anxiety: *"I would prefer a paper [TPR] because places that are supposed to be totally secure are being breached. ... And [using technology to report] would give me more anxiety than necessary."* P29-SW also explained: *"[The thought of using an O-TPRS] makes me nervous ... it's kind of like a fear of [the] unknown. I know that going into the [police] station is a lot more vulnerable too, but I have confidence that confidentiality is kept in place due to their legal obligations. I don't fully agree that when things are online that it's completely confidential."*

The possibility of hackers accessing an O-TPRS also leads to **the fear that perpetrators** [49] could see the O-TPR details. Access to such information by the perpetrator could lead to the **re-victimization of the survivor.** P16-SR explained this fear: *"Servers get hacked, and people can see that information. And sometimes there's not anything that you can do to stop that [from happening.] That's what skews me. [Your sexual assault information] can get into the hands of the wrong person."*

4.1.2 Lack of competency with using technology

Unfamiliarity with using any form of technology was another reason participants were not keen on trusting technology. P10-SWSR explained this challenge: *"I wouldn't be comfortable [using an O-TPRS] just because I'm not really comfortable with technology, so I don't see myself downloading a [TPR] app. ... Just when I [decide to report], I would not think of [using] something I am not comfortable with."*

4.1.3 Lack of anonymity assurance

According to participants, with O-TPRS, there was no assurance of anonymity of their personal information. Participants needed a guarantee that the information submitted through an O-TPRS would remain anonymous. They compared the anonymity a P-TPRS provided to that of an O-TPRS. In the P-TPRS, the third-party center representative takes off the cover sheet and sends the anonymized TPR to the police (see Section 2.1 for how the P-TPRS works). Though the O-TPRS also promises the same level of anonymity, participants found it hard to believe that their report would be anonymized. P22-SWSR explained this concern: *"If I go to a hospital and [I] fill out [a P-TPRS], [the nurses] can remove the cover sheet and then give [the anonymized P-TPRS] to the police ... something about that [process] feels safer [than an O-TPRS]. ... If I didn't have to [put] my own information [online] when making a report, then that would be better."*

4.1.4 The traceability of online reporting

There were concerns about the traceability of activities carried out on the internet. Participants believed that activities done on the internet left a lot of traces. Further, participants feared that sensitive sexual assault information submitted online could be traced back to them. P16-SR explained this problem: *"I would be scared to use an app or a website [as an O-TPRS] because ... once [the sexual assault information] is on the internet, it's on the internet. ... Even if you deleted the app, and then [people] go through your iCloud history you can see all the app that's uninstalled and installed. There's a lot of trail that can be traced back [to you] and that would be my number-one concern."*

Participants compared the traceability problem of an O-TPRS to the P-TPRS. P3-SR, for instance, stated: *"I know everything can be traced, so if I send [the sexual assault information online] to the people that are supposedly the third party, that are keeping my confidentiality, there's still a*

trace somehow. But if I write this down [on a P-TPRS], and I hand in this paper, there's no trace at all."

This concern was associated with **the fear that perpetrators could see the O-TPRS.** This *emotional concern* was prominent in the scenarios where the survivors knew the offenders. P22-SWSR explained this challenge: *"In my situation, I know the person that [assaulted me]. It's someone that I see from time to time. If there's some way for the offender to access this [online] form and then [the offender] can check the IP address that it was sent from and then it gets tied back to me, then I'm worried that there's going to be some ... kind of revenge. ... I [have the] fear that somehow [the online report is] going to be tied back to me. And then the person that did [the sexual assault] is going to know [and] get mad."* The issue also leads to the **re-victimization of the survivor.**

4.1.5 The dual use of technology

It was sometimes hard for participants to come to terms with the fact that the technology that is used to aid sexual assault or harassment could be used to reduce the occurrence of such crimes. This challenge sometimes made it difficult for survivors to trust the use of technology in reporting sexual assault: *"[Using technology to reduce sexual assault] is almost like an oxymoron. Because all we hear about is the sexual violence on the internet and people accessing porn on the internet and not as much of the reporting piece and safety."* (P18-SWSR). This disbelief of the participants was understandable, given how much sexual violence is technology facilitated [33–36, 52].

4.1.6 The possibility of false reporting through O-TPRS

An O-TPRS could be misused. A person could submit a false online sexual assault report, or could submit multiple times, thereby reducing the credibility of the platform. Regarding this possibility, P11-SW remarked: *"I could see people wanting a certain level of reassurance that someone didn't just go on [the O-TPRS] and, because they were mad at their ex or something, [submit an O-TPR form]."* This problem was a major concern for the police. P1-SW, who is a police officer, explained: *"I'd be afraid of people misusing [the O-TPRS], either as a prank, kids playing a joke on somebody, or even for malicious reasons. If someone was out to get somebody else, then they could make this [online] third-party report. And if it would go to the police and be reported in the police databank, then there wouldn't really be any other corroborating information, it would just be sort of that mark on the database."* Regarding the possibility of such pranks happening with a P-TPRS, P1-SW commented: *"It's harder to lie to another person than it is on the computer."* While Liu [41] predicted the possibility of false allegations when using technology to report sexual assault, our findings provide empirical evidence that Liu's concerns are shared by TPRS stakeholders.

4.1.7 Lack of trust in apps compared to websites

The type of technology used for the O-TPRS influenced participants' decision to trust the system. Participants were more willing to trust websites than smartphone apps because they believed websites were a more secure option. For instance, P14-SR explained why she would rather use a website: *"Apps are still so new on so many levels, it's so easy to get an app with just one tiny little bug in it and that's [the attacker's] entryway to take all your information."*

Further, participants associated the use of apps with unserious use cases or activities. P34-SW explained: *"My only concern is when I think of an app I tend to think of it as something fun, almost enjoyable ... [For instance, you can say] 'Oh, I have an app to go grocery shopping,' 'Oh I have an app to do my banking,' 'Oh, I have an app to report my sexual assault ...' You see what I mean? [Reporting through an app] takes away a little bit of that seriousness. [It takes away] the severity of [the sexual assault]. So that disturbs me. Whereas [using a website] you can do many different things online. [A website] just seems a bit more appropriate."* For P33-SW, her mental model regarding apps was geared towards using apps for fun activities.

Sometimes using an O-TPRS (either an app or a website) reduced the seriousness of the crime. P10-SWSR explained this concern: *"Reporting sexual assault online could be ... a de-sensitive experience. Currently, you report online for things like breaking into your car. I just feel like the severity of a human right violation being able to be typed [online] maybe can minimize someone's experience."*

Since apps are mostly used on phones, participants were concerned that the safety of the information on the app depends on keeping the phone safe. P14-SR expressed this concern while explaining why she would not use an app: *"[My sexual assault information] is not a personal information I want [on] my phone [because my phone] can be taken from me. ... It just takes one minute for someone to creep your phone, or your phone didn't lock right, or doesn't have a lock. Somebody can hack your phone because you read a [malicious] email on your phone. [For a website, the hackers] have to go directly for the website."* For P14-SR, a compromise of her phone security also meant a compromise of the app.

Using a phone to access the O-TPRS (either through a website or an app) could lead to **unauthorized people having access to the sexual assault information.** If someone sees the information on the phone, that information is no longer anonymous. Such a person could be one's partner or child, or even the perpetrator. P16-SR explained: *"If you had a partner, and they went through your phone and they saw that you had [O-TPRS] opened on your browser or app, and then they go through [the saved report] ... some people live in not so great relationships where there is not a lot [of] trust ... That can put [the survivor] in danger. That's scary for me [because] some women don't have that option to keep their phone."* If

it is the perpetrator who stumbles on this information, this could lead to **re-victimization of the survivor**.

Further, participants thought that seeing an app about sexual assault on one's phone could lead to a survivor's **reliving the experience through constantly seeing the app**. P14-SR explained this *emotional concern*: *"I don't want an app on my phone about my experience. Every time I see it, I am going to think of [the sexual assault incident]."* P10-SWSR further stated: *"Anytime you open your phone, you might see the app and then you just remember that you were assaulted and you have to finish this [sexual assault] application."* The presence of the app on the phone would be a constant reminder to survivors that the sexual assault took place.

4.1.8 The misuse of personal information for targeted advertisement

Information kept online can be misused by the O-TPRS. Because of the common practice of marketers using online information to serve ads, participants were concerned that the O-TPRS could use their personal information for ads. P26-SW expressed this concern and remarked: *"[If the O-TPRS is using my information for ads] I think that's where I would lose comfort in online [TPRS]. [The knowledge] that [my sexual assault information] is somewhere, as a data point to me, and then, suddenly my ads are coming up with 'take self-defense courses,' 'wear modest clothes,' or something. ... I would lose comfort in [the O-TPRS] for sure."*

4.1.9 Lack of control

Participants believed they were more in control when they used P-TPRS. There were concerns because of the errors that could occur when using technology, and participants believed they had no control over any of these errors. P25-SWSR expressed this concern in comparison with P-TPRS: *"If you're sending [a sexual assault report] online, there's always room for technology error [or] the form not going through properly. However, if a person is supported by a counselor or ... [a sexual assault support] agency in doing this, there can be some follow-up by that counselor with the police to say, 'Hey, did you get this third-party report?' ... just to confirm that [the police] did receive [the O-TPR form]."*

4.1.10 Concerns about the unlimited input in UI

While there were many user interface concerns, we report only the concern over *unlimited input*, which appears to have privacy and security repercussions. The information provided by the survivor because of unlimited input could lead to **re-victimization of the survivor** through court proceedings. The O-TPRS provides survivors with unlimited document space and time to type details about the sexual assault incident (see Section 2.2 on how O-TPRS works). However, this format could lead to issues for survivors. P11-SW explained this concern: *"I worry about [the survivor's] inner thoughts being documented in a way that could be used against them in real life. [For instance,] if I was assaulted at 3 [am] and I'd been drugged ... and I thought I had this [O-TPRS], I'm [going to] get this information in right away ... and then I hit send. Nobody else is [there to say], 'Hey, maybe, you need care right now. You need to be [in a] more grounded [59] place before you actually press send.' ... Having some guidance to say, 'You know, the police will ... understand you better when you're in a different spot.' That's my only [concern], because I worry about that information becoming part of some legal document or the public record. I've seen in court how words and things can be spun [against the survivor]."*

P21-SWSR explains this issue further: *"[The input in the O-TPRS] could be used against [the survivor] in a court of law [since the O-TPRS allows survivors] to be adding to [the O-TPR form] for several months after the assault. ... [For instance, you] get a survivor who's at home, feeling bad, and ... she's [going to write] something really horrible blaming herself. [She could say,] 'If I hadn't been at the bar, nothing would have happened,' 'I should kill myself, maybe ... I'll take the children with me' ... and those are the sorts of things women say or think in the middle of the night. But in the depths of depression, that might spill out. And then if this becomes a court case, the defense attorney gets hold of that and he's going say, 'Well look even here, you said it was your fault.' ... I think if people can talk about things over the course of months, it's going to be more [of an] opinion and feeling than factual. And that scares me [about O-TPRS]."*

4.2 Emotional Concerns

Various emotional concerns are related to technological concerns. These emotional concerns are *anxiety, fear of perpetrators seeing the O-TPRS, re-victimization of survivors, unauthorized people having access to the sexual assault information, and reliving the experience through constantly seeing the app*. We discussed these concerns in previous sections. In this section, we discuss emotional concerns that have not previously been addressed.

4.2.1 Lack of human support

Having no human interaction was a major reason participants were not comfortable to trust and use an O-TPRS. Participants believed that online systems lacked empathy, which made it difficult to trust an O-TPRS fully. P15-SWSR highlighted this concern: *"It's draining to fill out [your sexual assault story] on a[n] [online] form rather than conveying the story to a person. ... At least with people, they can [express] empathy, or it's like you're telling it to a person versus a computer screen ... [that's] like talking to a wall."*

In some cases, not having human interaction can lead to **re-traumatization for the survivors.** P25-SWSR, for example,

remarked: *"I think that this [online] form can be traumatizing for people trying to fill this out on their own. ... Just having a support person near them, even if they're not helping them to fill out the form, but they're close by so that if grounding [59] or some crisis support is necessary, there's someone around to do that with that person."*

Human support could be in various forms. Some participants were open to having an online audio or video form of support while filling out an O-TPRS. P22-SWSR explained that *"Having the option on the [O-TPRS] to be able to chat or to call somebody will be great. ... At times like that, questions can be very confusing ... you're disoriented and traumatized and it can be really hard. So knowing that somebody can walk you through it if you're not face to face with somebody ... [that] would be a great asset."* Other participants, however, believed that nothing could replace face-to-face human support. P8-SR, for instance, commented: *"[An O-TPRS is] missing the human link. You need the human link. The one thing that really works is the fact that you're face-to-face with a real person who's exhibiting empathy towards you and is concerned about you and would help you overcome what happened to you. ... I like walking into a place and seeing this empathetic face and then having someone offer me [a tissue] if I'm going to lose it."*

4.2.2 Having no human in the loop

Having a human in the loop was important to prove the legitimacy of the report. Participants who were police officers were concerned about trusting anonymous reports if there was no human involved. P1-SW, for instance, stated: *"An O-TPRS [doesn't have] either the check of a nurse or counselor or something from the social work side. ... Generally when someone's telling a nurse or a counselor something, I put more weight on that as opposed to just an anonymous [report that someone] typed out on their computer and sent it in. ... It's just easier for me to put weight behind it if [the survivor has] actually gone through and spoken to a person face-to-face as opposed to just over the internet."*

5 Discussion

5.1 Limitations

Our sample could have been more balanced and diverse. It had more female (86%) participants, though statistics show that more women experience sexual assault [50, 60]. Most of the participants (86%) were also recruited through sexual assault centers. In addition, the involvement of more than one researcher in the data collection and initial coding would have reduced personal bias. Furthermore, as with any interviews and focus groups, the data were self-reported and may have been affected by a number of systematic biases such as halo effect, social desirability, and acquiescence response bias [21].

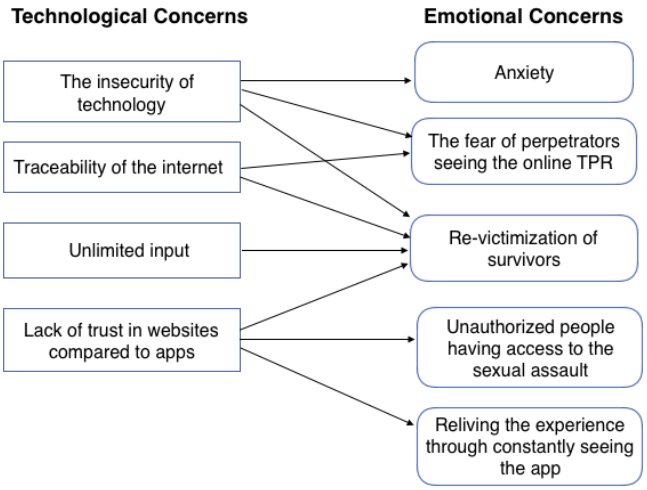


Figure 3: The relationship between technological and emotional concerns

Nonetheless, we believe that the results of our study can serve as a basis for further research on how O-TPRSs can be designed to support survivors of sexual assault.

5.2 Survivors vs. police: balancing their needs

For the sake of clarity, we define privacy and anonymity. Anonymity can be seen as a type of privacy. Privacy and anonymity are related but can be differentiated in some contexts. Webb [64] defines online privacy as the ability to "control who (if anyone) sees what activities you engage in online. In other words, 'they' can see who you are, but not what information or websites you access or seek." The author further defines anonymity as, "when you opt to have your online actions seen, but keep your identity hidden. [This means that] 'they' can see what you do, but not who you are." In line with Webb, we define privacy in an O-TPRS as the ability of the survivor to control who can see that the survivor used an O-TPRS. We define anonymity as the ability of the survivor to make sure that others cannot learn that the survivor has used an O-TPRS, even though others might know that someone used the system. Privacy means knowing a O-TPRS user's *identity* but *not* their *actions* in the system. Anonymity means knowing *actions* of a user in the O-TPRS system but *not* user's *identity*. Table 1 illustrates these definitions.

	Know my actions	**Do not know my actions**
Know my identity	No privacy and no anonymity	Privacy but no anonymity
Do not know my identity	Anonymity but no privacy	Privacy and anonymity

Table 1: Privacy and anonymity of survivors in an O-TPRS.

To understand how privacy and anonymity relate to our

findings, we make the following definitions. We define *identity* as a survivor. We define *action* as using an O-TPRS. We define the actors as the perpetrator, the police, or family and friends that the survivor has chosen not to disclose their sexual assault experience to (assuming the perpetrator doesn't fall into the latter category).

Our findings suggest that the O-TPRS should provide these properties:

Privacy protection from the perpetrator: Even though the *perpetrator* knows the person is a *survivor*, the *perpetrator* must not know that the survivor is using or has used an *O-TPRS*. (See examples in Sections 4.1.1, 4.1.4, 4.1.7, and 4.1.8.)

Anonymity protection from the police: The *police* must not know who the *survivor* is, while the *police* know that a "survivor" filled out an *O-TPRS report.* (See examples in Section 4.1.3.)

Privacy and anonymity protection from others: The survivor's *family and friends* must not know that the person has experienced a sexual assault. In addition, the *family and friends* also must not know that the person used or is using an *O-TPRS.* (See examples in Sections 4.1.1, 4.1.2, 4.1.7, and 4.1.8.)

Because both privacy and anonymity are related, a compromise of one could lead to the compromise of the other. There are many concerns that need to be addressed in designing an O-TPRS. When using an O-TPRS, the anonymous reporting of sexual assault is completed after a survivor submits an O-TPR form to the police (see Figure 2). The two main actors in the O-TPRS are the survivor and the police. The survivor must trust that the O-TPRS has anonymized the O-TPR form before sending it to the police. However, our results suggest that survivors find it difficult to trust that the O-TPRS can preserve their privacy and anonymity (for instance see Sections 4.1.1 and 4.1.4). The police must also trust that the report received from the system is not a false allegation. The police find it challenging to trust that the anonymous reports from the O-TPRS are from survivors (see Section 4.2.2). Therefore, survivors' need for privacy and anonymity is pitted against the police's (1) need to know the identity of the survivor and (2) the concern that anonymity could increase false reporting. The challenge for the O-TPRS designers is that without finding a solution that can satisfy these two stakeholders, it is unlikely that either will trust an O-TPRS. We discuss these concerns in depth in the following sections and explain how they affect survivors and the police.

5.3 Trust of survivors in an O-TPRS

5.3.1 Before sending the O-TPR form to the police

An O-TPRS requires both privacy and anonymity. Survivors want to send anonymous reports to the police. That means that the police should be able to identify that they have received a report from a survivor without being able to trace the report to the person who submitted it. Survivors also want to maintain their privacy by having control over who sees that they are using an O-TPRS.

The survivor should be able to trust that unauthorized people will not discover that the survivor is using or has ever used an O-TPRS. The O-TPRS has to be designed so it is not obvious on the survivor's device. Further, it should be unknown to the perpetrator that the O-TPRS will report sexual assault. This requirement could be achieved by using a pseudonym for the O-TPRS app or website; however, this design could lead to usability issues for the survivors because survivors would have to remember the pseudonym for the app.

Several proposals for addressing this problem have been put forward. For instance, for survivors of domestic violence, Arief et al. [4] suggest the design of an app that could automatically erase the parts of the survivor's browser history that shows that the survivor searched for online help resources for domestic violence. The authors suggest that the app could be "hidden behind an innocent front end, such as a game app or an image gallery app." According to the authors, this design will prevent the perpetrator from recognizing that the app erases the survivor's history. A similar design could also be useful for an O-TPRS; however, such a solution will be ineffective if the perpetrator knows the pseudonym of the app. For instance, in their work on how technology aids perpetrators in stalking intimate partner violence victims, Freed et al. [25] outline many ways in which perpetrators can gain access to survivors' phones. Some ways include forcefully compelling survivors to unlock their phones, or strictly monitoring their activities. If a sexual assault survivor lives in an unconducive situation, (for instance, Section 4.1.7 and P22-SWSR in Section 4.1.4), having an O-TPRS app on their phones, even in disguise, may bring harm to the survivor.

Survivors could also forget to close the O-TPRS, or the perpetrator might see them filling out the O-TPR form. The O-TPRS should be able to provide ways by which a survivor's privacy is protected if they leave their phone or computer unattended while filling out the form (see Section 4.1.7). The O-TPRS would also need to provide a way of easy escape on the app or the website if the perpetrator walks in on the survivor while they are filling the O-TPR form. Some sensitive websites have an escape button provided. These buttons allow people to exit the site quickly if they feel uncomfortable while reading the website's content or if it becomes unsafe to continue reading (for instance see [2]). Such designs could be looked into for O-TPRS apps and websites. Research needs to be done to determine how best such escape buttons could be placed on an O-TPRS and if they will be as effective.

It could be problematic for survivors if perpetrators know that an O-TPRS app was downloaded or the website was visited. By default, computers and phones save the history that an app was downloaded, or a website was accessed.

This default setting is a challenge for survivors (see Section 4.1.4). If the perpetrator see this information, it could cause re-victimization of the survivor. For survivors of domestic violence, Arief et al. [4] suggest an app that automatically erases the survivor's web history. However, in abusive situations where the perpetrators check the survivors' web and installation history, we believe such a design could lead to more problems for the survivor. This problem could arise because the perpetrator may suspect that the survivors are trying to hide their activities by erasing their history.

Some technological solutions help people to surf the internet anonymously. For instance, to browse the web anonymously, people could use the incognito mode of their browser [15], or they could also make use of a Tor browser [8]. An option to hide survivors' online history could be for survivors to access the O-TPRS only in incognito mode or through a Tor browser. However, these designs require a certain level of familiarity with technology, and survivors may not find such designs usable (see Section 4.1.2). Further, incognito mode won't help in a scenario when the perpetrator has installed a key logger or is eavesdropping the traffic between the survivor's computer and the internet [1]. In addition, the Tor network is linked with so many illegal activities such as human trafficking and illegal sex trade [38], and as cited by P18-SWSR in Section 4.1.5, it may be hard for survivors to trust that such systems can help reduce sexual assault.

Another option could be the inclusion of a process to verify a survivor's identity on an O-TPRS. This verification process could be done through an authentication system. Depending on the name supplied to the O-TPRS system, this design may not provide privacy because the presence of the app or website on a person's device may reveal to others that the person is a survivor. An authentication system may not fully protect the survivor's anonymity because whatever option is used to verify the survivor's identity could be an identifying factor of the survivor. This identifying factor could be the survivor's email address or biometric information. If a password system is used, this design may be problematic if survivors forget their passwords. If the survivor receives email to reset their login details, the perpetrators could see emails or email notifications, which compromises the survivor's privacy and anonymity. Further, if an authentication system is used, the O-TPRS would have to ensure that the police cannot access such identifying information without the survivors' consent.

5.3.2 After survivors send the O-TPR form to the police

After the O-TPR form has been sent to the police, the survivor's anonymity and privacy still need to be protected (see Section 4.1.1). Further, unauthorized individuals should be unable to discover that the survivor sent the information to the police (for instance, see Section 4.1.4). Protecting survivors' anonymity can be achieved by having security in place. Such a system will need high level of security, which is hard to afford especially for small organizations looking into developing O-TPRSs [26]. It is also difficult to measure how much security is good enough to protect a system. As argued by Hurlburt [37], security may never be good enough. The author explain further that for a secure system to be impenetrable by anyone, the system can probably not be connected to the internet, and humans will have to be taken out of the loop [37]. The O-TPRS will hold very sensitive information from survivors. Therefore, whatever security measures the system employs, such measures should have a low likelihood of being breached. Any compromise of the O-TPRS could lead to distrust of the system and, even worse, further victimization of survivors (see Section 4.1.1, 4.1.7). The O-TPRS operator will also have to convince survivors that such measures are good enough to protect their information.

5.4 The police trusting O-TPRS reports

The police want to be able to verify that the person who sends an O-TPR form is a survivor (see Sections 4.1.6, 4.2.2). However, it is unclear how this requirement can be achieved without violating the survivor's anonymity. One of the purposes of using an O-TPRS is to keep survivors anonymous to the police (see Section 2 and Table 1). Verifying the survivor's identity would violate their anonymity. In the P-TPRS, the presence of a representative at the TPR center may provide some assurance that the person making a report is a survivor (see Section 2.1). The police may trust that the report is valid because they trust the representative [12, 45].

Several solutions exist that provide verification of system users. Examples of such solutions include the completely automated public Turing test to tell computers and humans apart (CAPTCHA) [63]. However, current solutions such as CAPTCHA don't solve this problem, as CAPTCHA is designed to check if the user of a system is a human or not. CAPTCHA cannot verify whether the user of O-TPRS is a survivor or someone making a false report.

The cost of making a false report is low with O-TPRS. As explained in Section 2, a person is identified as a serial offender if three different survivors report them as an offender. Both O-TPRS and P-TPRS carry a possibility of false reporting. Nevertheless, the cost to a person who wants to create multiple false claims with P-TPRS is much higher. Such a person would have to convince two other people to walk into a sexual assault center at various times and accuse the same person of assault. With O-TPRS, the cost of making such false reports is smaller. A person could simply download the O-TPRS app or use the website and get two others to do the same. Alternatively, a person could make a report two more times from different accounts, known in distributed systems as Sybil Attack [22].

O-TPRS could lead to an increase in false reporting. Although sexual assault is an underreported crime, reducing the current barriers to reporting might lead to an increase in re-

porting. In addition, as explained by P1-SW in Section 4.1.6, the use of O-TPRSs might also lead to an increase in false reporting. This is a major challenge, as this problem might reduce the credibility of real reports made through O-TPRS. This challenge is similar to swatting attacks where swatters make false reports to the police about an ongoing crime [6]. Similarly, in an O-TPRS, the possibility of false reporting could reduce the credibility of real reports.

A solution used to mitigate a similar challenge in other systems is the use of a password-based authentication to identify users uniquely. As discussed earlier, this solution, however effective, could reduce the anonymity of O-TPRS users. Further, users could easily create multiple email addresses to make false reports. It is unclear what measures can be put in place to deter illegitimate users while maintaining ease of use for legitimate users to report their sexual assault. Future research could investigate how O-TPRSs can implement a form of verification or CAPTCHA system for survivors. This system should be able to verify that the person reporting is a survivor. In addition, the system should not introduce the additional bottleneck of having human verification or reducing survivors' anonymity.

However, it should be noted that the motivation for making multiple or false reports seems weak. Although any report made will be registered in a database, and three reports would trigger follow-up from the police, as explained in Section 2, that follow-up would simply be an invitation to make a formal report, which the survivor was free to do at any time anyway.

5.5 The provision of human support

The importance of human support when reporting a sexual assault was discussed by many participants (see Sections 4.2.1 and 4.2.2). Participants explained that when using an O-TPRS, it would be important for survivors to have humans in the process for two reasons: 1. To ensure that the survivor receives the support needed to complete and submit the form to the police. Many participants wanted human support when filling out an O-TPR. It is unclear if this finding is primarily because most of our participants were already receiving support from sexual assault centers and therefore could not imagine using an O-TPRS without a support worker. It may be important to carry out further research to investigate if survivors who do not receive support from sexual assault centers will be comfortable using an O-TPRS without human support. 2. To ensure that the survivor is in the right mental state to make a report of a sexual assault [59]. For instance, sometimes survivors deal with flashbacks or disassociation from the present moment and need support before, during, and after making a report [59].

To provide support for survivors, an option could be to provide human support via a video or audio call on the O-TPRS. While some participants thought this option would be useful, others suggested they would need face-to-face interaction (see Section 4.2.1). This design also doesn't address the problem of verifying that the survivor is ready to make a report [59]. It would be difficult for a human to verify over a video or audio call that a survivor was in the right mental state to make a report. This verification is important because on the O-TPRS, the survivor could write about their feelings rather than limiting the input to the factual details about the assault, and these details might be used against the survivor in the court of law (see Section 4.1.10). Further research is necessary to identify unique solutions to ensure that the survivor is ready, before submitting a report to the O-TPRS.

5.6 Balancing unlimited and limited input

There should be a balance between providing the survivor with too little or too much time and document space to complete a report. Too much time and document space in the O-TPRS could result in a survivor providing details that could be used against them (see Section 4.1.10). Implementing a document space limit on the O-TPRS may be helpful, however more research needs to be done to identify how much space is too much or too little and how such restraints may affect survivors' willingness to use the O-TPRS. Further, implementing a time limit could defeat the purpose of letting survivors complete an O-TPR form at their own convenience.

6 Conclusion

Our paper presents privacy and security challenges in designing an O-TPRS. It introduces many questions that need to be answered in order for survivors and police to trust and use an O-TPRS. Our research serves as a starting point towards designing O-TPRSs to increase sexual assault reporting and the arrest of perpetrators. We presented our findings to Vesta, and the organization is taking this report into consideration in the development of their O-TPRS. We hope these results can start a discourse in the research community and lead to solutions for designing effective online reporting systems for sexual assault survivors.

7 Acknowledgements

This research has been supported by the funding from Vesta Social Innovation Technologies Inc. and Mitacs Accelerate program. We thank members of the sexual assaults centers and all participants who were involved in the study. We thank members of the UBC Laboratory for Education and Research in Secure Systems Engineering (LERSSE) who provided their feedback on the reported research and the earlier versions of the paper. We thank our anonymous reviewers for all the feedback and suggestions they provided for improving the paper. Stylistic and copy editing by Eva van Emden helped to improve readability of this paper.

References

[1] Ruba Abu-Salma and Benjamin Livshits. Evaluating the end-user experience of private browsing mode. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[2] ACCESS. Sexual abuse. https://www.assaultcarecenter.org/en/sexual_abuse/, 2020. Accessed: 2019-02-26.

[3] Ildemaro Araujo and Iván Araujo. Developing trust in internet commerce. In *Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research*, pages 1–15, 2003.

[4] Budi Arief, Kovila PL Coopamootoo, Martin Emms, and Aad van Moorsel. Sensible privacy: How we can protect domestic violence survivors without facilitating misuse. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, pages 201–204, 2014.

[5] J Elin Bahner, Anke-Dorothea Hüper, and Dietrich Manzey. Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, 66(9):688–699, 2008.

[6] Laura-Kate Bernstein. Investigating and prosecuting swatting crimes. *US Att'ys Bull.*, 64:51, 2016.

[7] Shannon Brennan and Andrea Taylor-Butts. *Sexual assault in Canada, 2004 and 2007*. Canadian Centre for Justice Statistics Ottawa, Ontario: Statistics Canada, 2008.

[8] Tor Browser. Defend yourself. protect yourself against tracking, surveillance, and censorship. https://www.torproject.org/download/, 2020. Accessed: 2019-02-27.

[9] Clayton M Bullock and Mace Beckson. Male victims of sexual assault: Phenomenology, psychology, physiology. *Journal of the American Academy of Psychiatry and the Law Online*, 39(2):197–205, 2011.

[10] British Columbia Canada. British Columbia third party reporting protocol. https://endingviolence.org/wp-content/uploads/2019/10/TPR-Guidebook-2.0-July-2019.pdf, 2019. Accessed: 2019-02-26.

[11] British Columbia Canada. Third party reporting for victims of sexual offences. https://www2.gov.bc.ca/gov/content/justice/criminal-justice/bcs-criminal-justice-system/reporting-a-crime/victim-or-witness-to-crime/third-party-reporting-for-victims-of-sexual-offences, 2020. Accessed: 2019-02-26.

[12] Lemuria Carter and France Bélanger. The utilization of e-government services: Citizen trust, innovation and acceptance factors. *Information systems journal*, 15(1):5–25, 2005.

[13] CBC. RCMP looks to expand third-party reporting for sexual assault cases. https://www.cbc.ca/news/politics/rcmp-sexual-assault-reporting-1.4402828, 2017. Accessed: 2019-06-27.

[14] Yingyu Chen and Sarah E Ullman. Women's reporting of sexual and physical assaults to police in the national violence against women survey. *Violence Against Women*, 16(3):262–279, 2010.

[15] Google Chrome. Browse in private. https://support.google.com/chrome/answer/95464?co=GENIE.Platform%3DDesktop&hl=en, 2020. Accessed: 2019-02-27.

[16] Deborah Cohen and Benjamin Crabtree. Qualitative research guidelines project. http://www.qualres.org/, 2006.

[17] Adam Cotter. *Sexual misconduct in the Canadian Armed Forces, 2016*. Statistics Canada, 2016.

[18] Adam Cotter and Pascale Beaupré. Police-reported sexual offences against children and youth in canada, 2012. *Juristat: Canadian Centre for Justice Statistics*, page 1, 2014.

[19] Danielle Couch, Pranee Liamputtong, and Marian Pitts. What are the real and perceived risks and dangers of online dating? Perspectives from online daters: Health risks in the media. *Health, Risk & Society*, 14(7-8):697–714, 2012.

[20] Beata Cybulska. Sexual assault: Key issues. *Journal of the Royal Society of Medicine*, 100(7):321–324, 2007.

[21] Diane Dodd-McCue and Alexander Tartaglia. Self-report response bias: Learning how to live with its diagnosis in chaplaincy research. *Chaplaincy Today*, 26(1):2–8, 2010.

[22] John R Douceur. The Sybil attack. In *International workshop on peer-to-peer systems*, pages 251–260. Springer, 2002.

[23] Sarah M Flanagan, Sheila Greenfield, Jane Coad, and Susan Neilson. An exploration of the data collection methods utilised with children, teenagers and young people (ctyps). *BMC research notes*, 8(1):61, 2015.

[24] Sarah G Forrestal, Angela Valdovinos D'Angelo, Lisa Klein Vogel, et al. Considerations for and lessons learned from online, synchronous focus groups. *Survey Practice*, 8(2):1–8, 2015.

[25] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. "A stalker's paradise" How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[26] Michael P Gallaher, Brent R Rowe, Alex V Rogozhin, and Albert N Link. Economic analysis of cyber security. Technical report, Research Triangle Inst (RTI) Research Triangle Park NC, 2006.

[27] David Gefen, Elena Karahanna, and Detmar W Straub. Trust and TAM in online shopping: An integrated model. *MIS quarterly*, 27(1):51–90, 2003.

[28] Gregory Guest, Kathleen M MacQueen, and Emily E Namey. *Applied thematic analysis*. Sage Publications, 2011.

[29] Gregory Guest, Kathleen M MacQueen, and Emily E Namey. Introduction to applied thematic analysis. *Applied thematic analysis*, 3:20, 2012.

[30] Karin Hammarberg, Maggie Kirkman, and Sheryl de Lacey. Qualitative research methods: When to use them and how to judge them. *Human reproduction*, 31(3):498–501, 2016.

[31] Patricia L Hardré. When, how, and why do we trust technology too much? In *Emotions, Technology, and Behaviors*, pages 85–106. Elsevier, 2016.

[32] Gert Helgesson, Mats G Hansson, Johnny Ludvigsson, and Ulrica Swartling. Practical matters, rather than lack of trust, motivate non-participation in a long-term cohort trial. *Pediatric diabetes*, 10(6):408–412, 2009.

[33] Nicola Henry and Anastasia Powell. The dark side of the virtual world. In *Preventing Sexual Violence*, pages 84–104. Springer, 2014.

[34] Nicola Henry and Anastasia Powell. Beyond the 'sext': Technology-facilitated sexual violence and harassment against adult women. *Australian & New Zealand Journal of Criminology*, 48(1):104–118, 2015.

[35] Nicola Henry and Anastasia Powell. Embodied harms: Gender, shame, and technology-facilitated sexual violence. *Violence against women*, 21(6):758–779, 2015.

[36] Nicola Henry and Anastasia Powell. Sexual violence in the digital age: The scope and limits of criminal law. *Social & Legal Studies*, 25(4):397–418, 2016.

[37] George Hurlburt. "Good enough" security: The best we'll ever have. *Computer*, 49(7):98–101, 2016.

[38] Eric Jardine. *The Dark Web dilemma: Tor, anonymity and online policing*. Centre for International Governance Innovation and Chatham House, 2015.

[39] Holly Johnson. Why doesn't she just report it? Apprehensions and contradictions for women who report sexual violence to the police. *Canadian Journal of Women and the Law*, 29(1):36–59, 2017.

[40] West Coast LEAF. We are here: Women's experiences of the barriers to reporting sexual assault. http://www.westcoastleaf.org/our-publications/we-are-here-womens-experiences-of-the-barriers-to-reporting-sexual-assault/, 2018. Accessed: 2019-06-27.

[41] Heidi Liu. When whispers enter the cloud: Evaluating technology to prevent and report sexual assault. *Harv. JL & Tech.*, 31:939, 2017.

[42] CH Logie, Ramona Alaggia, and Marie-Jolie Rwigema. A social ecological approach to understanding correlates of lifetime sexual assault among sexual minority women in toronto, canada: Results from a cross-sectional internet-based survey. *Health education research*, 29(4):671–682, 2014.

[43] Helen Luce, Sarina Schrager, and Valerie Gilchrist. Sexual assault of women. *American family physician*, 81(4):489–495, 2010.

[44] Margaret J McGregor, Ellen Wiebe, Stephen A Marion, and Cathy Livingstone. Why don't more women report sexual assault to the police? *CMAJ*, 162(5):659–660, 2000.

[45] D Harrison McKnight, Michelle Carter, Jason Bennett Thatcher, and Paul F Clay. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information systems (TMIS)*, 2(2):12–32, 2011.

[46] Enid Montague and Onur Asan. Trust in technology-mediated collaborative health encounters: Constructing trust in passive user interactions with technologies. *Ergonomics*, 55(7):752–761, 2012.

[47] David L Morgan. *Focus groups as qualitative research*, volume 16. Sage publications, 1996.

[48] Federal Bureau of Investigation. National incident-based reporting system. https://www.fbi.gov/services/cjis/ucr/nibrs, 2013.

[49] Media Kit on Sexual Assault. Media kit on sexual assault: Perpetrators. https://www.inspq.qc.ca/en/sexual-assault/understanding/perpetrators, 2020. Accessed: 2019-02-26.

[50] Community Ontario Ministry of Children and Social Services. Statistics: Sexual violence. http://www.women.gov.on.ca/owd/english/ending-violence/sexual_violence.shtml, 2020. Accessed: 2019-02-27.

[51] Samuel Perreault. Criminal victimization in canada, 2014. *Juristat*, 35(1):1–43, 2015.

[52] Anastasia Powell and Nicola Henry. Technology-facilitated sexual violence victimization: Results from an online survey of australian adults. *Journal of interpersonal violence*, 34(17):3637–3665, 2019.

[53] Martine B Powell and Rita Cauchi. Victims' perceptions of a new model of sexual assault investigation adopted by victoria police. *Police practice and research*, 14(3):228–241, 2013.

[54] Andrea Quinlan. Suspect survivors: Police investigation practices in sexual assault cases in ontario, canada. *Women & Criminal Justice*, 26(4):301–318, 2016.

[55] Brian A Reaves. Felony defendants in large urban counties, 2009-statistical tables. *Washington, DC: US Department of Justice*, 2013.

[56] Cristine Rotenberg. From arrest to conviction: Court outcomes of police-reported sexual assaults in canada, 2009 to 2014. *Juristat: Canadian Centre for Justice Statistics*, pages 1–57, 2017.

[57] Cristine Rotenberg. Police-reported sexual assaults in canada, 2009 to 2014: A statistical profile. *Juristat: Canadian Centre for Justice Statistics*, 2017.

[58] Emily F Rothman, Deinera Exner, and Allyson L Baughman. The prevalence of sexual assault against people who identify as gay, lesbian, or bisexual in the united states: A systematic review. *Trauma, Violence, & Abuse*, 12(2):55–66, 2011.

[59] SARSAS. Coping with flashbacks. https://www.sarsas.org.uk/grounding/, 2020. Accessed: 2019-02-26.

[60] Sharon G Smith, Xinjian Zhang, Kathleen C Basile, Melissa T Merrick, Jing Wang, Marcie-jo Kresnow, and Jieru Chen. *The national intimate partner and sexual violence survey: 2015 data brief–updated release*. National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, 2018.

[61] Christina Underhill and Murrey G Olmsted. An experimental comparison of computer-mediated and face-to-face focus groups. *Social Science Computer Review*, 21(4):506–512, 2003.

[62] VESTA. Vesta social innovation technologies. https://www.vestasit.com, 2020. Accessed: 2019-02-27.

[63] Luis Von Ahn, Manuel Blum, Nicholas J Hopper, and John Langford. CAPTCHA: Using hard AI problems for security. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pages 294–311. Springer, 2003.

[64] James Webb. Anonymity vs privacy vs security understanding the shades of safety online. https://highspeedexperts.com/online-security-privacy/anonymity-vs-privacy-vs-security/, 2020. Accessed: 2019-02-26.

[65] Paul Willis. Learning to labour (London, Saxon House). *WillisLearning to Labour1977*, 1977.

[66] Jie Xu, Kim Le, Annika Deitermann, and Enid Montague. How different types of users develop trust in technology: A qualitative analysis of the antecedents of active and passive user trust in a shared technology. *Applied ergonomics*, 45(6):1495–1503, 2014.

Appendices

A Participants' Demographics

ID	Age	Gender	Survivor/Support Worker	Interview/Focus Group	Educational Level
P1	36	M	SW	I	Bachelor's
P2	63	F	SR	F	Bachelor's
P3	48	F	SR	F	College
P4	33	F	SWSR	F	Bachelor's
P5	67	F	SR	F	Bachelor's
P6	80	F	SR	F	College
P7	36	F	SWSR	F	College
P8	74	F	SR	F	High school
P9	60	F	SR	F	High school
P10	25	F	SWSR	I	College
P11	44	F	SW	I	Master's
P12	52	F	SWSR	F	MBA
P13	27	F	SR	F	High school
P14	22	F	SR	F	High school
P15	24	F	SWSR	I	Master's
P16	19	F	SR	F	High school
P17	19	F	SR	F	High school
P18	47	F	SWSR	F	College
P19	46	F	SWSR	F	Bachelor's
P20	20	F	SWSR	F	College
P21	63	F	SWSR	F	Bachelor's
P22	21	F	SWSR	I	College
P23	31	F	SR	I	College
P24	19	F	SWSR	F	Bachelor's
P25	29	F	SWSR	F	Bachelor's
P26	39	F	SW	F	Bachelor's
P27	51	M	SW	I	Bachelor's
P28	51	F	SWSR	I	College
P29	26	F	SW	F	Bachelor's
P30	37	F	SW	F	College
P31	62	F	SW	F	College
P32	35	F	SW	F	Master's
P33	22	F	SW	F	High school
P34	49	F	SW	F	Bachelor's
P35	26	F	SW	F	Bachelor's

Table 2: Demographics of participants. SR, SW, I, and F represent survivor, support worker, interview, and focus group, respectively.

B Saturation Graph

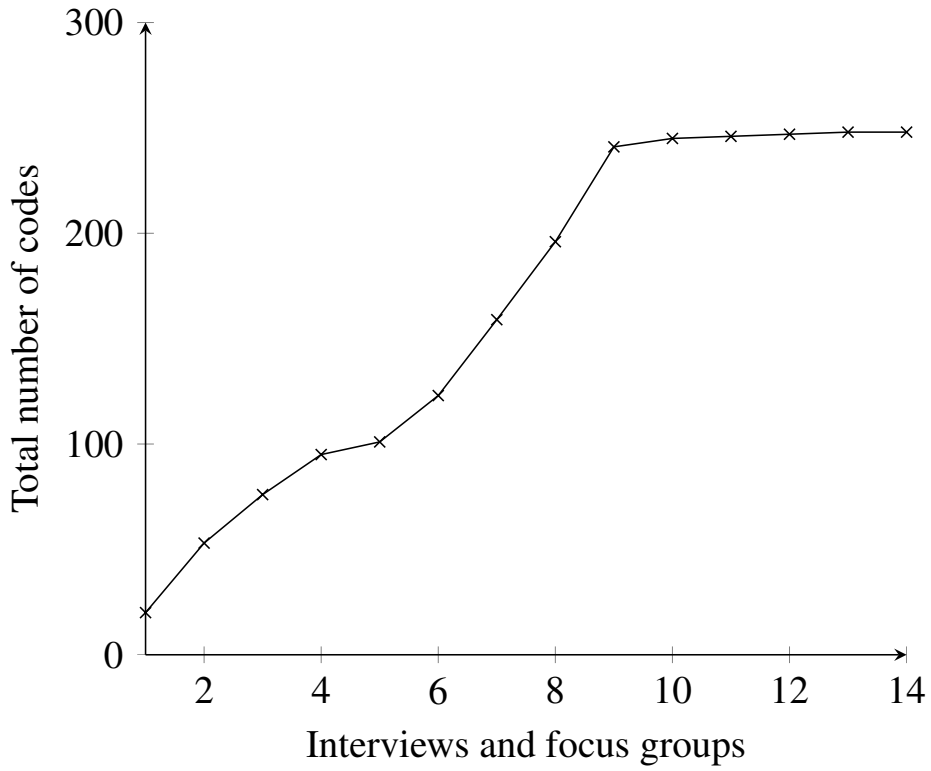


Figure 4: Number of codes after interviewing each participant

C Questions from the P-TPRS shown to Participants from Page 2 and 3

1. Date of Assault
2. Time of Assault
3. Location of Assault
4. Description of Complainant:
    - Male
    - Female
5. Age
6. Height
7. Weight
8. Build
9. Hair Colour
10. Style
11. Length

(A) Offender's Name: (if known)

(B) Offender's Address:

(C) Description of Offender :

- Male
- Female
- Colour
- Race
- Age
- Height
- Weight
- Build
- Hair Colour
- Style
- Length
- Facial Features
- Facial Hair
- Complexion
- Eye Colour
- Glasses
- Circumcised
- Scars/Tattoos/Birthmarks Etc.
- Clothing Worn at Time of Sexual Assault
- Distinguishing Characteristics

(a) Vehicle Information (Licence #, Make, Model, Colour, Damage, Anything Distinguishable)

(b) Details of Offense: (EXPLAIN IN COMPLAINANT'S OWN WORDS)

D Sample of the O-TPRS Prototype Shown to Participants



Figure 5: O-TPRS Homepage

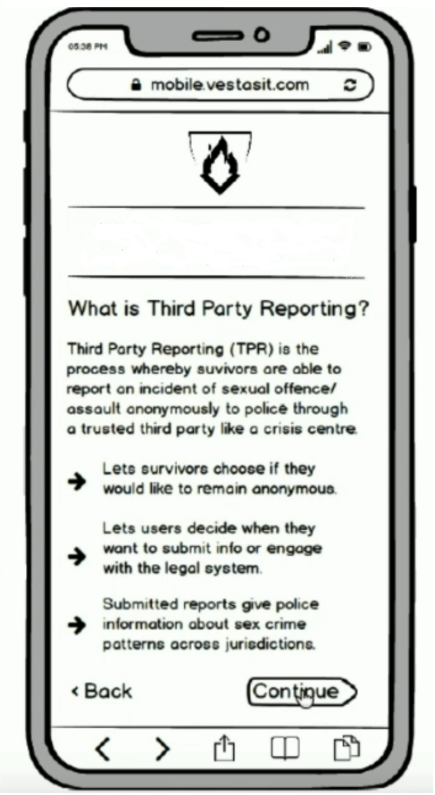


Figure 6: Introduction to TPRS



Figure 7: O-TPR Form Page 1



Figure 8: O-TPR Form Page 2

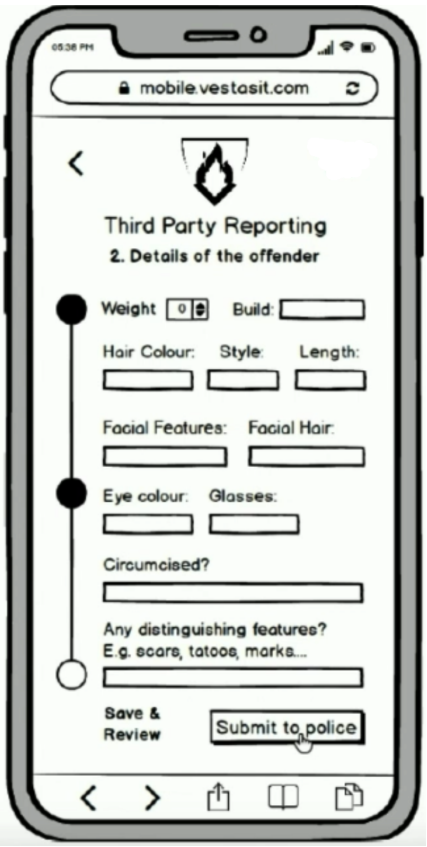


Figure 9: O-TPR Form Page 3

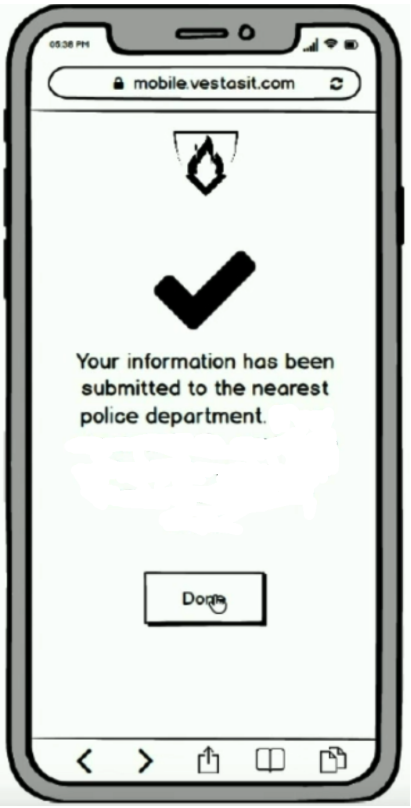


Figure 10: Submission Page

Replication: On the Ecological Validity of Online Security Developer Studies: Exploring Deception in a Password-Storage Study with Freelancers

Anastasia Danilova
University of Bonn
danilova@cs.uni-bonn.de

Alena Naiakshina
University of Bonn
naiakshi@cs.uni-bonn.de

Johanna Deuter
University of Bonn
johannadeuter@posteo.de

Matthew Smith
University of Bonn, FKIE Fraunhofer
smith@cs.uni-bonn.de

Abstract

Recruiting professional developers for studies can be challenging and one major concern for studies examining security development issues is their ecological validity—does the study adequately reflect the real world? Naiakshina et al. [28] examined the ecological validity of a password storage study conducted with students [29, 30] by hiring freelancers from Freelancer.com. In the hope of increasing the ecologically validity, Naiakshina et al. used a deception study design wherein freelance developers were hired for a regular job using a company front created for the study, instead of openly telling the freelancers that they were taking part in a study. Based on their results, Naiakshina et al. propose the use of online freelancers to be examined further, to supplement other recruitment channels such as CS students and GitHub users. The deception in their study was used with the aim that results would reflect the real work of online freelancers. However, deception needs to be used with careful consideration, which can entail additional study design work and negotiations with ethical oversight bodies. In this paper, we take a closer look at the deception used in Naiakshina et al.'s study. Therefore, we replicate Naiakshina et al.'s work but announce and run it as a study on Freelancer.com. Our findings suggest that for this password storage study deception did not have a large effect and the open recruitment without deception was a viable recruitment method.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

1 Introduction

Security is an issue many software developers struggle [2, 5, 11, 28–30, 47]. User studies with developers are a useful tool to examine misconceptions and issues of developers when faced with security issues. While human computer interaction (HCI) research has made great progress in investigating the security behavior of end users, it still lacks methodological research for the human factor of *software developers* [4]. One major issue with software developers is their limited availability as subjects for research studies. Professional developers often cannot set aside time, may not be locally available or have hourly rates that researchers cannot afford [3–5, 24, 25, 38, 49]. An option to recruit professionals for studies is cooperating with companies as can be seen in [13, 15]. However, it can be difficult to find companies willing to join such studies. Reasons we have encountered are, that companies are hesitant to allocate time of their developers to a study, both for time and cost reasons, as well as worries about disclosure of information about their company as part of the publication process. Therefore, many previous security developer studies were conducted with computer science (CS) students [3, 8, 29–31, 49] or developers recruited on GitHub [5, 20, 35, 47, 48]. To supplement these recruitment options, Naiakshina et al. [28] proposed to use platforms such as *Freelancer.com* for developer recruitment.

In 2017 and 2018, Naiakshina et al. [29, 30] conducted a qualitative and a quantitative password-storage study with CS students to provide more insights into developer's security behavior. The participants were asked to complete the user registration functionally of a university social networking platform. Half the participants were *prompted* for secure password storage in the task, while the other half were merely told the study is about API usability (*non-prompted*). The results showed that not a single participant stored user passwords securely without being prompted. Some students, however, noted that they would have saved the user passwords securely if they had been working on a project for a real company. In order to find out whether the previous results were a study

	CS Students [29, 30]	Freelancers [28]	This Study: Replication of [28] with Freelancers
Independent Variables (IV)	IV1: Security prompting (yes/no) IV2: Framework (JSF/Spring)	IV1: Security prompting (yes/no) IV2: Payment (100/200 euros)	IV1: Security prompting (yes/no) IV2: Payment (100/200 euros) IV3: Framework (JSF/Spring)
Study Context and Task	University researchers University social networking platform	Start-up Sports photo social networking platform	University researchers Sports photo social networking platform
Study Deception	No	Yes	No
Recruitment	University	Individually on Freelancer.com	Project on Freelancer.com
Post Security Request	No	SecRequest-P if plain text submission	SecRequest-P if plain text submission SecRequest-G if not implemented industry security standards

Table 1: Overview of password-storage studies for most relevant aspects.

artifact, Naiakshina et al. conducted a follow-up study with freelancers [28], which they did not announce as an academic study but as a real project. In the corresponding pilot study, the authors found that owing to the university context of task from the previous study, freelancers believed that they were working on university homework and also did not implement any security. To avoid this potential study bias, the authors invested additional study design work to provide a more realistic scenario to make the freelancers believe they were working for a real company and the code would be used in the wild. First, they removed the university social network context and posed as a start-up that had just lost a developer in their team and was searching for a new one. The job freelancers were asked to complete was to create code for a social networking platform for a sports photo sharing website. To make the scenario more believable, the researchers created a fake multi-page online web presence for the start-up. Second, the authors had to manually contact freelancers individually, instead of advertising the job on the service, and letting the freelancers apply for the job. The latter would have made recruitment easier, however, the Freelancer.com platform shows who has been hired for projects advertised in this way and pilot studies showed high dropout rates. This was due to freelancers being confused and suspicious about a single job being given to many freelancers. However, using the deception study design, Naiakshina et al. concluded that the CS students and the freelancers behaved similarly with regard to their password storage practices in this particular study setup and thus, suggested Freelancer.com as a promising platform for developer recruitment which warrants further examination.

In this paper, we specifically want to examine the use of deception in the context of this study. Deception can be a useful tool, if disclosing the study purpose would lead to a significant change in behavior in the participants, e.g., Naiakshina et al. were concerned that freelancers would not implement security in the same way in a study setting as they would for a real customer. However, deception should be used only after careful consideration. Three issues are relevant in our context 1) One of the fundamental principles of human subjects research is *informed consent* and deception can impact this principle. 2) If deception is used, participants who take part in multiple studies or have heard of deceptive studies might second guess what the study is about, potentially affecting their behavior in an unknown manner. 3) Deception studies can require additional study design work, e.g., in the case of Naiakshina et al. the creation of a fake company web presence and more elaborate scenarios.

In order to gain more insights into study methodology and ecological validity of developer studies, we replicated the freelancer study of Naiakshina et al. [28] except for the use of deception. While the study task was kept the same, we created a new profile on Freelancer.com, where we introduced ourselves as a university group conducting scientific research and the job as a study. Due to this we were able to post and manage our study as a single project and manage all freelancers from there, reducing the additional study design work needed to run the study.

In addition to examining the effect of deception, we add further insights into the effect of different application programming interfaces (API) offering cryptographic libraries. For this we replicated the comparison of two frameworks JSF and Spring, as examined in the studies [29, 30],

Finally, we investigated whether providing participants with password storage guidelines has an effect on the security of the submitted solutions. Table 1 provides an overview of the previous lab studies with computer science students [29, 30], the freelancer study [28], and the present study.

2 Related Work

There is a wide variety of studies investigating end-user password creation, password creation rules and their effects and password usage [11, 16, 22, 23, 26, 34, 36, 37, 39, 42–45]. Although developers play an important role in secure password storage too, rather little work has been done with them. To offer more insights into developers' security behavior, Naiakshina et al. conducted password-storage studies with CS students and freelancers [28–30]. A detailed description of the papers is available in the Introduction section. In order to explore the necessity of study deception as done in [28], we replicated the freelancer study of Naiakshina et al. by announcing our study context to participants. Section 2.1 outlines the work conducted on the ecological validity of study

design. Section 2.2 provides an overview of security developer studies in general and in the context of password storage. Finally, Section 2.3 summarizes studies conducted focusing on freelancers.

2.1 Ecological Validity

Acar et al. [4] argued that the examination of ecological validity–whether studies reflect the real world–is of great importance. Most work in this field was primarily conducted with end users. For example, Redmiles et al. [33] compared field data with survey reported data about software updates initiated with end users. They found that self-reported data in some cases varied from field data.

Furthermore, Wash et al. [45] compared 134 self-reports of end users' password behavior with their actual behavior and found only a weak correlation of self-reported intentions with reality. In [46], Wash et al. aimed at finding out which security behaviors were accurately self-reported by the end users. For this, they collected survey responses and behavior data from 122 participants. The authors concluded that security self-reports oftentimes do not reflect users' actual behaviors, e.g., if the behavior involves awareness.

Fahl et al. investigated the ecological validity of a password study [17]. They compared the study-observed behavior of 645 participants with their real-life password choices. They conducted an online and laboratory research under priming and non-priming conditions. The authors found that around 20% of the participants behaved differently in the study compared to their real-life password behavior. They concluded that ecological validity is an important criterion, as it can reveal a high index of the irrelevance of laboratory studies to the real-life behavior.

Simultaneously, Mazurek et al. [27] studied the guessability of passwords used by members of a university in comparison to passwords that were previously collected in experiments or were leaked from low-value accounts. Having the same password policy, the authors found the real university passwords to be more similar to the passwords from research studies than when comparing university passwords to the leak passwords.

In order to increase the ecological validity of developer studies, Stransky et al. [40] designed an online platform which enables developers to participate in a study using their own equipment and allows them to participate from anywhere they would like to. The authors used the platform to conduct two studies and found that participants created code that was equally good as the code created in previous lab studies and noticed that participants were willing to spend more time when working online.

Finally, Naiakshina et al. [28–30] investigated whether deception task design–prompting participants to secure password storage–would change their security behavior in comparison to non-prompting. The study results showed that prompting has an effect on the security of participants' solutions. We replicated the freelancer study of Naiakshina et al. [28] in order to explore what impacts removing the deception emulating ecological validity might have.

2.2 Security Developer and Password Studies

Balebako et al. [7] conducted semi-structured interviews with 13 app developers. They investigated their view on security- and privacy-related topics. Many participants stated that both were part of their development process but that they were not their top priority. The authors found that security and privacy are positively correlated.

Acar et al. [3] invited 54 Android developers to take part in a lab study comparing different kinds of resources (free choice of resources, Stack Overflow only, official Android documentation only, books only). The participants only using Stack Overflow wrote significantly less secure code than the others, while the ones using only books wrote significantly less functional code than the others.

In a further study [2], Acar et al. compared the usability of five different APIs. They asked 256 GitHub users to solve various tasks concerning symmetric and asymmetric encryption. Twenty percent of the users who solved the task functionally believed to have solved it securely while it was not. The researchers recommended better documentation with secure, easy-to-use code examples.

With regard to password policies, Bonneau and Preibusch [12] analyzed 150 websites which offer free user accounts for various purposes. They found a great variety of security implementations. Websites with few security-critical features had the worst security practices. Other websites storing more sensitive data, such as financial data, implemented better security.

In 2007, Prechelt [32] arranged a contest where teams of web developers took part using different web development platforms (Java EE, Perl, PHP). They all had 30 hours to implement the same requirements for a web-based application. Their results were compared along various factors like usability, functionality and security. They found that PHP was in many aspects "at least as safe" as the other platforms and that it tended to have the smallest within-platform variations. Finifter and Wagner [18] conducted further analysis on the code of [32]. The authors investigated the relation between programming language and its number of vulnerabilities and the frameworks' support for security features and number of vulnerabilities. They did not find a relation between choice of programming language and application security, but they noticed that the developers almost never made use of the frameworks' built-in security support, e.g., for password storage. Like Finifter and Wagner, we also investigated whether freelancers would make use of the Spring frameworks' built-in libraries for secure password storage.

In a further study, Acar et al. [5] invited 307 GitHub users to work on security related tasks (e.g., password storage) in

Python and to take part in a survey afterwards. The authors found a positive correlation between performance regarding security and functionality and years of programming experience.

Another password-storage study was conducted by Wijayarathna and Arachchilage [47] with 10 developers. The authors explored usability issues of the Bouncycastle API to provide insights on how to develop, design and improve security/cryptographic APIs. They identified 63 issues in the SCrypt implementation of Bouncycastle.

2.3 Developer Studies with Freelancers

Ahmed and van den Hoven [6] discussed the freelancers' responsibility and ability to cause harm. They pointed out that most freelancers only do exactly what they are asked to do, which can cause security issues with employers who do not have a computer science background–an observation which was also found in Naiakshina et al.'s studies with freelancers [28] and students [29, 30] indicating that this issue cannot be reduced to freelancers only. Another problem Ahmed and van den Hoven found was that freelancers use malicious code from the Internet without testing it to ensure functionality and security. The researchers wanted to encourage freelancers to accept their responsibility. In our study, we acknowledge these aspects by providing participants password storage standard sources, if they submitted non-secure solutions.

In a further study, Bau et al. [9] compared the websites developed by start-up developers with websites they asked freelancers from Elance.com and Freelancer.com to develop. They wanted to investigate how the employment status, the developer's security knowledge and the programming language influence web application security. Nineteen start-ups and 8 freelance developers were invited and all of them were interviewed and took part in a security quiz. Their analysis showed that the code written by freelancers had more weaknesses than the code written by the start-ups. There was a huge discrepancy between the freelancers' security knowledge and their implementations. Furthermore, they found that code written in PHP had more injection vulnerabilities than code written in other programming languages.

While Bau et al. referred to freelancers as being rather unreliable, in another freelancer study, Yamashita and Moonen [50, 51] conducted a study with 85 freelancers on code smells and acknowledged the flexibility, the access to a wide population, and the low costs of Freelancer.com. Furthermore, in contrast to Bau et al., Naiakshina et al. [28] reported freelancers to be very dependable in their study. Most of the subjects delivered their solutions within the time frame they promised; they were also reliable in their communication and showed a high interest in the study results.

3 Methodology

With the aim of improving the ecological validity of their study, Naiakshina et al. [28] concealed the context of their scientific research and hired freelancers for a "real project." The authors invested additional study design work into the deception by creating a fake start-up with a web presence, creating a fake profile on Freelancer.com, designing a task description as authentic as possible and contacting and hiring individual developers based on their skills.

To test whether similar results are achieved without the use of deception, we conducted a replication-extension [10] of Naiakshina et al.'s [28] freelancer study, which adopted the study design from the original CS student studies [29, 30]. While we replicated the study of Naiakshina et al. with freelancers [28], we also extended it by the methodology (study announcement) and the framework variable, which we adopted from the previous student studies [29, 30] as well as an additional security request (see Table 1). The task was to complete a Java registration functionality that facilitated the storage of user properties (including user passwords) from a web form in a database. We used the task description given in [28] and hired participants from Freelancer.com. Similar to the previous studies, *prompting* was one variable in our study: half of the participants were tasked to securely store the passwords while the other half was not explicitly asked to do so.

When the project was published on Freelancer.com, the participants bid on it, and we contacted them via private messages. To communicate with the participants, we used the playbook from [28] and extended it when new cases appeared. The changes we made are given in the Appendix E. After completing the programming task, participants were asked to fill out an online survey to obtain their opinions on the used frameworks, their demographics, and their feedback about the task. In the following section, we provide a detailed description of all the design changes we made in contrast to the previous study [28].

3.1 Study Design Changes

Recruitment. One major change from the previous work was the public posting of the project on the freelancer platform as part of a scientific study. In the freelancer study of Naiakshina et al. [28], the researchers sent private messages with the project offer to freelancers who had mentioned Java knowledge in their profiles. Due to the limited filter features, the researchers needed to manually inspect freelancers' profiles to verify whether they actually had the required knowledge. Eighty of the 340 selected subjects did not have the required knowledge. Additionally, the remaining 260 freelancers were contacted by the researchers, but a total of 211 did not accept the offer for different reasons, such as their lack of experience or time. In our study, developers had to submit an application

for the project and thus, it was ensured that all the applicants were available for the study. Our public announcement for the study is available in the Appendix E.1.

Study Announcement. In the previous study, the researchers presented themselves as a start-up company and revealed their academic aims only at the end of the programming task. To make the project as realistic as possible, the original task presented in [29, 30], needed to be changed from a university setting to a company setting. The authors created a fake company profile on Freelancer.com and a web presence for that company and shared that with the subjects. While we used the same task description as used in the previous freelancer study, we omitted all the other steps and introduced ourselves as academic researchers conducting a scientific study.

Framework. While in the original freelancer study only JSF was used, we randomly assigned the participants to use either JSF or Spring as introduced in the previous student studies [29, 30]. The JSF participants had to implement secure password storage on their own. In contrast, Spring offers supporting libraries.

Security Requests. In their CS student studies, Naiakshina et al. [29, 30] accepted the participants' initial solutions and evaluated them based on a security score. To sum up, participants received a score of 0-7 points for security, based on their implementation choices for the password storage security, such as the hashing algorithm, salt generation, iterations etc. (see Section 3.4). In the freelancer study, the authors extended the security score and included a security request:

- *SecRequest-P(laintext)*: If the submissions included plain text password storage, subjects were asked to revise their submissions and to securely store user passwords.

We adopted this procedure and extended it by a further security request:

- *SecRequest-G(uideline)*: If the security score of participants' solutions was less than 6 points, we asked them to store the passwords by following industry's best practices.

As in the previous studies by Naiakshina et al., we accepted 6 points, instead of the full 7, for security, because Springs' default implementation of the bcrypt scores 6 points; thus, no one received the full 7 points by using memory-hard hashing functions. To investigate whether developers could obtain the full 7 points when given the appropriate source, we provided our participants with website links to the password security guidelines of the Open Web Application Security Project (OWASP) [1] and National Institute of Standards and Technology (NIST) [21] for the SecRequest-G. The exact wording and an illustrated procedure of the security requests can be found in the Appendix A.

Compensation. In the freelancer study, Naiakshina et al. [28] tested the impact of a payment variable on security by recruiting subjects either with a payment of €100 or €200. After revealing the study context, freelancers were invited to fill out a survey for additional €20. To allow researchers to conduct scientific studies in an economical manner, we began recruiting participants with a lower compensation amount of €120 for the project, including the programming task and the survey. Naiakshina et al. did not find a significant effect between the two payment levels on security. Therefore, after facing difficulties to recruit over 18 participants with €120, we started offering €220.[1]

Performance Based Payment. Based on the insights obtained from our pilot study (see Section 3.2), our payment method resembled the prior freelancer password storage study [28] but was split into compensation milestones. Since we included up to three possible iterations of the code review and possible security requests, we divided the payment process based on three milestones: €50 for the first code submission, €50 for the final code release after our review including possible security requests, and €20 the survey completion. Accordingly, participants received €100, €100, and €20 respectively for each milestone for a payment of €220.

3.2 Pilot Study

We recruited two freelancers via private messages to participate in our pilot study. We offered €220 to each. One participant reported that he wanted further instructions instead of just the mention of OWASP and NIST. Therefore, we included the links to the security guidelines in the final study. Since one participant appeared to be bothered by the requests, we divided the payment process based on milestones to clearly indicate that there were additional steps in our study. One participant stated that it took him six hours to finish the task, but he submitted the solution after six days. The other participant worked for three and a half hours on the task. Since both the participants reported to have finished the task in a relatively short period of time we started the recruitment process with €120.

3.3 Participants

Recruitment. We posted our project in 5 iterations over a time-span of almost two months. On each public project, freelancers could place bids with their payment offer, which ranged from €120/€220 to €250. In total we received 101 applications, of whom we invited 73 to the study. Since we limited our payment to €220, participants with higher bids were not invited to the study. In total, we excluded 28 applicants for requirement reasons, such as participants with a bid higher than we offered (12), already participated in our

[1] As in the previous freelancer study [28], payment did not show an effect on the decision to include security in the initial solution (FET: $p = 0.55$, OR = 1.54, CI = [0.39, 6.19]). We also regarded the prompting vs. the non-prompting group and did not find any significant effect in both cases.

General Information:			
Gender	Male: 32	Female: 11	NA: 0
Age	min: 18, max: 46	mean: 28.95, median: 29	SD: 6.21
Country of Residence	China: 11, Hong Kong: 1, The Netherlands: 1, Palestine: 1,	India: 10, Gaza: 1, New Zealand: 1, Romania: 1,	Pakistan: 6, Russia: 3, Lithuania: 1, Vietnam: 1, Germany: 1, Malaysia: 1
Profession and Programming Experience:			
Profession	Freelance Developer: 28 Employed full time: 1	Industrial Developer:12 Undergraduate full-time Student: 1	Finance Professional:1
General Development Experience [years]*	min: 1, max: 20	mean: 7.42, median: 7	SD: 4.47
Java Experience [years]*	min: 1, max: 16	mean: 6.19, median: 5	SD: 3.69

* = There were no significant demographic differences between the groups.

Table 2: Demographics of participants (n = 43)

study (9), and other reasons (7), such as missing Java skills in their profile or large time frames. Forty-three accepted our study invitation and participated in the study.

The first 3 iterations were posted with a payment of €120 and received a total of 60 bids from which we invited 46 freelancers to the study. Fourteen were removed for requirement reasons. Eight participants did not answer to the study invitation and another 8 were not interested anymore after reviewing the study material. Two wanted a higher payment for the project, one had no Java skills, two declined without seeing the study materials and one did not believe that we were conducting a study. Finally, 24 agreed to participate in our study. In the last 2 iterations, we offered a payment of €220. We received 41 bids and contacted 27 potential participants, of which 19 agreed to participate in the study. Fourteen freelancers were removed for requirement reasons. Six participants did not respond and one told us he is not interested anymore. Another participant wanted a higher payment for the project. In total 43 freelancers participated in our study and were randomly assigned to one of the 4 conditions: Spring-Prompting (FSP), Spring-Non-Prompting (FSN), JSF-Prompting (FJP), and JSF-Non-Prompting (FJN).

Demographics. Table 2 summarizes the demographics of our participants. While the demographics were comparable to Naiakshina et al.'s [28] freelancers in general, more female participants were involved in this study. 74% (32 of 43) were male, and 26% (11 of 43) were female. Most of the participants claimed to be from China (11), India (10), and Pakistan (6). Their ages ranged between 18 and 46 years (mean: 28.95, median: 29, SD: 6.21). The general programming experience of the participants ranged from 1 to 20 years (mean: 7.42, median: 7, SD: 4.47). For Java, the programming experience was reported to be between 1 and 16 years (mean: 6.19, median: 5, SD: 3.69). Almost all the participants reported to be experienced in developing web applications (41 of 43) and desktop applications (27 of 43). Thirty-eight participants reported to have a university degree. The freelancers had the option of indicating a minimum hourly wage in their Freelancer.com profile. The lowest rate among our participants was €5/hour, while the highest was €46/hour (mean: 21.71, median: 19, SD: 11.3, NA: 2).

3.4 Evaluation

Code Analysis. When releasing the milestones, we accepted only functional solutions. We adopted the extended version of the security scale [29] for the password storage security used by Naiakshina et al. [28–30] to score the code submissions. To sum up, we used a binary variable *secure* that indicated whether a participant included at least some kind of security. An ordinal variable *security score* was used to assess the security of the solution according to the password security scale [29] from 0-7; the *security score* considered the hash algorithm, iteration count for key stretching and salt generation. Submissions in which the user passwords were stored as plain text in the database received 0 points. Base64 and symmetric encryption are not suitable methods for secure password storage, and thus, any submissions that included these methods were rated being insecure (0 points). The security scale can be found in the Appendix B.

Two coders independently reviewed all the programming code submissions and evaluated them for security. Disagreements were resolved by consulting a security expert and discussing the algorithm specifications. We had 7 cases of disagreement, e.g., if one coder assessed the iterations incorrectly or misread the hash length. However, after a discussion all cases were resolved. With the rigid scoring system and the strict algorithm specifications, full agreement was achieved among the researchers.

Statistical Testing. We evaluated the same hypotheses as Naiakshina et al. [28, 30]. We were able to examine the effect of prompting vs. non-prompting (H-P1), framework (H-F1), years of Java experience (H-G1), and password storage experience (H-G2) on security. We also used the same statistical tests as in [28, 30]. All the tests on the same dependent variable were corrected using the Bonferroni-Holm correction. We denoted all corrected tests with "family = N," where N is the family size, and reported both the initial and corrected p-values (cor-p). An additional description of the hypotheses and a summary of our statistical analysis can be found in the Appendix C and D.

Qualitative Analysis. The open-ended questions from the follow-up survey were analyzed by two researchers using inductive coding [41]. The two researchers independently searched for codes categories and themes emerging in the raw

data. The codes were compared and the inter-coder agreement was calculated by using the Cohen's kappa coefficient (κ) [14]. The agreement was 0.83. A value above 0.75 is considered a good level of coding agreement [19]. We found a large number of similar codes as in the previous freelancer study [28]. In order to provide novel insights, we only report new codes and findings in this work.

4 Limitations

Sample. Similar to Naiakshina et al. [28], our sample consisted of developers from Freelancer.com. This sample is not be representative for all developers. Other freelancer hiring services could be used by other developers, and the results may vary. Further, since we publicly announced the project, participants needed to actively contact us, leading to a possible self-selection bias.

Recruitment Payment. We used two payment levels to recruit participants. This might have led to a self-selection bias. However, we tested the effect of payment on security in the initial solution and did not find any significant effect between both payment levels.

Deception. Similar to Naiakshina et al.'s previous password storage studies, this work examined the prompting vs. non-prompting effect in the task description. This resulted in concealing the security-focused research from half of our participants. However, due to our security requests, participants were able to improve their submissions. We received only positive feedback from our participants.

Generalizability. Finally, our findings are based on a single example study and thus further studies are needed to see if our results replicate in other types of studies.

5 Ethics

The institutional review board of our university approved our project. The participants were sent a link to a consent form in the first message of the conversation. We informed them of our data-storage policies and that they could withdraw their data at any time. To treat all the participants equally, we compensated the participants who had initially been offered a lower pay amount (€120) with additional €100 at the end of our study. Thus, all our participants received €220 for their efforts.

6 Results

In this section, we present an analysis of the present study. Additionally, we compare the results of the present study with results from the previous studies [28, 30]. To enable this comparison, we investigated the effect of the same factors on whether participants decided to store user passwords securely in the database considering the initial submissions. Further, we compared submissions from the previous freelancer sample [28] with our sample.

6.1 Security

As in the previous freelancer study [28], our participants used three techniques to store user passwords in the database: (1) hashing (+ salting); (2) symmetric encryption; and (3) Base64 encoding. We rated the solutions according to the security scale introduced in Section 3.4.

Table 3 shows the summary of the initial password storage solutions and the solutions handed in after SecRequest-P (in bold). To submit their initial solutions, participants took on average 4 days (min: 2h 20 min, max: 29 days, median: 2 days, SD: 5.4 days). In total we received 23 non-secure and 20 secure initial solutions from our 43 participants. Seventeen of the 23 participants with non-secure submissions received SecRequest-P as they stored the user passwords in plain text. Most of these requests were sent to non-prompted participants (15/17). Including security prompting in the initial task description meant that there was (almost) no need to remind participants not to save passwords in plain text. For SecRequest-P, they needed on average 1.8 days (min: 15 min, max: 19 days, median: 2h 30 min, SD: 4.4 days). After SecRequest-P all participants except one at least hashed the user passwords.

Overall, 25 participants received SecRequest-G because their first or second submission achieved less than 6 points on the security scale. For SecRequest-G, participants needed on average 2 days (min: 15 min, max: 19 days, median: 1 day, SD: 3.7 days). We provide a deeper analysis of submissions after SecRequest-G in Section 6.5.

Participants needed on average 5.8 days for their final submissions (min: 2h 20 min, max: 32 days, median: 3 days, SD: 7.9 days). In contrast to the previous work [28], we wanted to have a more accurate time specification. Thus, we asked participants in the survey how much time they actually needed to finish the task. On average, they stated that it took them 14 hours and 30 minutes to complete the task (min: 1 h, max: 72 h, median: 10 h, SD: 14 h 50 min).

6.2 Prompting effect (H-P1)

As done with all the previous studies on password storage of Naiakshina et al. [28–30], we examined the effect of prompting for security in the task. We also found a significant effect of prompting on the security of participants' submissions (FET: $p = 0.006$*, $cor - p = 0.01$*, OR = 6.51, CI = [1.51, 33.18], family = 2). The majority of non-prompted participants submitted a non-secure solution (16 of 21); only 5 participants considered security. Of the 22 prompted participants, 15 at least hashed the passwords.

Participant	Prompting	Framework	Payment	Working Time	Include SecRequest-P	Active Working	Security Requests	Function	Length in bits	Iteration	Salt	Secure	Score	Copied
FJN1	**0**	**JSF**	**120**	**3 Days**	**3 Days**	**7h**	**P + G**	**SHA-1**	**160**	**1**		**0/1**	**0/2**	
FJN2	**0**	**JSF**	**120**	**3 Days**	**2 Days**	**4h**	**P**	**bcrypt**	**184**	2^{10}	**SR**	**0/1**	**0/6**	
FJN3	**0**	**JSF**	**120**	**18 Days**	**2h 30min**	**11h**	**P + G**	**MD5**	**128**	**1**		**0/1**	**0/1**	✓
FJN4	0	JSF	120	3 Days		8h		PBKDF2 (SHA-512)	512	20 000	SR	1	6	
FJN5	**0**	**JSF**	**220**	**2 Days**	**30min**	**12h**	**P + G**	**MD5**	**128**	**1**		**0/1**	**0/1**	
FJN7	0	JSF	120	3 Days		25h	G	MD5	128	1		1	1	
FJN8	0	JSF	120	7 Days		4h	G	PBKDF2 (SHA-512)	512	65 536	St	1	5	
FJN9	**0**	**JSF**	**220**	**12 Days**	**1 Day**	**30h**	**P + G**	**SHA-1**	**160**	**1**		**0/1**	**0/2**	
FJN10	0	JSF	220	3 Days		8h	G	pgCrypto(xdes)	64	725	pgC	1	3	
FJN11	0	JSF	220	2h 20min		4h		bcrypt	184	2^{10}	SR	1	6	
FJN12	**0**	**JSF**	**220**	**2 Days**	**35min**	**8h**	**P + G**	**MD5**	**128**	**1**		**0/1**	**0/1**	
FJP1	1	JSF	220	1 Day		5h		bcrypt	184	2^{10}	SR	1	6	
FJP2	1	JSF	120	3 Days		48h	G	sym. Encryption				0	0	
FJP3	1	JSF	120	1 Day		10h	G	PBKDF2 (SHA-1)	128	65 536	SR	1	5	
FJP4	1	JSF	220	1 Day		18h		bcrypt	184	2^{14}	MR	1	6	✓
FJP5	1	JSF	120	6 Days		16h		PBKDF2 (SHA-1)	160	20 000	SR	1	6	✓
FJP6	1	JSF	120	2 Days		3h	G	MD5	128	1		1	1	✓
FJP7	**1**	**JSF**	**120**	**3 Days**	**15min**	**10h**	**P + G**	**MD5**	**128**	**1**		**0/1**	**0/1**	✓
FJP8	1	JSF	120	4h		1h	G	MD5	128	1		1	1	✓
FJP9	1	JSF	120	12 Days		60h	G	Base64				0	0	
FJP10	1	JSF	120	1 Day		12h	G	sym. Encryption				0	0	
FJP11	1	JSF	220	2 Days		10h	G	sym. Encryption				0	0	
FSN1	**0**	**Spring**	**120**	**3 Days**	**5h**	**15h**	**P + G**	**SHA-512**	**512**	**1**	**SR**	**0/1**	**0/5**	
FSN2	**0**	**Spring**	**220**	**8h 30min**	**20min**	**4h**	**P**	**bcrypt**	**184**	2^{10}	**SR**	**0/1**	**0/6**	
FSN3	**0**	**Spring**	**120**	**7 Days**	**25min**	**10h**	**P**	**MD5**	**128**	**1**		**0/1**	**0/1**	
FSN5	0	Spring	120	1 Day		6h	G	sym. Encryption				0	0	
FSN6	**0**	**Spring**	**120**	**13 Days**	**19 Days**	**30h**	**P**	**bcrypt**	**184**	2^{12}	**SR**	**0/1**	**0/6**	
FSN7	**0**	**Spring**	**120**	**1 Day**	**1 Day**	**10h**	**P + G**	**SHA-1**	**160**	**1**		**0/1**	**0/2**	
FSN9	**0**	**Spring**	**220**	**1 Day**	**2h**	**6h**	**P + G**	**SHA-256**	**256**	**1**		**0/1**	**0/2**	
FSN10	**0**	**Spring**	**120**	**1 Day**	**1h 10min**	**6h**	**P + G**	**sym. Encryption**				**0/0**	**0/0**	
FSN11	**0**	**Spring**	**220**	**9h**	**20min**	**2h**	**P**	**bcrypt**	**184**	2^{10}	**SR**	**0/1**	**0/6**	
FSN12	**0**	**Spring**	**220**	**1 Day**	**3 Days**	**9h**	**P**	**bcrypt**	**184**	2^{10}	**SR**	**0/1**	**0/6**	
FSP1	1	Spring	220	29 Days		72h		bcrypt	184	2^{10}	SR	1	6	
FSP2	1	Spring	120	4 Days		20h	G	SHA-256	256	1		1	2	
FSP3	1	Spring	120	4h 30min		2h		bcrypt	184	2^{10}	SR	1	6	
FSP4	1	Spring	120	5 Days		10h	G	MD5	128	1 000	SR	1	4	
FSP5	1	Spring	220	2 Days		16h		bcrypt	184	2^{10}	SR	1	6	
FSP6	1	Spring	120	1 Day		12h	G	Base64				0	0	
FSP7	1	Spring	220	7 Days		20h		bcrypt	184	2^{10}	SR	1	6	
FSP9	1	Spring	220	1 Day		11h		bcrypt	184	2^{10}	SR	1	6	
FSP10	**1**	**Spring**	**220**	**2 Days**	**1 Day**	**35h**	**P**	**bcrypt**	**184**	2^{10}	**SR**	**0/1**	**0/6**	
FSP12	1	Spring	220	2 Days		7h		bcrypt	184	2^{10}	SR	1	6	
FSP13	1	Spring	220	1 Day		8h		bcrypt	184	2^{10}	SR	1	6	

Table 3: Evaluation of participants' initial and SecRequest-P submissions
Bold: Participants who at first delivered a plain text solution and thus, received the first security request (SecRequest-P). **Working Time** participants took to submit their initial solution. **Include SecRequest-P:** Time participants needed to add security after SecRequest-P (1 Day = 24 hours). **Active Working:** Self-reported active working time reported by participants for their final submissions. **Salt:** SR = SecureRandom, St = Static, pgC = pgCrypto, MR = Math.Random. **Copied:** Security code was most likely copied and pasted from the Internet.

6.3 Java and Password Storage Experience (H-G1, H-G2)

Similar to the previous freelancer study, we did not find any effect of Java experience on the security score of the submissions (Kruskal-Wallis: $\chi^2(12) = 8.46$, $p = 0.75$, $cor - p = 0.75$). Further, we only examined submissions which did include some security, but did not find any effect of Java experience on the security score either (group: secure = 1; Kruskal-Wallis: $\chi^2(7) = 3.83$ $p = 0.80$).

In addition to that, we asked our participants whether they had stored user passwords in a database before. Only 3 of 43 participants said that they had never stored passwords before. As in the previous studies [28, 30], we did not find any significant effect on their decision to store the passwords securely in our study (FET: $p = 0.59$, $cor - p = 0.59$, OR = 0.42, CI = [0.01, 8.63]).

6.4 Framework (H-F1)

Similar to the previous student study [30], we did not find that the framework had a significant effect on the security score when participants stored the passwords securely (group: secure = 1; Wilcoxon Rank sum: W = 32.5, $p = 0.16$, family = 2, $cor - p = 0.32$). The mean score for the JSF group was 4.18 (group: secure = 1; min: 1, max: 6 , median: 5, SD: 2.23). The Spring group received a higher mean of 5.33 (group: secure = 1, min: 2, max: 6, median: 6, SD: 1.41). In both studies the spring group achieved slightly higher scores, but since neither effects were statistically significant, future studies will have to decide whether the small improvement the spring framework might bring is worth conducting a study with more power.

Further, we investigated the reported usability of both APIs. We calculated the API usability scores suggested by Acar et al. [2] for the Spring group (min: 62.5, max: 92.5, mean: 74.17, median: 72.5, SD: 9.36) and the JSF group (min: 50, max: 100, mean: 68.75, median: 70, SD: 13.11). The

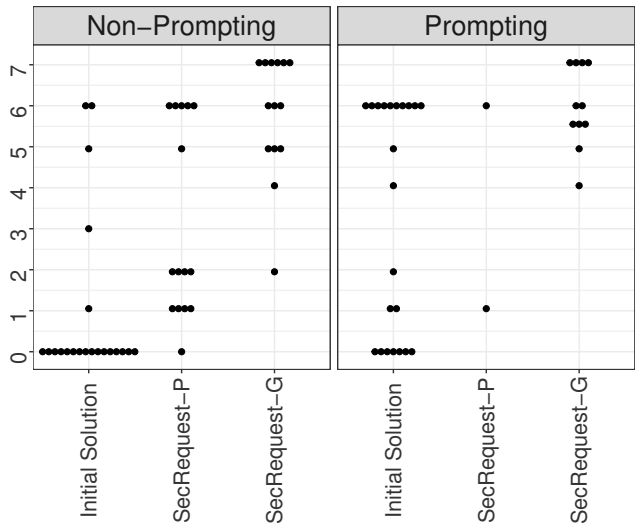


Figure 1: Distribution of scores before and after the security requests for the prompted and non-prompted group.
Each point stands for one participant.

score could range from 0-100 with 100 being the highest usability score that can be achieved. Thus, the Spring group achieved higher usability scores, however the difference was not statistically significant. (Wilcoxon Rank Sum: W = 160, $p = 0.08$).

Moreover, we tested for a correlation between API usability and the achieved security scores but did not find any significant correlation (Pearson: r = 0.08, $p = 0.57$).

In this study, we asked participants to evaluate the API usability after possible security requests, so we can be certain that all the participants used security mechanisms within the frameworks before giving us their assessment. In the student study [30] however, there were no follow up security requests and thus thus, there were participants who did not use any security mechanisms. Since their experience with a framework was reported based on functionality aspects only, we do not draw a direct comparison to our study.

6.5 Security Guidelines (NIST and OWASP)

Figure 1 shows the distribution of scores for the prompted and non-prompted groups for initial submissions and after participants received SecRequest-P and/or SecRequest-G. The figure visualizes how the distribution evolved after each request. The achieved scores rose after each request. Out of all 43 participants, 25 received SecRequest-G with web links to NIST and OWASP. The evaluation of these submissions is available in the Appendix (Table 6). The mean security score achieved was 5.86 (min: 2, max: 7, median: 6). Ten of the 25 participants used Argon2 to store the passwords and thus, achieved 7 points on the security score. Five of the 25 participants used bcrypt and achieved 6 points. The remaining 10 of 25 participants submitted solutions below 6 points, like PBKDF2 with insufficient parameters, SHA-1, or MD5.

In the follow-up survey—independently whether participants received SecRequest-G—we asked participants whether they know about and have experience with NIST or OWASP sources for user password storage. 58% (25 of 43) of all participants reported that they had heard of NIST or OWASP before. 47% (20 of 43) participants stated to have followed one (or both) of the guidelines in their submissions (including participants without SecRequest-G). Of the 25 participants who received SecRequest-G, 15 reported to have heard of NIST or OWASP. Consequently, 10 participants received instructions from us with NIST and OWASP sources, but indicated to not know the sources.

When comparing both guidelines, NIST offers a more theoretical and complex recommendation, while OWASP offers Argon2 example code in Java with clearer recommendations. We investigated whether our participants had copied and pasted code from the OWASP guideline and found that out of 10 participants who implemented Argon2 after our second security request, 8 copied and pasted the code from the OWASP guideline. We found the same comments in the programming code as on the website.

In the following we present an analysis of participants' experience with the guidelines based on their open-question answers in the survey and the chat communication.

Guideline Experience with SecRequest-G. Our participants mentioned that reading the guidelines has helped them to increase their knowledge. Participants found the guidelines useful, as they provided them with details, they were not aware of:

> "*I was unaware of Argon2 and the weakness in PBKDF2-with-HMAC and hadn't thought of a few things that were mentioned in the guidelines (max password length for DoS protection, Unicode normalization)*" (FJP3).

FSN5 reported to like OWASP because it instructed him to use existing hashing and salting algorithms instead of implementing them himself:

> "*One of the OSWAP design principles is to keep security simple. In the registration process I avoided implementing own salt and hashing algorithms [...]. This reduces chances of making security mistakes.*"

Guideline Experience without SecRequest-G. Ten of 43 participants (21%), who did not receive SecRequest-G, reported to have heard of NIST or OWASP before. Five of them (FJN4, FJP4, FSN2, FSN12, FSP5) stated that they did not follow the guidelines of the organizations in our task. As reasons FJP4 noted that he heard of the organizations, but never implemented their guidelines before and FSP5 stated:

		Company Freelancer [28]				Study Freelancer (only JSF)				Study Freelancer (JSF and Spring)			
		Non-secure	Secure	Score	Total	Non-secure	Secure	Score	Total	Non-secure	Secure	Score	Total
Initial submission	Prompting	8	13	$\mu = 2.19$ ($\sigma = 2.52$) min = 0, max = 6	21	5	6	$\mu = 2.27$ ($\sigma = 2.8$) min = 0, max = 6	11	7	15	$\mu = 3.32$ ($\sigma = 2.8$) min = 0, max = 6	22
	Non-Prompting	17	4	$\mu = 0.86$ ($\sigma = 1.96$) min = 0, max = 6	21	6	5	$\mu = 1.91$ ($\sigma = 2.59$) min = 0, max = 6	11	16	5	$\mu = 1$ ($\sigma = 2.07$) min = 0, max = 6	21
	Total	25	17	$\mu = 1.52$ ($\sigma = 2.33$)	42	11	11	$\mu = 2.09$ ($\sigma = 2.64$)	22	23	20	$\mu = 2.19$ ($\sigma = 2.68$)	43
After SecRequest-P	Prompting	2	1	$\mu = 2$ ($\sigma = 3.46$) min = 0, max = 6	3	0	1	$\mu = 1$ ($\sigma = 0$) min = 1, max = 1	1	0	2	$\mu = 3.5$ ($\sigma = 3.54$) min = 1, max = 6	2
	Non-Prompting	4	10	$\mu = 2$ ($\sigma = 2.29$) min = 0, max = 6	14	0	6	$\mu = 2.17$ ($\sigma = 1.94$) min = 1, max = 6	6	1	14	$\mu = 3.13$ ($\sigma = 2.36$) min = 0, max = 6	15
	Total	6	11	$\mu = 2$ ($\sigma = 2.33$)	17	0	7	$\mu = 2$ ($\sigma = 1.83$)	7	1	16	$\mu = 3.18$ ($\sigma = 2.31$)	17

Table 4: Comparison of (non-)secure solutions and the security score

"*I could develop it without these practices.*" Both used bcrypt in their initial solution and received 6 points in the security score. FSN11, FSP1, FSP3, FSP10, and FSP13 reported to have followed one of the guidelines. Four of them used bcrypt in their first submission and FSP10 used bcrypt in his second submission after SecRequest-P. FSP1 reported about NIST:

> "*NIST guidelines were easy to follow as they are in line with security best practices.*"

6.6 Sample Comparison

In this section we present a direct comparison between the previous freelancer study [28] and our replication study. In the following we denote this study as *Study Freelancer* and the original study as *Company Freelancer*, since a company deception was used as study design. Unlike in the Company Freelancer study, where only JSF was used as framework, this study also looked at Spring as a between groups condition. However, for the examination of the deception effect, we restrict our comparison to the JSF participants of our study, since only there a direct comparison can be made. Figure 2 displays a distribution of participants' initial submission security scores from the both studies Study Freelancer and Company Freelancer. For our study, Spring and JSF results are reported separately. Considering the prompted and non-prompted groups, the distributions of the obtained scores are fairly similar.

Table 4 summarizes participants' initial and SecRequest-P related security results of both the Study Freelancer and Company Freelancer. Since in the original study SecRequest-G was not introduced to participants, we do not consider it for the comparison of the both studies. Table 4 shows the count of participants from the two groups (prompted and non-prompted) and how many of the solutions were secure or non-secure. The proportions of participants in each group in both freelancer samples appear to be similar. Further, the mean score values from the company freelancers and study freelancers (only the JSF group) are similar as well.

We did not find any significant difference between the JSF security scores of the Study Scenario ($\mu = 2.09$, $\sigma = 2.64$) and Company Freelancer Scenario ($\mu = 1.52$, $\sigma = 2.33$) in the scores of the initial submissions (group: secure = 1 & group = JSF; Wilcoxon Rank sum: W = 86, $p = 0.73$, $r = 0.09$). However, a lack of a statistically significant finding does not mean there is none. With a large enough sample even small differences will result in a statistically significant finding. Power analysis is necessary to avoid Type II static errors (i.e., false negatives). In this study, power analysis would have established whether the lack of a statistically significant difference between the replicated and original results was meaningful, or if it was just an artifact of too few participants. Future studies will have to decide whether the differences shown above are worth re-examining with a larger sample size.

6.6.1 Implementation Time

We compared the implementation time to submit the initial solution of our study with the prior freelancer study [28] (min: 1 day, max: 8 days, mean: 3 days, median: 3 days, SD: 1.88). We did not find any significant difference between the implementation time of both JSF groups (Wilcoxon Rank sum: W = 457, $p = 0.83$). Further, we compared the time spent on the first security request. In [28], the participants needed on average 6.4 hours (sd 7.3 h, median 3.17 h). We did not find a significant effect in both freelancer studies either (W = 74, $p = 0.52$). The same caveats about the lack of power apply as above.

6.6.2 Study Announcement: Self-reflection

In the follow-up survey our participants were asked whether they would have stored the password securely if it had not been a study. Interestingly the answers split in half, 22 reported to would have stored the passwords securely, and 21 reported that they did not think that they would have stored them securely. Six of the 22 who reported that they would have performed differently if it was not a study, stored the

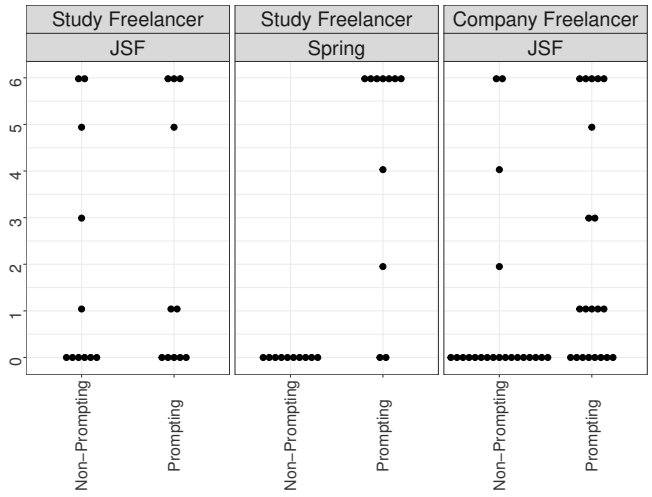


Figure 2: Initial submission score comparison of the previous freelancer study [28] (Company Freelancer) and our replication study (Study Freelancer).
Each point stands for one participant. In the Company Freelancer JSF group, Naiakshina et al. had almost twice as much participants as in the JSF group of this study, because we also investigated Spring as another value for the framework variable.

passwords in a sufficient way in the initial submissions. The results of the freelancer study of Naiakshina et al. [28] where the participants were not aware that they were participating in a study until they finished programming, shows that participants in developer studies can overestimate their own security awareness. We found that the reported details do not always fit the actual code submission.

7 Discussion

Methodological implications for developer studies: Our results suggest that freelancers are a useful sample for usable security developer studies. Response rates are higher than on GitHub. The studies can be conducted online and it is possible to reach developers from all over the world. Freelancers could provide more experience with real world projects and a wide range of age and experience. Since freelancers on the platform most likely do not know each other, the probability of participants communicating with each other about the study task and solutions is lower than for example, in a CS student sample. Additionally, freelancers are a convenient sample as they are looking for jobs and work with their own devices.

In comparison to the field study of Naiakshina et al. [28], where the study itself was concealed, our study was conducted by openly communicating the study context to the freelancers. Using the same sample size and same study protocol (minus the study deception), we got similar results as Naiakshina et al. We got similar effect sizes, directions, and statistically significant results for the same tests as they did and did not get any that they did not. In Table 5 a comparison of both studies and tests can be found. In both studies prompting lead to more security. The impact was significant in both samples, so we observed the same treatment effect (Company Freelancer: OR = 6.55 and Study Freelancer: OR = 6.51).

Neither study found a significant effect of previous password storage experience. However, in both cases only very few participants reported to have no password storage experience at all (Company Freelancer: n = 2 and Study Freelancer: n = 3). Therefore, the results should not be over-interpreted. Furthermore, Naiakshina et al. [28] did not find an effect of Java experience on security. In this study, we did not find an effect either. However, the direction of the correlation is the same in both studies. Future studies will have to decide whether to examine this effect with larger sample sizes.

In contrast to the previous freelancer study, we also tested the two frameworks Spring and JSF as done in the student study [30]. In [30], Naiakshina et al. did not find any significant difference between Spring and JSF. We did not find a significant difference between the two frameworks either. In both studies the mean security score was higher for the Spring framework but future studies with more participants would be needed to examine this effect further.

We conclude that for this study the removal of the deception element does not seem to have changed any outcomes and all relevant results gained from the original study with deception have also been gathered by this one without. Since deception should only be used when necessary, for this study we would not recommend to use it again in this specific context. While these results certainly do not generalize to all developer security studies, it is an important first indication that freelance developers recruited as part of a study behave similarly to when they are hired for a regular job. Therefore, our findings also offer an early indication that platforms such as Freelancer.com may be promising platforms for developer recruitment to supplement other channels such as CS students and GitHub developers.

Security guidelines: Acar et al. [3] showed in their programming experiment that using standard documentation without access to other sources such as the Internet lead to more secure code. However, the set-up of the experiment was rather artificial. In the real world developers use the Internet to find solutions while programming [30]. The standards are available but only a few developers use them or are even aware of them.

We found that concrete security policies with web links to the guidelines did increase the score of the password storage code. Ten participants were able to achieve full 7 points although no participants in the past studies and in the initial submissions achieved such a level for security. Even though some participants were able to use secure industry standards

IV	DV	Statistical Test	Company Freelancer [28]	Study Freelancer
Prompting	Secure	FET	p = **0.01*** OR = 6.55, CI = [1.44, 37.04]	p = **0.006*** OR = 6.51, CI = [1.51, 33.18]
Java Experience	Score	Kruskal-Wallis	p = 0.21, r = 0.12	p = 0.75, r = 0.04
Stored Passwords Before	Secure	FET	p = 0.17 OR = 0, CI = [0,3.87]	p = 0.59 OR = 0.41, CI = [0.04, 2.69]
Framework	Score	Wilcoxon Rank sum	- -	p = 0.16 group: secure = 1, r = 0.3

Table 5: Summary of all tests across the different samples
IV: Independent variable, **DV:** Dependent variable, **Company Freelancer:** Freelancers with study deception [28], **Study Freelancer:** Freelancers without study deception, **OR:** Odds ratio, **CI:** Confidence interval. The IV framework was not examined for freelancers in [28]. Significant tests are marked with *.

without being requested with the specific policies, a number of participants reported to know the guidelines. However, these participants were only able to score at most 6 points. This finding emphasizes the need for explicit encouragement of using security policies. The more support the policies offer the more secure the code can be.

Performance-based payment using milestones: Milestone payment is the recommended payment method on Freelancer.com. We split the payment into three milestones and chose the milestones for the initial submission and the (security) code review in a 1:1 ratio. It has to be investigated whether different ratios might result in different security results. Different performance rewards could help to increase security and the security awareness as well. We chose this ratio as we did not know how many security requests a participant would need as this depends on the password storage security in the initial and follow-up submission. In this study we chose the security to weight as much as the initial functional solution. We note that rewarding security performance as a study variable might lead to better initial solutions. In future research, this has to be evaluated with other scenarios and developer studies.

8 Conclusion

One major issue of usable security developer studies is their ecological validity. In the password-storage study of Naiakshina et al. [29, 30], CS students claimed that they would have behaved differently if they would have worked for a company. In a follow-up study, Naiakshina et al. [28] concealed the study context and hired freelancers for the same project. This form of deception requires additional study design work to maintain the deception. Frequent use of deception can cause problems as well. Both these issues make the use of deception for this kind of developer study something one would want to avoid if possible. Therefore, we replicated the deception study of Naiakshina et al. [28] without deception and compared the results. Overall the results were very similar leading us to propose the following recommendations:

- **When trying to get ecologically valid results for freelance developers, deception is not always necessary.** In our example running the study openly produced very similar outcomes compared to hiring freelancers for real. We got similar effect sizes and directions for the same tests as Naiakshina et al. in [28]. However, our study presents only one data point and further research is needed on the use of deception in other studies covering other security issues, tasks and scenarios, as well as with other types of developers.

- **Instruct developers to use security-guidelines.** We found that providing participants with specific guidelines for password storage can increase the security of their solutions drastically. Almost all our participants were able to implement secure password storage after being provided with specific security guidelines. If guidelines offer code examples, they are more likely to be implemented and included into the developers' code. Thus, we recommend designers of security guidelines to give specific code examples of secure code. We further recommend organizations to provide developers with specific security guidelines to receive software with state-of-the-art security standards. If guidelines might not be known to the employer, we recommend to include at least security prompting in the task to raise security awareness.

9 Acknowledgments

This work was partially funded by the ERC Grant 678341: Frontiers of Usable Security.

References

[1] Open web application security project (owasp). https://github.com/OWASP/CheatSheetSeries/blob/master/cheatsheets/Password_Storage_Cheat_Sheet.md. Last Accessed: November 14, 2019.

[2] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L Mazurek, and Christian Stransky. Comparing the usability of cryptographic apis. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 154–171, San Jose, CA, USA, 2017. IEEE, IEEE.

[3] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. You get where you're looking for: The impact of information sources on code security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 289–305, Piscataway, NJ, USA, May 2016. IEEE Press.

[4] Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. You are not your developer, either: A research agenda for usable security and privacy research beyond end users. In *Cybersecurity Development (SecDev), IEEE*, pages 3–8, Piscataway, NJ, USA, 2016. IEEE, IEEE Press.

[5] Yasemin Acar, Christian Stransky, Dominik Wermke, Michelle L. Mazurek, and Sascha Fahl. Security developer studies with github users: Exploring a convenience sample. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 81–95, Santa Clara, CA, 2017. USENIX Association.

[6] Malik Aleem Ahmed and Jeroen van den Hoven. Agents of responsibility—freelance web developers in web applications development. *Information Systems Frontiers*, 12(4):415–424, Sep 2010.

[7] Rebecca Balebako, Abigail Marsh, Jialiu Lin, Jason I Hong, and Lorrie Faith Cranor. The privacy and security behaviors of smartphone app developers. 2014.

[8] Titus Barik, Justin Smith, Kevin Lubick, Elisabeth Holmes, Jing Feng, Emerson Murphy-Hill, and Chris Parnin. Do developers read compiler error messages? In *Proceedings of the 39th International Conference on Software Engineering*, ICSE '17, pages 575–585, Piscataway, NJ, USA, 2017. IEEE Press.

[9] Jason Bau, Frank Wang, Elie Bursztein, Patrick Mutchler, and John C Mitchell. Vulnerability factors in new web applications: Audit tools, developer selection & languages. *Stanford, Tech. Rep*, 2012.

[10] Douglas G Bonett. Replication-extension studies. *Current Directions in Psychological Science*, 21(6):409–412, 2012.

[11] J. Bonneau. The science of guessing: Analyzing an anonymized corpus of 70 million passwords. In *2012 IEEE Symposium on Security and Privacy*, pages 538–552, San Francisco, CA, USA, May 2012. IEEE.

[12] Joseph Bonneau and Sören Preibusch. The password thicket: Technical and market failures in human authentication on the web. In *WEIS*, 2010.

[13] Deanna D Caputo, Shari Lawrence Pfleeger, M Angela Sasse, Paul Ammann, Jeff Offutt, and Lin Deng. Barriers to usable security? Three organizational case studies. *IEEE Security & Privacy*, 14(5):22–32, 2016.

[14] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[15] Cleidson R. B. de Souza, David Redmiles, Li-Te Cheng, David Millen, and John Patterson. Sometimes you need to see through walls: A field study of application programming interfaces. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, CSCW '04, pages 63–71, New York, NY, USA, 2004. ACM.

[16] Serge Egelman, Andreas Sotirakopoulos, Ildar Muslukhov, Konstantin Beznosov, and Cormac Herley. Does my password go up to eleven?: The impact of password meters on password selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2379–2388, New York, NY, USA, 2013. ACM, ACM.

[17] Sascha Fahl, Marian Harbach, Yasemin Acar, and Matthew Smith. On the ecological validity of a password study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, page 13. ACM, 2013.

[18] Matthew Finifter and David Wagner. Exploring the relationship between web application development tools and security. In *Proceedings of the 2Nd USENIX Conference on Web Application Development*, WebApps'11, pages 9–9, Berkeley, CA, USA, 2011. USENIX Association.

[19] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[20] Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke, Christian Stransky, Sebastian Möller, Yasemin Acar, and Sascha Fahl. Developers deserve security warnings, too: On the effect of integrated security advice on cryptographic (api) misuse. In *Fourteenth Symposium*

on Usable Privacy and Security (SOUPS 2018), pages 265–281, Baltimore, MD, 2018. USENIX Association.

[21] Paul A Grassi, James L Fenton, EM Newton, RA Perlner, AR Regenscheid, WE Burr, JP Richer, NB Lefkovitz, JM Danker, Yee-Yin Choong, et al. Nist special publication 800-63b: Digital identity guidelines. *Enrollment and Identity Proofing Requirements*. https://pages.nist.gov/800-63-3/sp800-63b.html, 2017.

[22] Ameya Hanamsagar, Simon S Woo, Chris Kanich, and Jelena Mirkovic. Leveraging semantic transformation to investigate password habits and their causes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 570, New York, NY, USA, 2018. ACM, ACM.

[23] Saranga Komanduri, Richard Shay, Patrick Gage Kelley, Michelle L Mazurek, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Serge Egelman. Of passwords and people: Measuring the effect of password-composition policies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2595–2604, New York, NY, USA, 2011. ACM, ACM.

[24] Katharina Krombholz, Wilfried Mayer, Martin Schmiedecker, and Edgar Weippl. "i have no idea what i'm doing" - on the usability of deploying HTTPS. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 1339–1356, Vancouver, BC, 2017. USENIX Association.

[25] Thomas D. LaToza, Gina Venolia, and Robert DeLine. Maintaining mental models: A study of developer work habits. In *Proceedings of the 28th International Conference on Software Engineering*, ICSE '06, pages 492–501, New York, NY, USA, 2006. ACM.

[26] Peter Mayer, Jan Kirchner, and Melanie Volkamer. A second look at password composition policies in the wild: Comparing samples from 2010 and 2016. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 13–28, Santa Clara, CA, 2017. USENIX Association.

[27] Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur. Measuring password guessability for an entire university. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS '13, pages 173–186, New York, NY, USA, 2013. ACM.

[28] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel von Zezschwitz, and Matthew Smith. "if you want, i can store the encrypted password": A password-storage field study with freelance developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 140:1–140:12, New York, NY, USA, 2019. ACM.

[29] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why do developers get password storage wrong?: A qualitative usability study. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 311–328, New York, NY, USA, 2017. ACM.

[30] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, and Matthew Smith. Deception task design in developer password studies: Exploring a student sample. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 297–313, Baltimore, MD, USA, 2018. USENIX Association.

[31] Duc Cuong Nguyen, Dominik Wermke, Yasemin Acar, Michael Backes, Charles Weir, and Sascha Fahl. A stitch in time: Supporting android developers in writingsecure code. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 1065–1077, New York, NY, USA, 2017. ACM.

[32] Lutz Prechelt. Plat_forms: A web development platform comparison by an exploratory experiment searching for emergent platform properties. *IEEE Transactions on Software Engineering*, 37(1):95–108, Jan 2011.

[33] Elissa M Redmiles, Ziyun Zhu, Sean Kross, Dhruv Kuchhal, Tudor Dumitras, and Michelle L Mazurek. Asking for a friend: Evaluating response biases in security user studies. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 1238–1255. ACM, 2018.

[34] Sean M Segreti, William Melicher, Saranga Komanduri, Darya Melicher, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L Mazurek. Diversify to survive: Making passwords stronger with adaptive policies. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 1–12, Santa Clara, CA, USA, 2017. USENIX Association.

[35] Awanthika Senarath and Nalin Asanka Gamagedara Arachchilage. Understanding software developers' approach towards implementing data minimization. *arXiv preprint arXiv:1808.01479*, 2018.

[36] Richard Shay, Saranga Komanduri, Adam L. Durity, Phillip (Seyoung) Huh, Michelle L. Mazurek, Sean M.

Segreti, Blase Ur, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Can long passwords be secure and usable? In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2927–2936, New York, NY, USA, 2014. ACM.

[37] Richard Shay, Saranga Komanduri, Adam L. Durity, Phillip (Seyoung) Huh, Michelle L. Mazurek, Sean M. Segreti, Blase Ur, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Designing password policies for strength and usability. *ACM Transactions on Information and System Security (TISSEC)*, 18(4):13:1–13:34, May 2016.

[38] Dag IK Sjoberg, Bente Anda, Erik Arisholm, Tore Dyba, Magne Jorgensen, Amela Karahasanovic, Espen Frimann Koren, and Marek Vokác. Conducting realistic experiments in software engineering. In *Proceedings international symposium on empirical software engineering*, pages 17–26, Piscataway, NJ, USA, 2002. IEEE, IEEE Press.

[39] Elizabeth Stobert and Robert Biddle. The password life cycle: User behaviour in managing passwords. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 243–255, Menlo Park, CA, USA, 2014. USENIX Association.

[40] Christian Stransky, Yasemin Acar, Duc Cuong Nguyen, Dominik Wermke, Doowon Kim, Elissa M Redmiles, Michael Backes, Simson Garfinkel, Michelle L Mazurek, and Sascha Fahl. Lessons learned from using an online platform to conduct large-scale, online controlled security experiments with software developers. In *10th USENIX Workshop on Cyber Security Experimentation and Test (CSET'17)*, 2017.

[41] David R Thomas. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, 27(2):237–246, 2006.

[42] Blase Ur, Jonathan Bees, Sean M. Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Do users' perceptions of password security match reality? In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 3748–3760, New York, NY, USA, 2016. ACM.

[43] Blase Ur, Patrick Gage Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L. Mazurek, Timothy Passaro, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Julio López. Helping users create better passwords. *USENIX*, 37(6):51–57, 2012.

[44] Blase Ur, Fumiko Noma, Jonathan Bees, Sean M. Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. I added '!' at the end to make it secure: Observing password creation in the lab. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 123–140, Ottawa, 2015. USENIX Association.

[45] Rick Wash, Emilee Rader, Ruthie Berman, and Zac Wellmer. Understanding password choices: How frequently entered passwords are re-used across websites. In *Twelfth Symposium on Usable Privacy and Security (SOUPS)*, pages 175–188, Denver, CO, 2016. USENIX Association.

[46] Rick Wash, Emilee Rader, and Chris Fennell. Can people self-report security accurately?: Agreement between self-report and behavioral measures. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2228–2232, New York, NY, USA, 2017. ACM.

[47] Chamila Wijayarathna and Nalin A. G. Arachchilage. Why johnny can't store passwords securely?: A usability evaluation of bouncycastle password hashing. In *Proceedings of the 22Nd International Conference on Evaluation and Assessment in Software Engineering 2018*, EASE'18, pages 205–210, New York, NY, USA, 2018. ACM.

[48] Chamila Wijayarathna and Nalin AG Arachchilage. Am i responsible for end-user's security? Baltimore, MD. USENIX Association.

[49] Khaled Yakdan, Sergej Dechand, Elmar Gerhards-Padilla, and Matthew Smith. Helping johnny to analyze malware: A usability-optimized decompiler and malware analysis user study. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 158–177, San Jose, CA, USA, 2016. IEEE, IEEE.

[50] Aiko Yamashita and Leon Moonen. Do developers care about code smells? an exploratory survey. In *Reverse Engineering (WCRE), 2013 20th Working Conference on*, pages 242–251. IEEE, IEEE, 2013.

[51] Aiko Yamashita and Leon Moonen. Surveying developer knowledge and interest in code smells through online freelance marketplaces. In *User Evaluations for Software Engineering Researchers (USER), 2013 2nd International Workshop on*, pages 5–8, San Francisco, CA, USA, 2013. IEEE, IEEE.

APPENDIX

A Security Requests

Figure 3 visualizes the security request procedure. When we were sent a solution where the password was stored in plain text or where the participant received less than 6 points in the *security score*, we sent the following messages:

- **SecRequest-P:** Participant handed in plain text code:
  R: *I saw that the password is stored in clear text. Could you also store it securely?*
- **SecRequest-G:** Security score < 6:
  R: *Thank you for submitting your solution. Now I have one further request. I noticed, that you did not follow industry best practices, e.g., NIST (National Institute of Standards and Technology) or OWASP (Open Web Application Security Project), to securely store the end-user password. Could you please revise your submission and ensure that you follow industry best practices? You can find some information on OWASP on this website:* `https://github.com/OWASP/CheatSheetSeries/blob/master/cheatsheets/Password_Storage_Cheat_Sheet.md` *and information on NIST on this website:* `https://pages.nist.gov/800-63-3/sp800-63b.html` *in section 5.1.1.2.*

B Security Scale

We based the evaluation of participants' submissions on the security score of Naiakshina et al. [29]:

1. The end-user password is salted (+1) and hashed (+1).
2. The derived length of the hash is at least 160 bits long (+1).
3. The iteration count for key stretching is at least 1 000 (+0.5) or 10 000 (+1) for `PBKDF2` and at least $2^{10} = 1\,024$ for `bcrypt` (+1).
4. A memory-hard hashing function is used (+1).
5. The salt value is generated randomly (+1).
6. The salt is at least 32 bits in length (+1).

C Hypotheses

Due to the adjusted study design, we were able to test only four of the seven main hypotheses from [30]. While it was possible to track security attempts in a lab setting, in an online study this information was not accessible. We did, though, consider the subset *secure = 1 (achieving security)* for our analysis. Hypotheses from [30]:

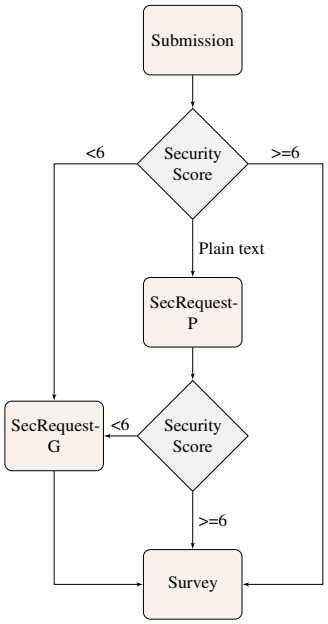


Figure 3: Security request procedure
Participants received SecRequest-P, if they submitted a plain text solution and SecRequest-G if the solution scored less than 6 points. Participants with plain text solutions thus, could receive both requests.

- H-P1 - Priming has an effect on the likelihood of participants attempting security.
- H-F1 - Framework has an effect on the security score of participants attempting security.
- H-G1 - Years of Java experience have an effect on the security scores.
- H-G2 - If participants state that they have previously stored passwords, it affects the likelihood that they store them securely.

D Summary of Statistical Analysis

Table 7 summarizes all hypotheses from our analysis of the study.

E Playbook

We used the same playbook Naiakshina et al. used in [28]. We extended and adapted it by several relevant aspects. **P** indicates the participant and **R** indicates the researcher. Because of space limitation, we mention only playbook extensions here.

Participant	Prompting	Framework	Payment	Include SecRequest-G	Function	Length in bits	Iteration	Salt	Secure	Score	Copied	NIST	OWASP
FJN1	0	JSF	120	1 Day	SHA-1	160	1		1	2			✓
FJN3	0	JSF	120	1 Day	bcrypt	184	2^{10}	SR	1	6			
FJN5	0	JSF	220	50 min	PBKDF2 (SHA-1)	256	10 000	St	1	5	✓	✓	
FJN7	0	JSF	120	1 Day	MD5	128	1	SR	1	4			
FJN8	0	JSF	120	2 Days	PBKDF2 (SHA-512)	512	65 536	St	1	5			✓
FJN9	0	JSF	220	19 Days	Argon2i	256	2	SR	1	7	✓	✓	
FJN10	0	JSF	220	1h 30min	PBKDF2 (SHA-1)	128	65 536	SR	1	5			✓
FJN12	0	JSF	220	15min	Argon2i	256	10	SR	1	7			✓
FJP2	1	JSF	120	2 Days	PBKDF2 (SHA-1)	512	1 000	SR	1	5.5			
FJP3	1	JSF	120	5h 30min	Argon2i	256	20	SR	1	7			✓
FJP6	1	JSF	120	2 Days	Argon2i	256	40	SR	1	7	✓		
FJP7	1	JSF	120	35min	PBKDF2 (SHA-1)	128	4	SR	1	4	✓	✓	✓
FJP8	1	JSF	120	3h 30min	PBKDF2 (SHA-1)	512	1 000	SR	1	5.5	✓	✓	
FJP9	1	JSF	120	2 Days	PBKDF2 (SHA-1)	512	1 000	SR	1	5.5	✓	✓	
FJP10	1	JSF	120	1 Day	Argon2i	256	40	SR	1	7	✓		
FJP11	1	JSF	220	1 Day	bcrypt	184	2^{12}	SR	1	6			
FSN1	0	Spring	120	2 Days	Argon2i	256	20	SR	1	7	✓	✓	✓
FSN3	0	Spring	120	6 Days	Argon2i	256	4	SR	1	7	✓		✓
FSN5	0	Spring	120	4 Days	bcrypt	184	2^{10}	SR	1	6			✓
FSN7	0	Spring	120	1 Day	Argon2i	256	40	SR	1	7	✓		
FSN9	0	Spring	220	1 Day	bcrypt	184	2^{10}	SR	1	6		✓	✓
FSN10	0	Spring	120	1h	Argon2i	256	40	SR	1	7	✓	✓	
FSP2	1	Spring	120	10h	PBKDF2 (SHA-1)	128	65 536	SR	1	5			
FSP4	1	Spring	120	3 Days	bcrypt	184	2^{10}	SR	1	6			
FSP6	1	Spring	120	14h	Argon2d	128	3	SR	1	7	✓		

Table 6: Evaluation of participants' submissions after SecRequest-G

Include SecRequest-G: Time participants needed to add security after SecRequest-G. **Salt:** SR = SecureRandom, R = Random, St = Static. **Copied:** Security code was probably copied and pasted from the Internet. **NIST/OWASP:** Participant stated that he/she followed the security guidelines of NIST/OWASP programming the task.

H	Sub-sample	IV	DV	Test	O.R.	CI	p-value	$cor-p$-value
H-P1#	-	Prompting	Secure	FET	6.51	[1.51, 33.18]	0.006*	0.01*
H-G1#	-	Java experience	Score	Kruskal-Wallis	-	-	0.75	0.75
H-G2#	-	Stored passwords before	Secure	FET	0.42	[0.01, 8.63]	0.59	0.59
H-F1#	secure = 1	Framework	Score	Wilcoxon rank sum	-	-	0.16	0.32
E-A1	secure = 1	Java experience	Score	Kruskal-Wallis	-	-	0.80	-
E-A2	-	Framework	API usability	Wilcoxon rank sum	-	-	0.08	-
E-A3	-	Score	API usability	Pearson Cor.	-	[-0.21, 0.38]	0.57	-
S-C1	secure = 1& group = JSF	Study sample	Score	Wilcoxon rank sum	-	-	0.73	-
S-C2	group = JSF	Study sample	Implementation time initial submission	Wilcoxon rank sum	-	-	0.83	-
S-C3	group = JSF	Study sample	Implementation time time for SecRequest-P	Wilcoxon rank sum	-	-	0.52	-

Table 7: Summary of statistical analysis

IV: Independent variable, **DV**: Dependent variable, **O.R.**.: Odds ratio, **CI**: Confidence interval, **E-A**: Exploratory analysis, **S-C**: Sample Comparison. All tests were conducted on security values of the *initial solutions* before participants received any security requests. H-P1 and H-G2 as well as H-G1 and H-F1 are corrected with Bonferroni-Holm correction (cor-p value). # = Hypothesis of Naiakshina et al. [28, 30], * = Significant Tests.

E.1 Study Announcement and Study Offer

We are researchers from the University of Bonn working in the field of software usability. We are always looking for software developers and system administrators who are interested in taking part in one of our studies (programming and surveys). All data will be processed pseudonymously and stored anonymized after the study; there will be no identifying information published in any form. If you are interested in participating in studies - Please contact us!

After the participants placed a bid on the project, we contacted them via the private chat at Freelancer.com with the following message:

R: *Hello XYZ, we are happy that you want to take part in our study. The payment will be divided into 3 milestones: 50 euros (100 euros) for your initial code release, additional 50 euros (100 euros) for the final code release after our review and further additional 20 euros for completing our survey about your programming experience and your experiences with this task. If this is fine for you and you want to take part in the study, we need you to sign a consent form. Please go to the following website to do so: LINK Your study-ID is: XXX Thanks in advance. Kind regards, ...*

When the freelancer signed the consent form, we sent the second message and a ZIP file with task and code:

R: *Hello XYZ, you agreed to take part in our study. Thank you for signing the consent form! Now I will send you the code as a ZIP file. You will find the task in there too. Please have a look at it and tell me if you want to do it. Then I will award you with the project and create the milestones. Kind regards, ...*

After that message, we got mostly three kinds of reactions:

- The freelancer agreed to take part, we awarded him/her

and wished him/her *Happy coding!*

- The freelancer did not react anymore:
  Hello XYZ, what do you think? Are you still interested to take part in the study?
- **P**: *May I check the code and get back to you tomorrow/later/. . . ?*
  R: *Yes, sure! Take your time.*

E.2 Deadline

Some of the participants asked us for a final deadline. Since we did not want to rush them, we did not set one, but asked them to tell us how much time they needed to finish it. When a freelancer did not ask for a deadline, we decided to contact him/her after 10 days to ask for an update.
We had two cases where the freelancers did not answer several questions for updates. In that case we set a deadline and ended the study, when they exceeded it.

- **P**: *What is the final delivery?*
  R: *What do you think how much time you need to solve the task?*
  P: DATE
  R: *Ok, that's fine. Thank you.*
- Deadline exceeded:
  R: *Hey XYZ, could you give us a status update? Kind regards, . . .*
- 10 days after task was sent and in case no deadline was set:
  R: *Hey XYZ, could you give us a status update? What do you think how much time you need to solve the task?*
- If no reaction after 3 weeks:
  R: *Hello, please send your solution till* date in one week. *If you decided to no longer participate in the study, I would be very glad if you could let me know. Thank you and kind regards, . . .*

E.3 Password-related Questions

The following questions concerned password storage:

- Before submitting initial solution:
  P: *Should I implement security/secure password storage?*
  R: *Whatever you would recommend!*
- **P**: *Is *** fine?*
  R: *Whatever you would recommend/use!*
- **P**: *I cannot find password encryption in the requirements, can you tell me where it is written? I might have missed it.*
  R: *It is not in there, that is true. But could you add it?*
- After SecRequest-G:
  P: *Should I implement all the rules mentioned in NIST document? There are so many rules in NIST.*
  R: *You don't have to implement all rules, but please concentrate on secure password storage in a database on the back-end site. For us it's most important that the password is saved securely.*

E.4 General Questions

Also in the general communication we often received similar questions.

- **P**: *Can I build it from scratch?*
  R: *You can solve the task as you prefer.*
- **P**: *If I have some questions for your project, can I ask you?*
  R: *Sure!*
- **P**: *Do you have a server where we can upload this code for you to test?*
  R: *None that I have access to. Would it be possible for you to send me a video or screenshots so I can see that it is working on your computer?*
- Participant did not work on our database:
  R: *Could you also make it work on our database?*
- **P**: *Once the user registers, do we need to send verification email also and once he clicks on that, we will make user status as active?*
  R: *No, we only need the data to be stored in our data base for now!*
- **P**: *I need to see the ER diagram.*
  R: *We do not have that yet. Is it necessary?*
  P: . . .
  R: *Could you please create a single table for now and I will talk to my mentor about the rest?*
- **P**: *Do you require the login functionality as well? Should I implement is as a further task? What else except registration will be needed?*
  R: *No, thank you. Please only program the registration functionalities.*
- **P**: *The password is in the database, so users won't be able to access it.*
  R: *And what if someone gets access to the database?*
- **P**: *Could you tell me what (. . .) is for? / Could you help me with (. . .)?*
  R: *Since we are conducting a study and all participants should have the same requirements, I cannot help you with specific questions about the code. I'm sorry!*

- Participant is confused (after SecRequest-G):
  R: *You will not receive any further request. You can choose, which industry standard you would like to follow; OWASP or NIST. So that you do not have to read the whole NIST guideline, you can read all necessary information in section 5.1.1.2. This section approximately complies with the length of the provided OWASP source. It is up to you to choose one standard. Afterwards you only have to fill out a subsequent survey, which concludes the study.*

E.5 Receiving the Solution and Survey Request

After receiving the final solution, we wrote:
Thank you, for sending your result! I will look into it.
We checked the remote database for code examples. If the freelancer had not worked on it, we wrote: *Could you also make it work on our database?*
If the freelancer could not make it work on our database, we asked for pictures and a video that showed that the code was working. We also checked it for functionality and if everything worked, we asked the freelancers to take part in our survey.
Message: *Hello XYZ, thank you for sending us your results. Like announced in the study description we would like to invite you to a concluding survey. You can find it here: LINK Kind regards, . . .*

E.6 Exit Communication

After the freelancers finished the survey, we released the last milestone and sent them the following message: *Thank you for your participation! We are happy about your feedback, but we would like to kindly ask you to not mention our study content in order to ensure the validity of our study. Thank you again!* And they all replied that they would not mention it.
Many of the freelancers asked for a good or a five-star review, which we gave them: *Yes, we did. It was nice working with you.*
Also many of them asked, if we had further projects in which they could take part.
R: *At the moment we unfortunately have only one project.*

E.7 Review

We gave all participants the same review:
Very good communication, delivered on time. It was nice working with him/her!

Innovation Inaction or In Action? The Role of User Experience in the Security and Privacy Design of Smart Home Cameras

George Chalhoub[1], Ivan Flechais[1], Norbert Nthala[2], Ruba Abu-Salma[3]
[1]*University of Oxford*
[2]*Michigan State University (MSU)*
[3]*University College London (UCL) & Inria*
[1]*{george.chalhoub, ivan.flechais}@cs.ox.ac.uk,* [2]*nthalano@msu.edu,* [3]*r.abu-salma@cs.ucl.ac.uk*

Abstract

Smart homes are under attack. Threats can harm both the security of these homes and the privacy of their inhabitants. As a result, in addition to delivering pleasing and aesthetic devices, smart home product designers need to factor security and privacy into the design of their devices. Further, the need for user-centered security and privacy design is particularly important for such an environment, given that inhabitants are demographically-diverse (e.g., age, gender, educational level) and have different skills and (dis)abilities.

Prior work has explored different usable security and privacy solutions for smart homes; however, the applicability of *user experience (UX)* principles to security and privacy design is under-explored. In this paper, we present a qualitative study to explore the development of smart home cameras manufactured by three companies. We conduct semi-structured interviews with 20 designers and their collaborators, and analyze these interviews using Grounded Theory. We find that UX was seen as helpful by our participants in fostering innovation in the design of privacy solutions. However, UX was not used or considered in the design of security solutions due to an explicit need for established, tried-and-tested solutions (i.e., previous traditional security solutions that were seen as effective and reliable to fix certain design problems). Drawing from the findings of our study, we propose a model of UX factors influencing security and privacy design of smart home cameras. We also extract a set of recommendations to improve the security and privacy design of smart cameras. We finally outline several areas for future investigation.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

1 Introduction

Homes are increasingly becoming instrumented, connected, and smart. Devices (including light bulbs, doorbells, door locks, thermostats, and coffee makers) are designed to be Internet-connected and offer greater convenience, functionality, and energy efficiency. However, the rise in the adoption and use of smart home devices is accompanied by new security and privacy threats [1]. Most smart home devices have always-on sensors that collect different types of data and then transmit the data over the Internet to various destinations [2, 3]. The data can be used to spy on or create fine-grained inferences of device users and other home inhabitants. Smart devices also increase the technical complexity of the security infrastructure in smart homes; non-expert home users are expected to protect themselves and their families from various attacks. Increasing numbers of attacks, which have seen attackers taking control of smart homes (e.g., [4,5]), emphasize the need for protecting smart homes. When such attacks happen, smart home device manufacturers often blame users for behaving insecurely (e.g., choosing weak passwords to secure smart home mobile applications [5]), while users ascribe the shortfalls to manufacturers.

Smart home devices do not only affect the privacy and lifestyle of users of these devices, but also those of every inhabitant of the home. In one example, a husband decided to unplug a smart home camera that his wife placed in their house (to check in on her family while she was away) because he felt that the camera was staring at him while he was making coffee [6]. This example illustrates, as Hassenzahl and Tractinsky [7] state, that product attributes must be linked to the needs and values of users (and bystanders in this instance). This entails considering the affective consequences of technology on people, as well as the situatedness and temporality of the product (e.g., the actual or anticipated experience of a user—and a bystander—with the product). Hassenzahl and Tractinsky argue that experience is a combination of elements including the product and internal states of the user (e.g.,

mood, expectations, active goals), which extends over time. The elements interact and modify one another and, hence, understanding the *user experience (UX)* of security and privacy design is important for the successful adoption and use of smart home devices. The importance of UX in the design of smart home devices has long been recognized and advocated for [8–10]. UX encompasses more than usability: it covers emotions, psychological responses, beliefs, perceptions, behaviors, and accomplishments [11].

Prior work has investigated how to achieve security and privacy in smart homes, focusing mainly on technical aspects (e.g., [12, 13]). Other work has also explored the security and privacy of smart homes in a user-centered way, specifically investigating users' knowledge of threats, attitudes, and expectations (e.g., [14]). While studies suggest that factoring UX into security and privacy design can be challenging [15, 16], research is needed to explore the practices of designers (or manufacturers) of smart home devices.

To make a step in this direction, we conducted qualitative user studies with three companies that developed smart home security cameras. We interviewed 20 participants from these companies (n=6, n=8, n=6) who were involved in the design process of smart cameras. Our aim was twofold: (1) to understand how companies factored UX into the design of security and privacy solutions and (2) to investigate how UX influenced the design of security and privacy solutions. We summarize our key findings and contributions below:

1. Product design teams used UX as a means of designing innovative privacy solutions.
2. UX was not used to design innovative security solutions due to an explicit need for established security solutions.
3. Data protection regulations triggered security and privacy considerations, but some regulations were regarded as impractical from a UX perspective.
4. Conflicting interests among departments represented in the design team impeded the UX design of security and privacy solutions.

The rest of the paper is structured as follows: we give an overview of relevant literature in Section 2. We describe our methods in Section 3. We present and discuss our results in Section 4 and Section 5, respectively. Finally, we present our design recommendations in Section 6.

2 Related Work

2.1 Security and Privacy in Smart Homes

Several studies have explored users' experiences, values, needs, and concerns in relation to smart home surveillance (e.g., data collection, use, and sharing) [17–20]. Zeng et al. [14] interviewed 15 smart home users and found that their understanding of threats depended on the sophistication of their mental models. Malkin et al. [21] surveyed 116 users of smart speakers and found that they were protective of the audio command history of children and guests, and that they strongly opposed third-party data tracking. Malkin et al. [22] also surveyed 591 smart TV users and found that they disagreed with their data being shared with other parties despite a lack of understanding of regulations that protected their rights. Geeng and Roesner [23] interviewed 18 smart home users to investigate multi-user interactions and found tensions during installation, normal use, and long-term use. Abdi et al. [24] conducted interviews with 17 smart assistant users and found that they had limited understanding of data storage and sharing. Naeini et al. [25] conducted a vignette study with 1,007 users to investigate privacy preferences, and found that users were more comfortable with data collected publicly, and that they would more likely consent to providing data if it were perceived as beneficial. Apthorpe et al. [26] surveyed 1,731 smart home users to measure the acceptability of third-party data sharing. They provided insights into existing privacy norms and extracted best design practices.

Other studies have investigated the concerns and perceptions of bystanders—such as visitors or co-habitants—who do not make the choice to install smart home devices. Yao et al. [27] ran focus groups and design activities with 18 participants and found three factors impacting the privacy perceptions of bystanders. Bernd et al. [28] proposed to use the framework of Contextual Integrity to research the privacy of domestic workers that are affected by smart home devices, and the design process of product teams who build such devices.

2.2 UX of Security and Privacy

As technology use evolves and becomes embedded in everyday life, the focus on usability (i.e., how easy, efficient, and effective technology is to use) becomes necessary, but insufficient. Broader issues need to be considered, such as social communication, contextual trust, and even aesthetic aspects of security and privacy design.

Dunphy et al. argue that it is crucial to understand how UX is factored into the security and privacy design of technologies [29]. However, there are several gaps in this space: Shava and Van Greunen [30] state there is a "missing link" between UX and usable security and privacy. Other researchers have also reported on the lack of scientific research into UX and usable security and privacy [31, 32].

2.3 UX in Smart Home Security and Privacy

There has been an increased focus on designing user-centered smart home devices [14, 23, 24, 33–36]. However, there has been little research into the role of UX in the security and privacy design of smart home devices [37]. Further, there has been little work exploring how designers and their collabora-

tors think about the UX of these devices [38,39]. In particular, there is a research gap in how designers consider UX during the security and privacy design of smart home devices [40,41]. Bergman and Johansson [42] conducted a structured literature review of 150 smart home research papers and found that there was no research into how product teams factored UX into the security and privacy design of smart homes. Without an in-depth understanding of designers' processes, challenges, and responsibilities, we argue that existing security and privacy issues in smart homes will persist.

Existing research has focused mostly on the practices of developers. Assal and Chiasson [43] interviewed 20 developers to explore real-life software security practices during the development lifecycle and found that security was not considered in the design stage. Similarly, Waldman [44] interviewed 36 product developers and found that product teams did not consider privacy in their decision-making. Further, previous research has found that many gaps existed among product teams on the one hand and security teams on the other hand, which included miscommunication and lack of security knowledge [45,46]. As a result, they found that some companies contracted developers who were security experts to act as an intermediary between product and security teams [47].

The literature suggests that security and privacy may pose UX challenges for smart home developers. Oh and Lee [16] analyzed reviews of quantified self applications and found that privacy was a key problem affecting UX, security, and privacy design processes. This was later confirmed by Bergman et al. [15], where they explored how 11 smart home companies captured UX requirements and found that security and privacy posed a UX challenge for designers. Rowland et al. [48] found that smart home designers often faced tensions between UX and security in smart homes (e.g., trade-offs between strong authentication and users' ease of interaction with smart devices). Unlike desktop computing, smart home applications span inter-connected physical and digital devices [49], increasing the complexity of factoring UX into the design of smart devices [50]. While UX and usability guidelines are established for desktop systems and web applications, guidelines that specifically target smart home devices are under-researched [49,51]. Moreover, smart home devices lack standardization and quality dimensions [52]. Some general UX rules are recommended for the design and implementation of smart home devices [53], but their effectiveness and suitability have not been explored in detail [54].

In summary, prior research has uncovered a variety of design-related security and privacy issues from the user perspective for which UX is critical (e.g., the need to consider aesthetic aspects of security and privacy). Researchers have argued for understanding and designing the UX of security and privacy [29]. However, the literature reveals that there has been limited work on frameworks, models, and scientific research bridging UX, security, and privacy. Our work takes a step to solve this problem by investigating the role of UX in the security and privacy design of smart home cameras.

3 Methods

We designed and conducted a qualitative user study of designers of smart home cameras based on approaches described in [55–57]. We interviewed 20 participants in the United Kingdom, focusing on understanding the design processes and practices of smart home cameras manufactured by three different companies A (n=6), B (n=8), and C (n=6). We aimed to investigate the design, development, and implementation of three security camera products that had been in production for years. We concentrated on the design of these products because smart home security cameras (i) have a growing adoption rate [58], (ii) are subject to increased security attacks [59], and (iii) are seen as particularly invasive by end-users [60,61]. Our institution's ethics committee approved this study.

3.1 Research Questions

Our work aims to address the following research question:

RQ. How do product design teams factor UX into the security and privacy design of smart home cameras?

To address our main research question, we explore the following sub-questions:

1. How do designers and their collaborators make decisions during the security and privacy design process of smart home cameras?
2. What are the different aspects of the design process of smart home cameras that explicitly deal with UX factors?
3. What are the challenges that different stakeholders face when factoring UX into the security and privacy design of smart home cameras?

3.2 Recruitment

To recruit our participants, we posted flyers and distributed leaflets in the United Kingdom, and advertised the study on online platforms (e.g., LinkedIn). We also recruited participants through snowball sampling, which allowed us to reach employees that were not easily accessible through other strategies. At the time of recruitment, interested participants were employees who were active at their company and responsible for the design, development, or maintenance of a smart home camera product. The participants we recruited from each company were all on the same development team and worked on the same product.

We asked interested participants to complete an online screening questionnaire (see Appendix A). We received 31 complete responses. In addition to asking demographic ques-

tions, we provided participants with a list of job titles and then asked them to choose the title that best described their position at the company (e.g., UX Designer, Security Engineer, Product Manager). We describe the demographics of our participants in Table 1 in Section 4.

Additionally, we asked participants to provide information about their company. We also asked them to specify the type of products that their company manufactured (e.g., security, lighting, or phone systems), as well as the specific products their company manufactured (e.g., cameras, hubs, voice assistants, lights). Finally, we asked participants to estimate the number of employees who worked at their company. The number allowed us to establish the company size (e.g., startup, mid-size, enterprise) since we were interested in targeting large-scale product development companies that were more likely to put effort into improving the UX of products [62].

3.3 Interview Procedure

We conducted semi-structured interviews with 20 employees working at companies that manufactured smart home devices: Company A (n=6), Company B (n=8), and Company C (n=6). We used the funnel technique [63] to structure our initial interview questionnaire (study script), starting with general questions and then drilling down to specific ones.

The interview started with general questions characterizing participants' role at the company (e.g., responsibilities, duration of employment), the type of products they designed or developed, and their perspectives on UX, security, and privacy. Members of design teams referred to different groups of people (e.g., device purchaser, device administrator, and device user in the house) as 'users' without distinction. We then asked questions related to requirements gathering and specification in the design phase, as well as questions about how UX was factored into the design process (e.g., UX in the security and privacy design process, UX design methods, techniques, and artifacts).

Finally, we asked specific questions related to the profession of participants: regulatory stakeholders were asked about data protection regulations, product liabilities, and regulatory affairs; management stakeholders were asked about roles and responsibilities related to security and privacy; security stakeholders were asked about security requirements, the design of security, security maintenance, and security breaches.

We conducted our interviews remotely using Skype and Zoom. We also audio-recorded and transcribed all interviews. Interviews lasted for an average of 52 minutes. Our interview questions can be found in Appendix C.

3.4 Pilot Study

After creating our initial interview questions (see Appendix B), we conducted a pilot study with four smart home product designers at a local conference. Two researchers recorded and analyzed the pilot interviews. We used the findings to identify potential problems (e.g., adverse events, time, cost) in advance prior to conducting the full-scale study. Drawing from our findings, we made the following changes:

- We refined our interview questions to reduce bias and improve their quality.
- We changed our data analysis method from Thematic Analysis to Grounded Theory because we aimed to (i) develop a substantive theory, (ii) deeply explore design processes, and (iii) derive grounded recommendations.
- We were better informed of the average duration of our interviews, which turned out to be around 50 minutes.

3.5 Data Analysis

We transcribed and analyzed all 20 semi-structured interviews using Grounded Theory, following Strauss and Corbin's procedure [64]. Grounded Theory enables the examination of topics and situations from many different angles, leading to comprehensive and deep explanations. It can uncover beliefs and meanings behind behaviors and events, through examining both rational and irrational aspects of behaviors [65].

Four researchers in total analyzed the transcripts. The primary researcher (who conducted the interviews) and a second researcher independently completed the initial coding of all interview transcripts. To verify the credibility of the initial codes, a third researcher cross-checked the codes against the interview transcripts. At the same time, the fourth researcher reviewed the initial codes and supporting quotes. The four researchers discussed any differences and generated a codebook of 155 codes. The researchers then grouped the codes into themes (axial coding) and categories (selective coding).

We observed data saturation [66–68] between the 18^{th} and the 20^{th} interview; i.e., no new codes emerged in interviews 18–20, and, hence, we stopped interviewing. After creating the final codebook (see Table 4 in Appendix D), we tested for inter-rater reliability. The average Cohen's kappa coefficient (κ) for all codes in our data was 0.81. Cohen's kappa values over 0.80 indicate almost perfect agreement [69].

3.6 Research Ethics

The University of Oxford Central University Research Ethics Committee reviewed and approved the study (CUREC/CS_C1A_19_049). Before each interview, we asked participants to read an information sheet and sign a consent form that presented all the information required. Participants had the option to withdraw at any point during the study.

3.7 Limitations

Security, privacy, and regulatory matters are sensitive issues in big organizations like the ones we interviewed (see Table 2). Our participants' corporate responsibilities as well as their company's reputation might have biased their responses. To mitigate this, we explained to our participants that data would be collected and processed in accordance with the General Data Protection Regulation (GDPR).

Further, self-reporting bias is common in user studies [70]. Some participants might not have responded accurately to our questions because they did not remember specific details or wanted to be viewed as socially acceptable. To maximize validity and minimize self-reporting bias, we avoided leading questions and relied on open-ended questions, inviting participants to provide in-depth answers in their own words.

Finally, our qualitative work is limited by the size and diversity of our sample. Following recommendations from prior work to interview between 12 and 20 participants [71], we interviewed 20 participants until new codes stopped emerging.

4 Results

In this section, we detail the findings of our study. We present our participant demographics (Section 4.1), and then discuss our key findings organized according to the main themes of our analysis, as illustrated in Figure 1. The main themes are:

- Development Process (Section 4.2);
- UX in Security Design (Section 4.3);
- UX in Privacy Design (Section 4.4);
- Innovation in Security and Privacy Design (Section 4.5);
- Trust (Section 4.6).

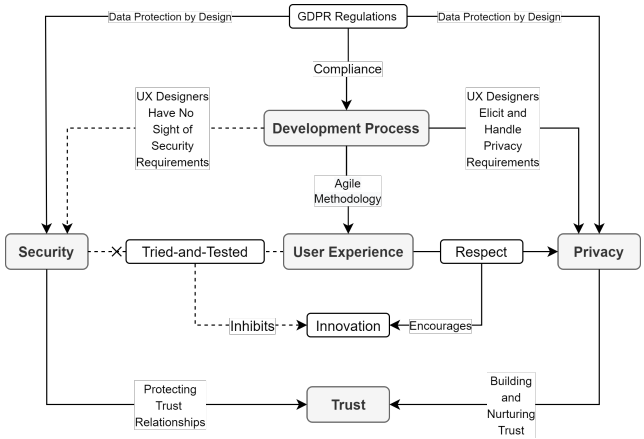


Figure 1: Drawing from the findings of our study, the figure describes a model showing how UX was factored into the security and privacy design of smart home cameras.

4.1 Participant Demographics

Table 1 summarizes the demographics of our sample (n=20). We interviewed 12 male and eight female participants. Ages ranged from 25 to 52. Ten participants had a college (or an undergraduate) degree, and ten had a graduate (or postgraduate) degree. We divided our participants (n=20) into six groups of stakeholders based on employment: security stakeholders (n=4), regulatory stakeholders (n=3), UX stakeholders (n=5), management stakeholders (n=4), software stakeholders (n=2), and hardware stakeholders (n=2).

	Gender	Age	Degree	Experience	Employment
A01	Female	46	M.A.	4 years	Product Manager
A02	Male	28	B.A.	2 years	UX Designer
A03	Female	42	M.Sc.	5 years	Security Manager
A04	Male	30	M.Eng.	2 years	Security Engineer
A05	Male	32	M.Sc.	2 years	Hardware Designer
A06	Male	44	M.A.	6 years	UX Director
B07	Female	52	J.D.	7 years	Legal Counsel
B08	Female	31	J.D.	2 years	Compliance Counsel
B09	Female	38	B.A.	4 years	Experience Designer
B10	Male	43	B.A.	7 years	Product Designer
B11	Male	27	B.A.	2 years	UX Designer
B12	Male	35	M.Sc.	4 years	Security Architect
B13	Male	28	M.Sc.	3 years	Mobile Developer
B14	Female	31	B.Sc.	4 years	Software Engineer
C15	Female	46	J.D.	6 years	Product Counsel
C16	Male	36	B.A.	6 years	Senior UX Designer
C17	Male	29	B.Eng.	2 years	Hardware Engineer
C18	Male	50	B.Bus.	8 years	Product Manager
C19	Female	34	B.Sc.	4 years	Security Engineer
C20	Male	25	B.Sc.	1 year	Software Developer

Table 1: Semi-structured interview participant demographics.

4.2 Development Process

All participants (designers and their collaborators) followed an agile product development process, which included requirements analysis, design, development, testing, and maintenance [72]. In this section, we report on the requirements analysis, design, and development stages that companies A, B, and C (see Table 2) followed to develop products PA, PB, and PC (see Table 3). We describe how GDPR influenced the development process of smart home cameras. We also describe the challenges that UX design activities and smart homes introduced to our participants.

Company	HQ	Employees	Product	Designed in
A	UK	500+	PA	UK
B	USA	1000+	PB	UK
C	USA	1000+	PC	UK

Table 2: Summary of companies.

Product	Type	Major Functionality
PA	Camera	Motion and sound detection with alerts.
PB	Doorbell	Real-time monitoring/audio features.
PC	Camera	Face recognition/detection of intruders.

Table 3: Summary of products.

4.2.1 Processes and Approaches

We briefly describe the processes and approaches of the design teams working at companies A, B, and C. All teams combined hardware and software development in agile and iterative design processes.

Company A. A cross-functional team that involved various stakeholders (e.g., senior UX/UI designers, software and mobile application developers, industrial designers, product managers) was in charge of questioning, exploring, defining, and making decisions related to the design of product PA (a camera). The team ran multiple workshops with designers and developers to explore various ideas and techniques, become familiar with common design patterns, and understand the product's business strategy. They followed a collaborative UX design process (multi-staged UX [73]).

Company B. A self-managing agile team was in charge of the design and development of product PB (a doorbell). The team was composed of different experts (e.g., designers, developers, engineers) who met on a regular basis to share data, communicate, collaborate, and discuss their progress. No managers were controlling or directing the team because team members decided how to prioritize their work, manage their team, and achieve the goals of the project. UX designers were in charge of eliciting functional and quality requirements by applying different UX activities related to the use cases of the product. The requirements were extracted from user needs identified by UX research (e.g., personas, prototypes, interviews).

Company C. A functional team—operating in a traditional organizational structure—adopted agile mindsets, principles, and practices and was in charge of the development lifecycle of product PC (a camera). The team consisted of senior UX designers, product managers, and software developers. The team leader reshuffled team members regularly depending on the project's needs and requirements. Team members met regularly and were familiar with each other's work processes. UX designers, developers, and content designers conducted user research and made explicit UX decisions during the early stages of their projects. The team elicited requirements during the design, development, and implementation phases.

4.2.2 GDPR and Compliance

All three companies were required to comply with GDPR [74], which mandated *Data Protection by Design* (DPbD) [75] practices (as reported by A3, B7, and C15). In practice, a DPbD approach requires companies to *"consider privacy and data protection issues at the design phase of any system, service, or product."* [76]

Delayed effect. GDPR came into force on May 25th, 2018, after the smart cameras of companies A, B, and C had been developed and released. Product Counsel C15 said that the devices produced before the enforcement date were non-compliant with GDPR. Similarly, Legal Counsel B7 said that the company's infrastructure that stored user data was not equipped to deal with GDPR requests. Security Architect B12 stated that making changes in the existing product architecture required an increased demand for labor, money, and effort.

Obtaining consent. GDPR requires smart home companies to obtain clear and valid consent from users to the use of their data. Due to the large amount of data exchanged in the ecosystem of company C's products, consenting to *all* uses of data was described as technically challenging by Product Manager C18. UX Designer A2, who was familiar with GDPR, stated that asking users to consent to all uses of data in their ecosystem would be detrimental to UX. A2 said: "*I think it would be too overwhelming for users to see every single piece of data that we collect.*"

Right to withdraw consent. Under GDPR rules, smart home users have the right to withdraw their consent at any time, which requires companies to delete user data. However, in the case of company C, Security Engineer C19 reported that their smart camera was often used with other company products as well as by third-party products (e.g., Amazon Alexa). C19 explained that the increased number of devices that shared customer data made complying with this regulation demanding. C19 described that their infrastructure "*is not designed to destroy the data just like this, with one click.*" C19 also mentioned that lack of control of data collected by third-party devices made complying with this regulation "*very challenging*." In particular, C19 stated that it was difficult to determine whether third-party devices; e.g., Amazon Echo, were GDPR-compliant due to the lack of clear guidelines showing how third-parties collected and processed user data.

Conflict between business and regulation. Different and conflicting design goals could arise during the design phase. Security Manager A3 reported dealing with a tension between commercial and regulatory stakeholders. The *Legal Department* wanted some of the data collected from users to remain stored on users' cameras (i.e., offline); however, the *Commercial Department* requested all data collected to be stored on the company's cloud servers. A3 explained: "*The legal team asked to keep the data local only, but at the same time the commercial team wanted us to collect it. I guess they wanted to monetize it.*" This reported conflict highlights the important role of regulation in smart homes.

4.2.3 UX Design Activities

Participants reported different challenges during different activities of the UX design process.

Identifying pain points. UX designers (n=2) investigated data monitoring and collection in smart home cameras (through conducting research) to appropriately *"identify the main sort of frustrations and pain points"* (A2). A2 interviewed smart camera users to research *"acceptable areas of monitoring"*; however, they could not identify the pain points resulting from monitoring. To address this problem, A2 interviewed psychologists and visited existing customers in their house. A2 was able to identify six major pain points related to video monitoring. For example, A2 found that customers were concerned about cameras spying on them—by passively collecting data without their knowledge.

Making UX design decisions. UX designers made recommendations, but did not always make design decisions. UX Designer B11 reported that despite conducting user research and suggesting UX-aware changes, they did not make any final decisions; they were instead made by the project manager. B11 said: "*Ultimately, it's always their decision at the end of the day, but it has always been difficult for my job to ensure that I'm providing the best experience possible.*"

Fully understanding user behavior. UX designers (n=3) mentioned that one persistent UX challenge they faced was to fully understand the security behavior of users. A6 commented: "*The reality is we won't know the exact behavior until the product is out.*" The difficulty of understanding real user behavior is a challenge that is well-known to the usable security and privacy community [77].

Making hardware design changes. Hardware stakeholders (n=2) faced issues when applying UX-related changes to existing products. In company A, industrial designers faced difficulties when implementing a privacy feature which would visualize the on/off state of smart cameras. Hardware Designer A5 explained that software developers were more flexible: *"while mobile developers are flexible, we have to make early decisions that are not easy to change."* Similarly, Hardware Engineer C17 said there was not enough time to build hardware sprints. The lack of flexibility (e.g., time, effort) made it difficult to apply UX changes to existing products.

4.2.4 Interoperability of Smart Home Devices

Participants reported two smart home interoperability challenges that occurred during design and development.

Heterogeneous devices. Participants (n=2) stated that the integration between heterogeneous devices and company products was important to companies. For example, most products in company A supported heterogeneous devices and services, such as Amazon Alexa, If This Then That (IFTTT) applications, and Apple's Siri (A1). Third-party services (e.g., Amazon Alexa) "*improve[d] [user] experience*" (A2); however, they created difficulties for security stakeholders. Security Engineer A4—who worked on encrypting the data exchanged between the company's ecosystem of products and Amazon Alexa—described the process as *"complex"* and *"time-consuming."* A4 specifically reported dealing with a legacy platform unable to send and receive encrypted messages with the API of Alexa Voice Service [78].

Securing connections between devices. Security Engineer C19 stressed that their company's smart home devices had *"solid security."* The challenge, however, was encrypting data exchanged among the increased number of devices in the company's ecosystem. C19 explained: *"Smart home devices are generally secure... The problem is the number of connections between all of those devices, they all have to be protected."* In addition, Product Designer B10 explained that the increased connections between devices, touch points, and objects would add to the complexity of this challenge. Complexity in smart homes was previously reported in the literature [79].

4.3 UX in Security Design

In this section, we present how design teams applied UX principles and practices to the design of security features that end-users interacted with. We found that UX was not factored into the design of security solutions due to lack of expertise and the misperception of security being a low-priority technical-only problem. In addition, we found that GDPR and security audits motivated UX considerations.

4.3.1 Alignments between UX and Security Design

Regulations and legal liabilities. We found that regulation triggered security design considerations. Although security was not explicitly factored into company B's design phase, regulations and legal liabilities required designers to consider some security requirements. UX designers in company B attempted to consider regulatory requirements in the design phase although they faced several obstacles (i.e., high-level guidelines). Legal Counsel B7 mentioned that the introduction of GDPR's *"Data Protection by Design"* requirements prompted doorbell design teams to implement new security features (i.e., stronger encryption during authentication).

Security audits looking into user behavior. All three companies conducted security audits to establish how well their information system conformed to security standards and frameworks. We found that security audits in companies B and C prompted security design considerations. During a security audit led by Security Architect B12, a security review was conducted to investigate the password strength of the accounts of PB's users. B12 found several instances of poor password

behavior, which prompted an evaluation of password strength as well as the creation of UX-aware password requirements (e.g., the addition of password strength meters).

4.3.2 Incompatibilities between UX & Security Design

Design of security features was not explicitly anyone's responsibility. Among our participants (n=20), no one took responsibility for the design of security features. Participants (n=5) who handled security design tasks said they were not accountable for security design issues. Those participants did not have UX design expertise. For instance, Product Manager A1, who made security design decisions based on their understanding of common security practices, said that the *Information Security Team* was responsible for all matters related to security, including security design. However, Security Engineer A4 from the *Information Security Team* dismissed any responsibilities related to the design of security features.

Security design was a low-priority concern. For some stakeholders (n=2), security design was acknowledged but perceived as low-priority. As a result, minimal efforts were made to introduce security stakeholders in design teams that handled UX design. Security Manager A3 explained that the budget of the *Information Security Team* was limited, and that adding security experts to the design team was a *"luxury"* they could not afford. Moreover, Product Designer B10 expressed similar thoughts when discussing the addition of a '*usable security expert*' to the design team.

Security was seen only as a technical problem. Many participants (n=15) described security as being only a technical problem that should be addressed from a technical perspective. As a result, participants expected that security to be exclusively handled by developers and security experts. This perception gave little to no consideration to social aspects of security. For instance, when we asked UX Designer B11 why security was not part of the design process, B11 stated that this was *"a development question."* Security Engineer A4 had a similar response: *"Designers do not have any security expertise and it doesn't make sense to expect them to handle security problems."* This finding is not novel, but confirms the existence of a long-term challenge in HCI where security is treated as a technical problem [80], regardless of ongoing efforts to bridge the gap between social and technical aspects of security design [77].

Security features were not designed by usable security or UX experts. In company C, sensitive features related to the connectivity of security cameras, firmware upgrades, and registration were designed by Software Developer C20. Similarly, Product Manager A1 – in company A – did not "*see the value*" of including security experts in the design team, and chose security features – such as authentication – based on their understanding of common security practices.

UX designers had no sight of security requirements. Security requirements were not always present in the UX design phase. Experience Designer B9, who played a core part in the design of the doorbell (PB), said that the requirements he was provided with did not include data protection or security requirements. Similarly, UX Designer A2 explained that the security of registering and processing data was discussed during the design phase. However, there were no requirements related to security design: "*There wasn't specific kind of UX work around data protection or user protection or something like that.*"

Security design considerations were ad hoc. For some participants (n=5), features handling sensitive information (e.g., authentication, software patches, access to video footage) created security design considerations on an ad hoc basis. For example, Mobile Developer B13 designed the software update development process for the doorbell mobile application. B13 strongly valued the design of update features because they realized that these features could be used to deliver security updates. In all five cases, ad hoc design security considerations were triggered by non-experts of security design: management stakeholders (n=3) and development stakeholders (n=2). This finding confirms Assal and Chiasson's study results [43], which suggested that ad hoc security considerations are fragile because non-experts of security design (e.g., developers) could fail to identify security-sensitive features.

Lack of security experts in design teams. The product design teams of companies A, B, and C did not include security experts. In company A, the *Information Security Team* was not involved in the design phase of their smart camera. Justifying the decision, Security Manager A3 stated that all company employees underwent annual training and followed the company's *"information security management framework."*

Security was only considered at the implementation stage. For some security stakeholders, security design was acknowledged but was not seen as a priority. The *Security Team* in company B prioritized working with the *Development Team* over the *Design Team*. Security Architect B12 who worked with the doorbell *Development Team* said: "*I know that we can get involved with designers, but well, it's more efficient to work with the development team.*"

Security was reactive and not proactive. Security stakeholders (n=2) reported that security design was treated as reactive, rather than proactive, in companies A and B. Companies preferred a reactive approach, in which they made security design considerations or changes based on security incidents reported by customers. For example, Security Architect B12 reported that their security teams had implemented multi-factor authentication as an option to secure user accounts after successful account hijacking attacks were reported.

4.4 UX in Privacy Design

In this section, we present how design teams applied UX principles and practices to the design of privacy features that end-users interacted with. We found that UX was factored into the design of privacy solutions in companies A and B through considerations of consent, transparency, and user control. However, in company C, UX was not considered in the design of privacy features due to lack of expertise and relying on a general understanding of privacy issues and product use.

4.4.1 Alignments between UX and Privacy Design

Giving users control. Companies A and B gave customers more control of their privacy settings, which UX designers reported to increase trust. For example, both companies implemented a privacy mode in their mobile application in order to allow users to stop camera monitoring. In company A, designers also aimed to make users "*feel in control*" by adding (1) a visible on/off feature that showed the current state of cameras and (2) a privacy mode to give users "*peace of mind*" (A2)—allowing users to automatically or manually disable cameras when using their mobile application. In company B, UX Designer B11 explained that they added a private mode because their customers shared their cameras with family members: "*We do have a privacy feature in our product which allows you to switch the [...] whenever users want to have privacy. [...] When we interviewed users, we realized that a lot of them share their camera with others, mostly family members.*"

Being transparent with users. Participants (n=7) reported that their company made numerous efforts to be transparent with users. Legal Counsel B7 described that the *Legal Department* at company B worked with UX designers to create user-friendly FAQ pages that explained how their company collected and processed user data, as well as the measures they took to protect data. Further, B7 mentioned that the company constantly reminded users of their right to get their data deleted (by sending regular reminder emails). On the other hand, company A, which had to deal with multiple security vulnerabilities in the past, had recently updated its data breach incident response plan to inform customers of data breaches (A3). UX Director A6 helped the company use best practices to ensure that affected users had the best experience possible.

Obtaining explicit consent from users. UX designers (n=2) described different projects that looked into obtaining explicit consent from users in relation to data collection and sharing. UX Director A6 described an on-going project looking into obtaining consent through visual indicators instead of text-heavy documentation that would be difficult for users to read. UX Designer B11 worked on developing user-friendly consent notifications for the camera's mobile application. In addition, developments were made to allow customers to change their own privacy settings based on their needs.

Ensuring smart home cameras were not 'creepy' or intrusive. UX designers (n=3) conducted user research with the aim to design smart home products that were not 'creepy' or 'intrusive'. In company A, the goal of UX designers was to ensure users felt comfortable with their camera (PA), and that it did not make users feel that it was a "*tool of surveillance*" (A6). To achieve their goal, UX Designer A2 interviewed psychologists and visited existing customers in their house to identify acceptable and non-intrusive "*areas of monitoring.*" In company B, Experience Designer B9 assisted in the design of a feature which allowed cameras to "*detect human activity based on geographic location*." B9 explained that the feature allowed users to automatically disable their smart home camera when they were at home and, hence, the device did not feel "*creepy*."

4.4.2 Incompatibilities between UX and Privacy Design

Designers of privacy features lacked expertise. In company C, privacy features were designed by stakeholders (n=2) who did not possess design or privacy expertise (e.g., developers, product managers). Software Developer C20 designed the privacy mode settings of the camera's mobile application during the development process. C20 made privacy design decisions based on their own understanding of sensitive data. Similarly, Product Manager C18 made privacy decisions related to a feature that allowed family members to disable video monitoring and notifications. Both stakeholders did not refer to any design or data protection guidelines. C18 said: "*We didn't follow any requirements, no. [...] I don't know why, I wasn't aware of any requirements.*"

Some privacy solutions were designed based on a general understanding of product use. Company C's *Product Design Team* appeared to deal with privacy design based on a general understanding of product use, rather than a thorough investigation of the specific context of use. For instance, Product Manager C18 believed that privacy concerns of users would better be dealt with by understanding users in a broad and wider context of user-centered design.

Privacy was not explicitly discussed during user research. The *Product Design Team* of company C did not explicitly discuss privacy during the design phase. Senior UX Designer C16 – who worked with product designers, engineers, and managers – said that privacy was not discussed during the user research phase when user interviews were conducted.

4.5 Innovation in Security and Privacy Design

We found that innovation cross-cut UX with security and privacy. In this section, we describe how innovation seemed to enhance the design of privacy solutions, but also to impede the design of security solutions.

4.5.1 Enablers of Innovation in Privacy Design

New privacy features were supported by qualitative and quantitative UX research. UX stakeholders (n=4) working at companies A and B adopted a mixed qualitative-quantitative approach to build new features that addressed user privacy (e.g., concerns, pain points, expectations) during the design phase. To design features related to camera monitoring, UX Designer A2 conducted qualitative interviews with users as well as observed users in their homes to address any privacy concerns. UX Director A6 conducted quantitative research by collecting and analyzing survey data to design a visible indicator that showed whether a camera was turned on or off. A6 also used existing quantitative data from Google Analytics to prioritize which privacy features to implement. Experience Designer B9 created detailed storyboards and personas to visualize how their doorbell would be used in users' homes and whether it would be intrusive.

New privacy features were evaluated through usability testing. UX stakeholders (n=2) conducted usability testing of new privacy features introduced by company B. This was used to ensure that new privacy features did not negatively affect customer experience. UX Designer B11 delivered usability testing results to the *Product Design Team* based on the analysis of mobile application prototypes. B11 mentioned that among these prototypes, some requirements were related to privacy features. B11 explained that users were observed interacting with and changing privacy settings. Similarly, Experience Designer B9 conducted usability testing of the doorbell privacy features and was able to identify issues that prompted design considerations.

4.5.2 Barriers to Innovation in Security Design

Security solutions were tried-and-tested. Security experts (n=3) mentioned that their companies' *Information Security Team* did not design their own security solutions. Instead, they used existing security solutions in their company's security protection paradigms, a practice known as *tried-and-tested* security. Additionally, non-experts also made security design choices supported by their own understanding of common security solutions. Product Manager A1, who worked with the *Design Team* that did not include security experts, chose the *"username and password"* authentication mechanism since it was familiar, widely-used, and accepted in industry.

New security solutions increased uncertainty. Participants (n=2) explained that incorporating tried-and-tested security solutions avoided uncertainties that arose out of the introduction of new security features. Product Manager C18 mentioned that new security solutions were likely to create usability concerns due to lacking information on how users would interact with such features. Similarly, Security Engineer C19 mentioned that attempts to introduce new security features were discouraged in the *Security Team*. C19 explained that introducing new security features would increase security risks due to lacking the knowledge required to design these features.

4.6 Trust

We found that trust heavily influenced UX design choices: product teams aimed to build customer trust through better privacy experiences, and also aimed to protect trust relationships with their customers through data protection policies.

Building and nurturing trust through privacy experiences. We found concerted efforts in the companies that aimed to build a culture of fostering trust. In company C, Product Counsel C15 explained that employees were encouraged to take an interest in and care about protecting user privacy. Similarly, in company B, Legal Counsel B7 described efforts put into creating a customer-first culture, where user privacy was not only seen in development processes but also discussed and encouraged culturally among product teams.

Protecting trust relationships through data protection policies. Product teams (n=5) used data protection policies to protect their company's reputation and build user trust. Many companies had established policies to deal with security vulnerabilities and attacks. For example, Security Manager A3 reported that his company adopted an incident response plan in case of a breach, in order to maintain its reputation, which we identified as a powerful motivator for companies to take security measures. Similarly, Security Architect B12 reported that their *Security Team* had invested in "*developing well-founded requirements*" for responding to security incidents, even when incidents resulted from users' incompetence (e.g., falling for a phishing attack, a compromised home router). Security Engineer C19 said their company drafted a "*responsible disclosure policy*" which dealt with managing security vulnerabilities reported by users.

Overall, our interview participants identified that customer trust was strongly linked to data protection: security was needed to mitigate loss of trust arising from exploiting security vulnerabilities. Further, user privacy was used by product teams to build and nurture trust relationships.

4.7 Summary

All product teams used an agile methodology to drive the development of their smart home products. We found that the practice of using tried-and-tested security solutions inhibited innovation in security design. In addition, the perception of security being only a technical problem, for which there were 'best-practice' technical solutions, limited the consideration of social and interactive aspects of security. In particular, it created a gap between UX considerations and security design (e.g., UX designers had no sight of security requirements).

Despite the gaps that we found in security design, our results show companies innovated in the privacy design space (e.g., company B created a novel geographic-based privacy feature). Our data shows that UX stakeholders in design teams elicited and handled privacy requirements. The practice of using UX design principles to respect user privacy (e.g., giving users control, avoiding creepiness and intrusiveness) seemed to encourage innovation in the privacy space. Moreover, we found that companies were motivated to preserve a trust relationship and build trust with their customers, as privacy or security failures (e.g., intrusive or vulnerable products) would undermine that relationship. Finally, regulations (e.g., GDPR) legally required design teams to consider data protection by design in their requirements.

5 Discussion

Our results uncover complex challenges and limitations that product designers faced: challenges arising from complying with GDPR; the importance and role of building trust; barriers to factoring UX into security design solutions. In this section, we use our findings to discuss the wider role of innovation in designing security and privacy solutions, as well as the implications of adopting a user-centered agile approach to data protection. We also highlight areas for future work.

5.1 The Role of Innovation in Security and Privacy Design

All novel issues related to smart home security and privacy point to significant challenges where innovative solutions are necessary. Despite recognizing the importance of security in the design process, our results show design and security teams are less innovative due to existing practices and perceptions. These practices include favoring tried-and-tested security solutions or procuring security solutions from reputable vendors. This finding highlights a desire to avoid novelty and a preference to 'follow the crowd' in the design of security.

Further, the perception of security as only a technical problem, for which there are "best-practice" technical solutions, limits design considerations for security solutions (e.g., authentication consisting of only username and password combinations). Many smart home devices are designed for operating in privacy-sensitive environments (e.g., personal spaces). Given the relative immaturity of the smart home device space, tried-and-tested solutions are not particularly suitable, and innovative solutions are required. For example, current designs do not accommodate the diversity of social aspects of smart home security and privacy (e.g., the nuance between a device being in a shared space in a flat-share vs. being in a shared space in a single-family household).

While we found no evidence of innovation in security design, our results show that efforts have been made to innovate in the privacy space. For instance, company B created a geographic location privacy feature which could detect human activity and make their doorbell less intrusive. One reason for this was that companies wanted to preserve their trust relationship with their customers, and privacy failures were seen as potentially "creepy" and "intrusive," which would undermine this relationship.

The current efforts of innovation in privacy design are a good first step, but more is needed. For example, the challenge of communicating and obtaining user consent in smart homes needs to be systematic (e.g., within the same device ecosystem) and coordinated (e.g., among device ecosystems). However, this is currently not the case and highlights the need for better communication and coordination between stakeholders and product teams.

While data protection regulations (e.g., GDPR) appear to be consistent with better UX design for privacy in smart homes, these regulations remain unclear as to whether the same could be true with regards to UX for security design. Security and privacy qualities of smart homes are not the same; however, both are qualities of data protection. It is not clear how much responsibility users should have to ensure the secure operation of their devices. However, some manufacturers blame breaches on users who do not adopt secure practices (e.g., failing to change default passwords). Regardless of where responsibility lies, manufacturers could put effort into improving security experience, making it easier for users to achieve their desired security outcomes. One option would be for data protection legislation to explicitly cover security experiences, as currently there are very few incentives for manufacturers to put additional effort into enhancing the UX of security.

Regardless of whether regulations should encompass UX aspects of both security and privacy, design standards, guidelines, frameworks, and APIs are other options which have not been explored from an innovation perspective. The tensions that exist between regulators and UX designers over communicating the use of data (e.g., despite being required by GDPR, UX designers do not typically ask users to consent to all uses of their data because—otherwise—it would be detrimental to UX) should invite us to find innovative solutions that satisfy both parties: regulators and users.

5.2 Security Design in Agile Development

Agile teams have historically treated security as a technical problem, ignoring its social and interaction aspects [81]. With that in mind, we argue that in an agile setting, security would still not be considered during the design stage and would, hence, remain an implementation problem. In company A, Security Manager A3 described their *Information Security Team* as "*the department of 'no' when it comes to enforcing security.*" This problem has been common in the past where

security teams blocked progress in agile environments with the attitude of "security says no" [82].

Moreover, agile development does not have built-in steps for explicitly dealing with security issues because it was not designed with security in mind [83]. This might explain our results which show that product design teams who used an agile development process did not explicitly consider security issues during the design phase. However, our results show that GDPR required design teams to follow DPbD requirements, in order to build legally-compliant products. We argue that this is a promising step toward better considerations of security design in agile teams, but this is accompanied with noteworthy challenges and barriers, especially in the context of smart home ecosystems.

5.3 Directions for Future Work

In this section, we outline areas for future investigation.

Innovation without hindering security. Our results show that tried-and-tested solutions were highly demanded in companies A, B, and C which preferred reliability and assurance (e.g., reusing best-security practices). Those practices were shown to hinder innovation; however, we believe more research is needed to explore the relationship between UX, innovation, and security. A key issue to uncover is what aspects of security design can be safely innovated, and how UX can be used to design more effective security experiences.

UX-aware data protection guidelines. Our findings show that data protection regulations (e.g., GDPR) influenced the design phase. Our participants reported that GDPR touched on facets of product design but often failed to translate into specific requirements, which caused disparities in the design process. While GDPR requires practitioners to factor security and privacy into the design process, it can bring more confusion to the design table: regulatory requirements have been reported to be high-level and impractical [84]. New techniques and tools are needed to address how data protection regulations and practices can factor the application of UX design principles.

Improving communication among different stakeholders. Our results show poor communication among multi-stakeholder teams where security design happens. In the absence of regular communication among stakeholders, the number of implicit assumptions made increases (e.g., in our study, Product Manager A1 selecting security features based on their own knowledge of common practices) [85]. Similarly, tensions among stakeholders also increase. For example, in Company B, UX Designer B11 was frustrated that they could not make UX-aware decisions. Expecting largely autonomous groups of stakeholders (e.g., security, legal, design, UX) with different goals, motivations, and constraints to speak the same language is unrealistic. Therefore, more research into this area should explore how to make different teams communicate effectively about factoring UX into the security and privacy design of smart home products.

6 Conclusion

Studies and recent events show that security and privacy of smart home products can have detrimental and life-threatening effects on people (e.g., compromised products have allowed attackers to spy on residents and control home networks). Design must consider users' motivations, perceptions, and expectations to enable users to effectively protect themselves when using these products.

While research suggests that factoring UX into security and privacy design is important, the practices of product designers in this space have not been empirically explored. To bridge this gap, we conducted three user studies involving 20 interviews with security camera designers. We analyzed the data using Grounded Theory and found that design teams used UX as a means of innovating in privacy design to address social aspects of privacy, in particular to avoid intrusiveness. However, UX was seen as undesirable for innovating in security design due to the belief that security was only a technical problem where tried-and-tested solutions were the only option. Based on our findings, we conclude with recommendations to improve design practices in smart homes:

Explicitly aim to innovate through UX of security. Tried-and-tested security solutions are preferred by design teams as they provide a measure of assurance that they are effective and reduce vulnerabilities. The challenges introduced by smart homes (e.g., diverse social contexts, varying levels of skill and ability, subtle tensions among stakeholders) are not addressed by current tried-and-tested security solutions. By exploring security through the lens of UX, new ways of simplifying and streamlining interactions can be uncovered. While these innovations may also lead to new security challenges, it is necessary to innovate in order to design better solutions that will eventually become tried-and-tested for smart homes.

Align security and privacy in UX. Our results show that UX of security design is not distinct in practice from UX of privacy design. Many technical aspects of security and privacy design have common principles and, thus, could be considered as part of a single UX domain—instead of being broken down into separate components, such that one is in scope and one is not.

Factor UX into the practice of data protection compliance. The compliance aspect of data protection regulations strongly motivates security and privacy considerations (e.g., DPbD). Our results show that UX can help identify issues with compliance, and suggest more workable alternatives.

Acknowledgments

This research was supported by the 2018-2019 Information Commissioner's Office's (ICO) Grants Programme. The authors would like to thank Elie Tom, Martin J. Kraemer, and the anonymous SOUPS reviewers for their valuable input. George Chalhoub is funded by Fondation Sesam. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ICO, or any sponsor.

References

[1] Junia Valente, Matthew A. Wyn, and Alvaro A. Cardenas. Stealing, Spying, and Abusing: Consequences of Attacks on Internet of Things Devices. *IEEE Security & Privacy*, 17(5):10–21, 2019.

[2] Noah Apthorpe, Dillon Reisman, Srikanth Sundaresan, Arvind Narayanan, and Nick Feamster. Spying on the Smart Home: Privacy Attacks and Defenses on Encrypted IoT Traffic. *ArXiv*, 2017.

[3] Jingjing Ren, Daniel J. Dubois, David Choffnes, Anna Maria Mandalari, Roman Kolcun, and Hamed Haddadi. Information Exposure From Consumer IoT Devices: A Multidimensional, Network-Informed Measurement Approach. In *Proceedings of the Internet Measurement Conference*, pages 267–279, New York, New York, United States, 2019.

[4] Tribune Media Wire. Couple 'Felt So Violated' After Hacker Took Over Smart Home Devices, Began Talking to Them. `https://wtkr.com/2019/09/23/felt-so-violated-couple-scared-after-hacker-targets-homes-smart-devices/`, September 2019.

[5] Joseph Cox and Samantha Cole. How Hackers Are Breaking Into Ring Cameras. `https://www.vice.com/en_us/article/3a88k5/how-hackers-are-breaking-into-ring-cameras`, December 2019.

[6] Kashmir Hill and Surya Mattu. The House That Spied on Me. `https://gizmodo.com/the-house-that-spied-on-me-1822429852`, 2018.

[7] Marc Hassenzahl and Noam Tractinsky. User experience — a research agenda. *Behaviour & Information Technology*, 25(2):91–97, 2006.

[8] Claire Rowland, Elizabeth Goodman, Martin Charlier, Ann Light, and Alfred Lui. *Designing Connected Products : UX for the Consumer Internet of Things*. O'Reilly Media, Inc., Sebastopol, California, 2015.

[9] Jong-bum Woo and Youn-kyung Lim. User Experience in Do-It-Yourself-Style Smart Homes. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing(UbiComp 2015)*, pages 779–790, New York, NY, United States, 2015.

[10] Gregory D. Abowd and Elizabeth D. Mynatt. *Designing for the Human Experience in Smart Environments*, chapter 7, pages 151–174. 2005.

[11] John McCarthy and Peter Wright. *Technology as Experience*. MIT Press, Cambridge, Mass, 2007.

[12] Andreas Jacobsson and Paul Davidsson. Towards a Model of Privacy and Security for Smart Homes. In *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT 2015)*, pages 727–732, 2015.

[13] Ali Dorri, Salil S. Kanhere, Raja Jurdak, and Praveen Gauravaram. Blockchain for IoT security and privacy: The case study of a smart home. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops 2017)*, pages 618–623, United States, 2017.

[14] Eric Zeng, Shrirang Mare, and Franziska Roesner. End User Security & Privacy Concerns with Smart Homes. In *Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security (SOUPS 2017)*, pages 65–80, 2017.

[15] Johanna Bergman, Thomas Olsson, Isabelle Johansson, and Kirsten Rassmus-Gröhn. An Exploratory Study on How Internet of Things Developing Companies Handle User Experience Requirements. In *Requirements Engineering: Foundation for Software Quality*, volume 10753 LNCS, pages 20–36, Germany, 2018.

[16] Jeungmin Oh and Uichin Lee. Exploring UX Issues in Quantified Self Technologies. In *2015 Eighth International Conference on Mobile Computing and Ubiquitous Networking (ICMU 2015)*, pages 53–59, 2015.

[17] Eun Kyoung Choe, Sunny Consolvo, Jaeyeon Jung, Beverly Harrison, Shwetak N. Patel, and Julie A. Kientz. Investigating Receptiveness to Sensing and Inference in the Home Using Sensor Proxies. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp 2012)*, pages 61–70, New York, NY, USA, 2012.

[18] Charlie Wilson, Tom Hargreaves, and Richard Hauxwell-Baldwin. Benefits and Risks of Smart Home Technologies. *Energy Policy*, 103:72–83, 2017.

[19] Charlie Wilson, Tom Hargreaves, and Richard Hauxwell-Baldwin. Smart Homes and Their Users: A Systematic Analysis and Key Challenges. *Personal and Ubiquitous Computing*, 19(2):463–476, 2014.

[20] Meredydd Williams, Jason R. C. Nurse, and Sadie Creese. Privacy is the Boring Bit: User Perceptions and Behaviour in the Internet-of-Things. In *2017 15th Annual Conference on Privacy, Security and Trust (PST)*, pages 181–18109. IEEE, 2017.

[21] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy Attitudes of Smart Speaker Users. *Proceedings on Privacy Enhancing Technologies*, 2019(4):250–271, 2019.

[22] Nathan Malkin, Julia Bernd, Maritza Johnson, and Serge Egelman. "What Can't Data Be Used For?": Privacy Expectations about Smart TVs in the U.S. In *European Workshop on Usable Security (EuroUSEC 2018)*, 2018.

[23] Christine Geeng and Franziska Roesner. Who's In Control? Interactions In Multi-User Smart Homes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI 2019)*, pages 1–13, 2019.

[24] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants . In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 451–466, USA, 2019.

[25] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. Privacy Expectations and Preferences in an IoT World. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 399–412, Santa Clara, CA, 2017.

[26] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. Discovering Smart Home Internet of Things Privacy Norms Using Contextual Integrity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2), 2018.

[27] Yaxing Yao, Justin Reed Basdeo, Oriana Rosata Mcdonough, and Yang Wang. Privacy Perceptions and Designs of Bystanders in Smart Homes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[28] Julia Bernd, Alisa Frik, Maritza L. Johnson, and Nathan Malkin. Smart Home Bystanders: Further Complexifying a Complex Context. In *Proceedings of the 2nd Symposium on Applications of Contextual Integrity*, 2019.

[29] Paul Dunphy, John Vines, Lizzie Coles-Kemp, Rachel Clarke, Vasilis Vlachokyriakos, Peter Wright, John McCarthy, and Patrick Olivier. Understanding the Experience-Centeredness of Privacy and Security Technologies. In *Proceedings of the 2014 workshop on New Security Paradigms Workshop (NSPW 2014)*, pages 83–94, 2014.

[30] Fungai Bhunu Shava and Darelle Van Greunen. Factors Affecting User Experience with Security Features: A Case Study of an Academic Institution in Namibia. In *2013 Information Security for South Africa*, pages 1–8. IEEE, 2013.

[31] Niels Raabjerg Mathiasen and Susanne Bødker. Threats or Threads - From Usable Security to Secure Experience? In *Proceedings of the 5th Nordic cConference on Human-Computer Interaction: Building Bridges*, volume 358, pages 283–289. ACM International Conference Proceeding Series, 2008.

[32] Panagiotis Zagouras, Christos Kalloniatis, and Stefanos Gritzalis. Managing User Experience: Usability and Security in a New Era of Software Supremacy. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 174–188, 2017.

[33] Norbert Nthala and Ivan Flechais. Informal Support Networks: An Investigation Into Home Data Security Practices. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 63–82, Baltimore, MD, 2018.

[34] Eric Zeng and Franziska Roesner. Understanding and Improving Security and Privacy in Multi-User Smart Homes: A Design Exploration and In-Home User Study. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 159–176, Santa Clara, CA, 2019.

[35] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. User Perceptions of Smart Home IoT Privacy. In *Proceedings of the ACM on Human-Computer Interaction*, number 2 in CSCW, 2018.

[36] Noah Apthorpe, Dillon Reisman, and Nick Feamster. A Smart Home is No Castle: Privacy Vulnerabilities of Encrypted IoT Traffic. *ArXiv*, 2017.

[37] George Chalhoub. The UX of Things: Exploring UX Principles to Inform Security and Privacy Design in the Smart Home. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA 2020)*, page 1–6, New York, New York, United States, 2020.

[38] Noura Aleisa and Karen Renaud. Privacy of the Internet of Things: A Systematic Literature Review. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, pages 5947–5956, 2017.

[39] Kai Zhao and Lina Ge. A Survey on the Internet of Things Security. In *2013 Ninth International Conference on Computational Intelligence and Security*, pages 663–667, 2013.

[40] George Chalhoub, Ivan Flechais, Norbert Nthala, Ruba Abu-Salma, and Elie Tom. Factoring User Experience into the Security and Privacy Design of Smart Home Devices: A Case Study. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA 2020)*, page 1–9, 2020.

[41] George Chalhoub and Ivan Flechais. "Alexa, Are You Spying on Me?": Exploring the Effect of User Experience on the Security and Privacy of Smart Speaker Users. In *Human Aspects of Information Security, Privacy and Trust*, Cham, 2020.

[42] Johanna Bergman and Isabelle Johansson. The User Experience Perspective of Internet of Things Development. Master's thesis, Certec, Department of Design Sciences, LUND University, Lund, Sweden, 2017.

[43] Hala Assal and Sonia Chiasson. Security in the Software Development Lifecycle. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)* , pages 281–296, 2018.

[44] Ari Ezra Waldman. Designing Without Privacy. *Houston Law Review*, 55(659):659, 2017. Available at SSRN: `https://ssrn.com/abstract=2944185`.

[45] Tayyaba Nafees, Natalie Coull, Robert Ian Ferguson, and Adam Sampson. Idea-caution Before Exploitation: The Use of Cybersecurity Domain Knowledge to Educate Software Engineers Against Software Vulnerabilities. In *Engineering Secure Software and Systems*, pages 133–142, 2017.

[46] Tayyaba Nafees, Natalie Coull, Ian Ferguson, and Adam Sampson. Vulnerability Anti-patterns: A Timeless Way to Capture Poor Software Practices (Vulnerabilities). In *Proceedings of the 24th Conference on Pattern Languages of Programs (PLoP2017)*, United States, 2018.

[47] Tyler W. Thomas, Madiha Tabassum, Bill Chu, and Heather Lipford. Security During Application Development: An Application Security Expert Perspective. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*, pages 1–12.

[48] Claire Rowland. UX and Service Design for Connected Products. `https://iotuk.org.uk/wp-content/uploads/2018/06/UX-and-Service-Design-IoTUK.pdf`, June 2018.

[49] Michael Onuoha Thomas, Beverly Amunga Onyimbo, and Rajasvaran Logeswaran. Usability Evaluation Criteria for Internet of Things. *International Journal of Information Technology and Computer Science*, 8(12):10–18, 2016.

[50] Jonathan Follett. *Designing for Emerging Technologies : UX for Genomics, Robotics, and the Internet of Things*. O'Reilly Media, Inc., Sebastopol, CA, 2014.

[51] Ali Asghar Nazari Shirehjini and Azin Semsar. Human Interaction with IoT-Based Smart Environments. *Multimedia Tools and Applications*, 76(11):13343–13365, 2017.

[52] Luigi Atzori, Antonio Iera, and Giacomo Morabito. The Internet of Things: A Survey. *Computer Networks*, 54(15):2787–2805, 2010.

[53] Timo Jokela. Assessments of Usability Engineering Processes: Experiences from Experiments. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS '03) - Track 9 - Volume 9*, HICSS '03, page 328, USA, 2003.

[54] Rosa Yáñez Gómez, Daniel Cascado Caballero, and José-Luis Sevillano. Heuristic Evaluation on Mobile Interfaces: A New Checklist. *The Scientific World Journal*, 2014:1–19, 2014.

[55] Sharan B. Merriam. *Case Study Research in Education: A Qualitative Approach*. San Francisco, 1988.

[56] Sharan B. Merriam. *Qualitative Research and Case Study Applications in Education*. Jossey-Bass Publishers, San Francisco, 1998.

[57] Dennis Basil Bromley and Dennis Basil Bromley. *The Case-study Method in Psychology and Related Disciplines*. Wiley Chichester, 1986.

[58] Jack Narcotta and William Ablondi. Smart Home Surveillance Camera Market Analysis and Forecast. *Strategy Analytics*, 2018.

[59] Matt Burgess. The IoT's Security Nightmare Will Never End. You Can Now Search Insecure Cameras by Address. `https://www.wired.co.uk/article/internet-of-things-security-camera-search-location`, November 2018.

[60] Kambiz Ghazinour and Emil Shirima. Privacy for Security Monitoring Systems. *USENIX Symposium on Usable Privacy and Security (SOUPS 2018)*, 2018.

[61] Anuroop Gaddam, Subhas Chandra Mukhopadhyay, and Gourab Sen Gupta. Trial & Experimentation of a Smart Home Monitoring System for Elderly. In *2011 IEEE International Instrumentation and Measurement Technology Conference*, pages 1–6. IEEE, 2011.

[62] Jasper van Kuijk, Heimrich Kanis, Henri Christiaans, and Daan van Eijk. Barriers to and Enablers of Usability in Electronic Consumer Product Development: A

Multiple Case Study. *Human-Computer Interaction*, 32(1):1–71, 2017.

[63] Charles F. Cannell, Peter V. Miller, and Lois Oksenberg. Research on Interviewing Techniques. *Sociological Methodology*, 12:389–437, 1981.

[64] Anselm Strauss and Juliet M. Corbin. *Grounded Theory in Practice*. SAGE Publications, Thousand Oaks, California, 1997.

[65] Anselm Strauss. *Basics of qualitative research : techniques and procedures for developing grounded theory*. SAGE Publications, Thousand Oaks, California, 1998.

[66] Clive Seale. Quality in Qualitative Research. *Qualitative Inquiry*, 5(4):465–478, 1999.

[67] Juliet Corbin and Anselm Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Inc., 2014.

[68] Greg Guest, Arwen Bunce, and Laura Johnson. How Many Interviews are Enough? An experiment with data saturation and variability. *Field Methods*, 18(1):59–82, 2006.

[69] Marry L. McHugh. Interrater Reliability: the Kappa Statistic. *Biochemia Medica*, pages 276–282, 2012.

[70] Victor Jupp, editor. *The SAGE Dictionary of Social Research Methods*. SAGE Publications, London Thousand Oaks, Calif, 2006.

[71] Kathy Charmaz. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. Constructing grounded theory. SAGE Publications, 2006.

[72] Nayan B. Ruparelia. Software Development Lifecycle Models. *ACM SIGSOFT Software Engineering Notes*, 35(3):8–13, 2010.

[73] Lassi A. Liikkanen, Harri Kilpiö, Lauri Svan, and Miko Hiltunen. Lean UX: The Next Generation of User-Centered Agile Development? In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational (NordiCHI 2014)*, pages 1095–1100, 2014.

[74] Paul Voigt and Axel Von dem Bussche. *The EU General Data Protection Regulation (GDPR) : A Practical Guide*. Cham, Switzerland, 1 edition, 2017.

[75] Lee A. Bygrave. Data Protection by Design and by Default : Deciphering the EU's Legislative Requirements. *Oslo Law Review*, 1(02):105–120, 2017.

[76] Information Commissioner Office. Data Protection by Design and Default. `https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-by-design-and-default/`, August 2019.

[77] Mary Ellen Zurko. User-Centered Security: Stepping Up to the Grand Challenge. In *Proceedings of the 21st Annual Computer Security Applications Conference*, ACSAC '05, pages 187–202, USA, 2005.

[78] Marcia Ford and William Palmer. Alexa, are you listening to me? An analysis of Alexa voice service network traffic. *Personal and Ubiquitous Computing*, 23(1):67–79, 2019.

[79] Zhi-Kai Zhang, Michael Cheng Yi Cho, Chia-Wei Wang, Chia-Wei Hsu, Chong-Kuan Chen, and Shiuhpyng Shieh. IoT Security: Ongoing Challenges and Research Opportunities. In *2014 IEEE 7th International Conference on Service-Oriented Computing and Applications*, pages 230–234. IEEE, 2014.

[80] Teng Xu, James B. Wendt, and Miodrag Potkonjak. Security of IoT Systems: Design Challenges and Opportunities. In *Proceedings of the 2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD 2014)*, pages 417–423. IEEE Press, 2014.

[81] Xiaocheng Ge, Richard F. Paige, Fiona Polack, and Phil Brooke. Extreme Programming Security Practices. In *Agile Processes in Software Engineering and Extreme Programming*, pages 226–230, Berlin, Heidelberg, 2007.

[82] James Stewart. Why 'Security Says No' Won't Cut it Anymore - Technology in Government. `https://technology.blog.gov.uk/2016/07/13/why-security-says-no-wont-cut-it-anymore/`, July 2016.

[83] Steffen Bartsch. Practitioners' Perspectives on Security in Agile Development. In *Proceedings of the 2011 Sixth International Conference on Availability, Reliability and Security (ARES 2011)*, pages 479–484, United States, 2011.

[84] Seda Gürses, Carmela Troncoso, and Claudia Diaz. Engineering Privacy by Design Reloaded. In *Amsterdam Privacy Conference*, pages 1–21, 2015.

[85] Ivan Flechais, M. Angela Sasse, and Stephen M. V. Hailes. Bringing Security Home: A Process for Developing Secure and Usable Systems. In *Proceedings of the 2003 Workshop on New Security Paradigms (NSPW 2003)*, pages 49–57, New York, United States, 2003.

A Screening Questionnaire

1. Select your gender:
   - ◯ Male
   - ◯ Female
   - ◯ Other
   - ◯ Prefer not to answer
2. Select your age group:
   - ◯ 18-24
   - ◯ 25-34
   - ◯ 35-44
   - ◯ 45-54
   - ◯ 55-64
   - ◯ 65-74
   - ◯ 75 or older
   - ◯ Prefer not to answer
3. What best describes your job at the company?
   - ◯ UI Designer
   - ◯ UX Designer
   - ◯ UX Director
   - ◯ Interaction Designer
   - ◯ Sales Manager
   - ◯ Software Developer
   - ◯ Security Officer
   - ◯ Legal Officer
   - ◯ Hardware Designer
   - ◯ Hardware Engineer
   - ◯ Product Manager
   - ◯ Other:
4. How long have you been working at your company?
   - ◯ No Experience
   - ◯ Less than 1 year
   - ◯ More than 1 year and less than 3 years
   - ◯ More than 3 years and less than 5 years
   - ◯ More than 5 years and less than 7 years
   - ◯ More than 7 years
5. What best describes your company?
   - ◯ Consultant Company
   - ◯ Product Company
   - ◯ Service Company
   - ◯ Platform Company
   - ◯ Other:
6. Select the smart-home product category (or categories) that your company deals with:
   - ☐ Phone Systems
   - ☐ Smart Lights and Dimmers
   - ☐ Temperature and Climate Control Systems
   - ☐ Security Access Control Systems
   - ☐ Other:
7. Select the type of device(s) that your company deals with:
   - ☐ Cameras
   - ☐ Door Locks
   - ☐ Home Theaters
   - ☐ Hubs
   - ☐ Voice Assistants
   - ☐ Leak Detectors
   - ☐ Lights
   - ☐ Motion Sensors
   - ☐ Power Outlets and Switches
   - ☐ Smoke Detectors
   - ☐ Thermostats
   - ☐ Other:
8. How many employees does your company have?
   - ◯ less than 25
   - ◯ 26-50
   - ◯ 51-100
   - ◯ 101-250
   - ◯ 251-500
   - ◯ 501-1000
   - ◯ More than 1000

B Pilot Interview Questions

1. What company do you work for? What does the company do? What is your role in the company?
2. Can you describe your product design process?
3. Do you consider UX when designing security and privacy solutions for your smart home products? If so, how?
4. What are the typical challenges that you face when designing smart home products? Are there any challenges specific to factoring UX into security and privacy design?
5. Is there anything in the design process that could help address user-centered security and privacy challenges in smart homes? If so, please elaborate.

C Main Interview Questions

Our interviews were semi-structured. We below describe our study script (divided into several sections). The last four sections describe specific questions that we asked to employees who had different responsibilities.

C.1 Characterizations

1. Would you tell us about your role in the company that you work at?
   (a) When did you join the company?
   (b) What are your responsibilities?
   (c) What is your specific role in the development or design process of smart home devices?
2. Would you tell us about the products that you develop?
   (a) Is there a specific product that you focus on developing?
3. Would you tell us about your users (or customers)?
   (a) How would you describe the typical customers that use your products?

C.2 Introductory Questions

1. How would you describe User Experience (UX)?
   (a) What UX characteristics do you regard as important?
   (b) What do you think the role of security/privacy in UX is?
2. Do you think there is a relation between UX and security/privacy?
   (a) (if yes) Could you describe this relation?
   (b) (if no) Could you explain why not?

C.3 Requirements Gathering and Specification

1. How do you identify or specify the requirements of a smart home camera before you design it?
   (a) What kind of requirements do you consider?
   (b) How do you prioritize requirements?
2. Do you handle any security/privacy requirements during the requirements gathering or specification process?
   (a) (if yes) How do you handle these requirements?
   (b) (if yes) Do you consider UX when addressing security or privacy requirements in the design process?
   (c) (if no) What do you think of designers who do so?

C.4 UX Design Process

1. Do you consider UX to be an important factor in the design process of smart home devices?
   (a) Where in the design process do you apply UX techniques?
   (b) Is there a specific UX team or role in a specific department?
   (c) How are decisions made when it comes to UX?
   (d) Do you factor UX into the security and privacy design of smart home cameras? If so, could you give us more details?
2. Does security or privacy play a role in the UX development process that you take part of?
   (a) (if yes) Could you explain the role?
   (b) (if no) What type of effect would it have if it did?
3. Do you collect user data for UX development?
   (a) (if yes) What type of user data do you collect? How do you collect it?
   (b) (if yes) What sort of data-driven methods do you use?
   (c) (if yes) Have you ever handled UX requirements within the context of security and privacy? Can you give us more details?
4. Do you have a UX requirements gathering process?
   (a) (if yes) Could you give us more details?
5. What design processes—including methods, techniques (e.g., storyboards), and artifacts (e.g., personas)—do you use in the context of data protection?
6. Regardless of whether you have a UX requirements gathering process, what do you think the best design practices are (e.g., programming patterns, artifacts)?

C.5 End-user Involvement

1. Do you engage end-users in the development of smart camera products or features?
2. (if yes) How do you involve end-users in the design phase?
   (a) To which extent do you involve end-users?
   (b) Do you consider data security or user privacy during the process? Why/Why not?
   (c) Does the type of product influence whether end-users can be involved? What about security or privacy risks?
3. (if no) What are your thoughts on involving end-users in the design of smart home cameras?

C.6 Ecosystem Considerations

1. Is your product part of an ecosystem of products used in smart home environments?
2. (if yes) Have you faced any obstacles when considering the ecosystem?
    (a) (if yes) What were they? Could you describe the main obstacles? How did you deal with them?
3. (if security/privacy was mentioned) Can you describe how do you deal with security and privacy?
4. (if security/privacy was not mentioned) What do you think of the role of security and privacy in a smart home ecosystem?

C.7 UX Challenges

1. What were the UX challenges that you faced during the design of smart cameras?
    (a) How did you overcome those challenges?
    (b) Were there any challenges without any solutions in sight?
2. (if security/privacy was mentioned) Could you give us more details of the security or privacy challenges?
    (a) How did you address those challenges?
3. (if security/privacy was not mentioned) Smart homes are associated with security/privacy threats. Have you ever experienced challenges specific to UX during the security/privacy design process of smart cameras?
    (a) (if yes) Could you describe the challenges you faced, and how did you address them?
    (b) (if no) What do you think of the role(s) of security/privacy and UX in smart home environments?

C.8 Security Stakeholder Questions

1. Can you tell us how security is taken into consideration at your company?
2. How do you ensure that your product is secure? Is there a process? How does it look like?
3. How do you identify security requirements?
4. Do you work/communicate with the design team? Do you get involved in the security and privacy design of smart cameras?
5. In general, who is responsible for designing the security/privacy features of smart cameras?
6. How do you update the firmware of smart cameras that you sell? Who is responsible for this task?
7. How often does security need to be maintained?
8. If a security breach happens in the physical products sold to clients, who will take responsibility?
9. If your company suffers from a data breach, how will you address this? Do you notify users?
10. How do you make sure that users who use your products are protected when it comes to breaches?

C.9 Regulatory Stakeholder Questions

1. Who deals with GDPR and Product Liability?
2. Do you deal with legislation?
3. How is data protection represented in your organization?
4. Do you interact with any regulatory bodies (e.g., ICO) when it comes to matters of data protection?
    (a) (if yes) What are these matters?
    (b) (if no) Do you think it would be useful to do so?

C.10 Management Stakeholder Questions

1. Are there any restrictions (e.g., legal, security, privacy) that make it harder for you to use customer data for product design or making decisions?
2. What data protection roles/responsibilities are there for:
    (a) Product management (and data management)?
    (b) Product design and development?
    (c) UX, usability, and experience-centered jobs?
    (d) Marketing and sales?
3. What does privacy mean in terms of your products?
4. How do you design for data protection when devices are shared among multiple users?

C.11 Concluding Remarks

1. Do you think there is anything in the design/development process that makes it easier to address user-centered security and privacy challenges in cameras?
2. We have reached the end of the interview. Thank you for talking to us!
    (a) Do you have any questions?
    (b) Do you have any comments you want to add?

D Codebook

Development	Stakeholder Limitations	Internal Security Audits	Obtaining Consent
AB Testing	Stakeholder Responsibilities	Low-Priority Security	Opt-Out Services
Acceptability	Storyboarding	No One's Responsibility	Privacy by Design
Agile Development	Surveys	No Sight of Security Requirements	Privacy Mode
Agile Sprint Reviews	Technical Requirements	Reactive Security	Privacy Requirements
Agile User Stories	Tensions Between Stakeholders	Security by Design	Privacy Settings
Analyzing Requirements	The Complexity of Smart Home Devices	Security Maintenance	Privacy Threats
Company Policies	Time Constraints	Security Mode	Privacy Vulnerabilities
Conceptual Sketches	Usability Testing	Security not Factored into Design	Putting Users in Control
Conflicts Between Stakeholders	Usefulness	Security Practices	Security Management Frameworks
Convenience Aspects	User Cases	Security Requirements	Subjective Awareness of Privacy
Cost Constraints	User Frustration	Security Settings	Transparency
Design Process	User Integrity	Security Threats	UX in Privacy Requirements
Development Process	User Interviews	Security Vulnerabilities	UX of Consent
Development Teams	User Involvement in Design	Subjective Awareness of Security	UX of Privacy
Diffusion of Responsibilities	User Observation	Subjective Security Decisions	Withdrawing Consent
End-User Incompetence	User Personas	Technical-Only Security	**Innovation**
End-User Tests	User Research	Understanding User Behavior	Best Practices
Focus Group Interviews	User Respect	Usable Security	Common Practices
Functional Requirements	Utility	Usable Security Experts	Designing New Devices
GDPR Vagueness	UX Best Practices	UX in Security Requirements	Designing New Features
Hardware Design	UX Challenges	Vulnerability Reporting Programs	Designing New Solutions
Heterogeneous Devices	UX Departments	**Privacy**	Lack of Innovation
Indirect User Feedback	UX Design Process	Collecting/Processing Data (GDPR)	Novel Security Uncertainty
Industrial Design	UX Guidelines	Creepiness	Privacy Innovation
Interoperability	UX Pain Points	Creepiness-Convenience Trade-off	Security Innovation
Legal Compliance	UX Policies	Data Protection by Design	Security Uncertainty
Less Time for Hardware Design	UX Requirements Gathering	Data Protection Practices	Tried-And-Tested Security
No Flexibility to Upgrade Hardware	UX Research	Data Protection Requirements	**Trust**
Non-Functional Requirements	UX Roles	Data Protection Responsibilities	Culture of Trust
Not Enough Hardware Sprints	**Security**	Data Protection Roles	Data Breach Response Plan
Product Liabilities	Ad-hoc Security	End-User Compliance	Experience of Harm
Project Constraints	Afterthought Security	Explicit Consent	Incident Response Plan
Requirements Gathering	Awareness Based on Outside Sources	GDPR Consent	Increasing Trust
Requirements Specification	Awareness Based on Social Influences	GDPR Impracticality	Informing Users of Their Rights
Scope Constraints	Encryption Between Devices	GDPR Delayed Effect	Motivation for Privacy
Stakeholder Authority	Evidence-Based Security Decisions	Handling Sensitive Data	Motivation for Security
Stakeholder Characteristics	External Security Audits	Help Pages and FAQs	Motivation for Trust
Stakeholder Communication	Fit-and-Forget Security	Making Devices not Creepy	Preserving Trust
Stakeholder Knowledge	Incidents	No Sight of Privacy Requirements	Trust Relationship

Table 4: Codebook (Grounded Theory).

An Ethnographic Understanding of Software (In)Security and a Co-Creation Model to Improve Secure Software Development

Hernan Palombo* Armin Ziaie Tabari* Daniel Lende Jay Ligatti Xinming Ou
University of South Florida, Tampa, FL, USA
Email: {hpalombo, aziaietabari, dlende, ligatti, xou} @usf.edu

Abstract

We present an ethnographic study of secure software development processes in a software company using the anthropological research method of participant observation. Two PhD students in computer science trained in qualitative methods were embedded in a software company for 1.5 years of total research time. The researchers participated in everyday work activities such as coding and meetings, and observed software (in)security phenomena both through investigating historical data (code repositories and ticketing system records), and through pen-testing the developed software and observing developers' and management's reactions to the discovered vulnerabilities. Our study found that 1) security vulnerabilities are sometimes intentionally introduced and/or overlooked due to the difficulty in managing the various stakeholders' responsibilities in an economic ecosystem, and cannot be simply blamed on developers' lack of knowledge or skills; 2) accidental vulnerabilities discovered in the pen-testing process produce different reactions in the development team, often times contrary to what a security researcher would predict. These findings highlight the nuanced nature of the root causes of software vulnerabilities and indicate the need to take into account a significant amount of contextual information to understand how and why software vulnerabilities emerge during software development. Rather than simply addressing deficits in developer knowledge or practice, this research sheds light on at times forgotten human factors that significantly impact the security of software developed by actual companies. Our analysis also shows that improving software security in the development process can benefit from a co-creation model, where security experts work side by side with software developers to better identify security concerns and provide tools that are readily applicable within the specific context of the software development workflow.

* Both are first authors. Names are ordered alphabetically.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

1 Introduction

It has long been recognized that human factors play a dominant role in ever-present software vulnerabilities, with substantial research devoted to this area [1–10]. These past efforts have used a variety of research methods including surveys, interviews, controlled experiments, studying code artifacts, and analyzing data collected from secure-coding competitions. It is also understood that there is a fundamental economic problem underlying software insecurity [11], and in general there often appears to be an unwillingness in industry to give code security equal importance as other business considerations, such as time to market and richness of features. It is therefore important to recognize that the (in)security of software produced by software companies is impacted not only by individual developers' knowledge and skills and the types of programming languages/environment they use, but also by the various incentives at play both in the market and at the organizational level. Thus, to produce real impact in secure software development, it is indispensable to study this problem in the context of where the process happens, i.e., in the software companies.

Recent work by Sundaramurthy et al. [12, 13] showed that by employing the anthropological research method of participant observation [14, 15], researchers successfully obtained deep insights into the challenges faced by security analysts in security operations centers (SOCs). Moreover, embeddings in the SOCs allowed researchers to produce both technical and non-technical interventions that improved SOC operations by uncovering and addressing the pain points in the overall work process and environment. Encouraged by the success in that work, we conducted an extensive ethnographic study in a software company, using the same method of participant observa-

tion. Two PhD students in computer science were trained in qualitative methods by an anthropologist and spent 1.5 years of total research time doing fieldwork in the company. They participated in everyday work activities in the company such as coding and meetings, and observed software (in)security phenomena both through investigating historical data (code repositories and ticketing system records), and through pentesting the developed software and observing developers' and management's reactions to the discovered vulnerabilities. The fieldworkers then shared their observations with a larger research team that includes the anthropologist and computer science professors collaborating on this research.

In this paper we report our findings on specific examples where incentive structures, organizational relationships, work flow, and other contextual factors shape how security considerations come into play (or not) during the development and updating of software. Some findings were not what we would have expected before the research and shed light on some deep relationships between secure software development and the various incentive structures inherent in the software industry. We present them in the hope of eliciting a broader look into the secure software development process, taking into account at times forgotten human factors that significantly impact the security of software developed in companies.

Finally, through the fieldwork, the embedded researchers were able to make interventions into the firm's software development process, providing methods and tools to help fix discovered vulnerabilities and prevent similar mistakes from occurring in the future. In doing so, we found that the fact that the fieldworkers were part of the software development team became a major reason that the methods and tools they provided actually worked within the context of the company and were taken up by the developers. This co-creation approach provides a model for how security experts may increase their ability to improve the security of software.

2 Research Methods

The main method utilized in this research was participant observation [14, 15]. This method was developed by anthropologists and sociologists as an effective way to study human behaviors and cultures through participating in daily activities and observing people's behaviors through long-term study (typically more than a year). These activities help researchers obtain a solid understanding of a particular culture and gain insights into subjects' activities, knowledge, and habits. By adapting this approach to work within a software company, we can provide an in-depth examination of the complexity of the software development process, the various incentive structures among the stakeholders impacting human behaviors, and the tight coupling of both technical and human factors that impact software security.

In this research, the participant observers were two computer science PhD students, each of whom underwent systematic training in qualitative research method under the guidance of the anthropologist (Lende) on our research team. Being CS students and possessing a substantial amount of security knowledge enabled them to get quickly immersed into the company's software development process and start observing practices that might have an impact on the software products' security. Being inside the company enabled them to observe both contemporary events as they unfolded, as well as past events studied through ticketing systems and checking the relevant code in the repositories. The students' role in the company – working as if they were an employee of the company – helped with two important assets of our research. First, their daily interactions with the developers while doing regular on-the-job tasks provided a unique angle to observe the subjects' authentic behaviors as they performed their job duties. Second, they not only acted as passive observers but as advocates of software security inside the company. This approach enabled the team to observe how the various stakeholders reacted to discoveries of security vulnerabilities, providing valuable insights into why those vulnerabilities were introduced in the first place and the constraints under which they could be fixed (or not).

Each researcher worked at the company 20 hours a week, spread across three week-days. One researcher worked for 12 months and the other for 6 months. The researchers were not paid directly by the company. However, the company provided both financial and in-kind contributions to this research. In general, the researchers' tasks included debugging existing implementations to find bugs' root causes, writing code fixes or implementing new features, performing code reviews, and software quality assurance. The researchers took field notes about their observations, including both security issues found in the software and everyday interactions with developers and other employees involved in the development process. Notes had two forms: descriptive and insightful. Descriptive notes were intended to be as informative as possible, avoiding personal judgments or opinions. Insightful notes aimed to capture "ah-ha" moments and provide reflective analysis of the situations experienced by the observers.

To derive research insights from the raw notes, we applied the general inductive approach [16], augmented by specific techniques for qualitative data analysis [17]. The initial step was to find patterns that emerged directly from the data themselves. In our research, this process happened via weekly meetings of the larger research team including both the fieldworker(s) and the professors, where comparisons could be made across researchers, discussions could address both the human and technical dimensions of software development in a company, and plans made for further exploration of interesting topics. Identifying themes and links between ideas proved central to the inductive analysis, as well as developing contextual analysis around key examples. Data analysis continued through the coding of field notes based on identified themes. These codes included themes related to software

security, human elements of the work, important explanatory concepts that emerged during the research, and data linked to the key examples. A more detailed description of the coding process as well as the codebook can be found in the Appendix. Research meetings then shifted to further developing our joint understanding of the data and identifying ways to explain the observed patterns, as well as potential solutions to how human and technical factors combined to shape (in)security.

It is important to highlight two unique aspects of our participant observation approach. First, participant observation is often a solo affair in the social sciences; having two embedded researchers permitted the examination of the company from two different but complementary perspectives. The researchers were assigned different tasks, had slightly different hours at the company, and developed relationships with company personnel at different points of time. This dual approach to participant observation increases the robustness and validity of the data from this research. Second, the research team consisted of experts in engineering and social science. This multidisciplinary team participated with the embedded researchers in developing the analysis over months, permitting the identification of themes and ideas that crosscut disciplines and had both theoretical and applied dimensions. This team-based approach to both data collection and analysis is a significant contribution to how this type of research can be done effectively.

3 Context

3.1 The Company and Its Products

At the company, the researchers worked in the same space as four other developers, four support engineers, two network engineers, one customer-facing onboarding specialist, the CTO, a marketing and sales manager, and other staff. The researchers' work focused on two products: a solution for controlling network access and a solution for allowing users to securely access networks remotely. The solutions configured third-party network devices (e.g., routers and access-points), enforced operator-defined access-control policies, and managed remediation flows. Typical customers were medium- and large-size organizations, and common users were IT staff who managed the organizations' networks. Organization end users attempting to connect to its network were prompted first by a captive portal that asked for credentials. Once authenticated, they were asked to remediate any issues that prevented them from complying with policy, e.g., they might be required to download and run a client-side monitoring agent and update their anti-virus software.

3.2 Development Process

The company followed general agile development principles. The development team held a scrum meeting every morning that lasted 15-30 minutes. In this meeting, each developer briefly commented about any progress accomplished or roadblocks encountered the day before and discussed the plan-of-work for the current day. This was an opportunity for developers and managers to give and receive feedback from each other. The meeting was led by the dev team lead. The CTO was usually in the room but did not lead the meeting.

Work was organized, prioritized, assigned, and tracked using ticketing and code management systems. In general, tickets were generated by developers, support techs, or customer-facing specialists, ranked in prioritization meetings held by the dev team lead and CTO, and assigned and tracked by the dev team lead. After implementation, tasks were moved into the peer-review stage in which other developers (often more experienced ones) reviewed any code changes, added pending tasks if necessary, and finally approved merge requests. After code changes were approved by all reviewers, tickets were reassigned for quality assurance and integration testing, which was often done by both developers and support/customer-facing specialists. When all tests had been passed, tickets were marked as "done" and merged into the code repository's development branch. When the set of target features for a release had been implemented, the team lead created a release candidate branch. Every release candidate was tested in-house one last time before being finally moved into release and installed on customer environments.

3.3 Study Participants

The main participants in the study were the four software engineers on the development team where the student researchers were embedded. The dev team lead was an experienced developer who had been at the company long-term and written many parts of the system. Two of the other developers had been with the company for several years and another had recently joined. One developer specialized in front-end development and two were full-stack developers. The researchers also interacted with other personnel at the company, including the CTO, via company meetings, work communications, and everyday activities such as breaks and lunches where people often "talked shop" in informal ways.

3.4 Research Ethics

In our research, the employees of the company (developers, support techs, and managers) were considered human subjects. The study was reviewed and approved by the Institutional Review Board (IRB). Researchers explained the study goals to participants and obtained verbal informed consent from participants. Field notes were anonymized, as well as discussions during weekly research meetings. This paper follows that same anonymization approach. Throughout the paper, we use the term application under study (AUS) to refer to a

specific application in the company's product suite. We also anonymized all product-specific terms in the paper.

One ethical dilemma that emerged during the research was what to do when security vulnerabilities were discovered. Given ethical standards among cybersecurity professionals, we made the decision to present these discoveries to the software development team. This process proved crucial to the further development of the research. Rather than simply observing what happened while continuing to work at the company, the researchers raised these security concerns, and where directed, actively worked on addressing them. This active engagement led the research team to a co-creation model, where research, programming, and security were all ongoing parts of what happened during the fieldwork.

4 A Historical Study of a Security Flaw/Feature: Silently Allow Failed Authentication

In this section we describe a security flaw (or feature) that was encountered by the first researcher soon after he started working in the company. Studying the origin and evolving usage of this feature provided a lens through which we observed a number of dilemmas the developers and company had to deal with. This could be helpful for understanding the root causes of other similar security flaws. We first describe this feature and the methods used to study it, then analyze how it was used in multiple instances, and finally draw some conclusions through our reflective analysis.

In this AUS, the process of authenticating users and devices into the network involved assigning an authentication state to every authentication attempt and subsequent authentication queries. Authentication queries were self-triggered by the AUS and configurable, i.e., AUS operators specified how frequently users and devices must authenticate and the AUS executed the corresponding assessments by querying authentication servers, the operating system on the client's device, or sometimes prompting the user directly through a web browser. The authentication states were: pass, fail, unreachable server, unknown username, and the silently-allow (SA) state. Once an authentication flow entered the SA state, it acquired an SA role, which was associated with a built-in policy that granted full access to the protected network. From a security perspective, SA was a dangerous authentication bypass mode, providing access without full authentication (hence, silently-allow). Specifically, an authentication attempt assigned to an SA state was treated as successful and granted full access to the network the AUS was expected to protect. Further, the SA state was relatively persistent. Once a device acquired an SA role, it remained in this state until forced to re-authenticate (by default every 4 hours), or when an administrator manually added policies that would forcefully change the role assigned to the device.

4.1 Specific Methods Adopted in This Study

In addition to the participant observation research methods discussed before, the researcher also used the company's ticketing system and code repositories to understand the rationale behind the SA feature, particularly why it was introduced in the first place and why it kept being used later. In the ticketing system, the researcher located all records where comments were made about the SA feature. Developers rarely used security-related terms in the tickets' description, so uncovering the different instances of SA required the researcher to correlate ticket information with the implementation code. After an initial SA instance was discovered, the researcher searched the code repositories for potential SA-related terms and variable names, formalized the authentication flows, and set up a lab for demonstrating that the SA state could be triggered at runtime. These led to the discovery of other SA instances not mentioned in tickets.

4.2 Observed SA Instances

The AUS had five configurable authentication back-ends: three databases (one legacy implementation, a second one for network guests, and a third one for admin user accounts) and two integration components (one connecting to LDAP servers, and a second connecting to SAML identity providers). The SA state could be triggered in each of these configurations, under one of the following scenarios: 1) the AUS failed to communicate with some authentication server, 2) the AUS was misconfigured, 3) an unexpected runtime exception occured, or 4) a back-end implementation was lacking.

4.2.1 Broken Integration

Broken integration with customer's backend authentication server was the main trigger of the SA state. In the LDAP authentication flow, the AUS communicated with Active Directory (AD) servers. The connection parameters must be pre-configured by network operators and stored in the AUS's internal database. At runtime, these parameters were pulled from the database and LDAP search queries were executed. If the network connection to the AD server failed (e.g., due to misconfiguration, a physical link failure, DNS down, or timeouts due to request overloads), no domain entries were found in the AD server, or any other runtime exception occured (e.g., invalid credentials or duplicate entries in the AUS database), then the device was assigned the SA role and obtained full access to the network.

Most developers and support techs were aware of this SA scenario and often framed it as a feature that alleviated the configuration burden for network operators. It was often justified by explaining that network operators were more interested in not blocking legitimate users into the network than in protecting their networks from intruders with stricter policies that could affect network usability.

For example, support techs said that there had been scenarios in which customers became very frustrated because policy changes had forced most users out of the network and required them to reauthenticate. According to interviewed participants, network operators would see these scenarios as "network outages" that would cause a high load of help-desk support calls for them, which were not well-received.

It remains unclear how many customers actually preferred this SA solution to one that would have required the customers to fix their own integration issues. A support ticket generated from a customer call four years before the study showed that at least one customer would not prefer this SA solution:

> *"Customer doesn't want failed authentication attempts due to LDAP errors to fall into SA but to try another server."*

There was scarce evidence on customer-facing documentation to support the claim that all customers were well aware of SA authentication. SA was only mentioned as a footnote in the release notes PDF of a version of the product that had been released five years before the study, which read:

> *"A system failure will not cause a network-wide outage and will silently-allow authentication for existing and new users attempting network access."*

When the researcher raised the concern that some customers may not be fully aware of SA, some participants explained that customers would be informed only if they asked about the SA role (which was only visible when triggered, on a secondary page on the UI). Since this SA instance was not seen as a security vulnerability but as a feature, no action was taken to remediate the issue.

4.2.2 Misconfiguration

The second way in which the SA state would be triggered was when the AUS was misconfigured. Specifically, a drop-down menu in the UI for policy creation in the administrative portal allowed operators to select an SQL authentication option that mapped to a non-functional legacy database. This would result in assigning an SA state to all authentication attempts, regardless of which credentials were entered by users. Developers explained:

> *"This authentication method is probably broken. I believe it has been deprecated a long time ago."*

When asked why the UI was still showing this option, they said, *"I don't know why but it should not be there."*

We were unable to find any written documentation other than the code to confirm if and when this authentication method had been officially deprecated, or if any customers were still using it. An internal testing ticket from three years back suggested that this authentication method had already been deprecated, but some customer-facing documentation still listed SQL authentication as a possible authentication method. The issue was documented on a ticket by the researcher but at the time of this writing was not yet prioritized for development.

4.2.3 Unexpected Runtime Errors Due to Implementation Bugs

The SA state could also be triggered by unexpected runtime errors. Several instances of authentication code were surrounded in try-catch blocks that would catch SQL and other runtime exceptions and set authentication state directly to SA. Exceptions were somewhat common across the AUS and sometimes caused it to halt operation. Some SQL exceptions occurred after upgrades that resulted in tables with missing attributes, or because the AUS had incorrect database permissions. Other runtime exceptions included null pointer exceptions and out-of-bound array access.

Like the broken-integration case, this instance of SA was an intentional choice. The code that implemented SQL server authentication was added more than 15 years ago and since then had been revisited a few times. Although not explicitly stated, it is possible that this SA instance was an ad-hoc solution for dealing with code complexity and legacy implementations, allowing the AUS to continue execution despite any incomprehensible bugs. The code preceeding the catch blocks looked complicated (with several sections commented out). From reading the code, it was hard to tell what code paths could be executed in each scenario.

4.2.4 Unimplemented Protocol Flows

Perhaps the most critical SA instance was an unimplemented SAML[1] authentication flow that allowed users directly into the network without even checking the credentials against the SAML identity provider. Unlike previous SA instances which were somewhat acknowledged by developers, developers said that they were unaware of this SA instance and believed previous developers who no longer worked at the company were responsible for this problem. When asked about the impact of the problem, some developers said that

> *"Very few customers are probably using SAML authentication on their networks."*

Yet, in a separate conversation, a support tech said that he knew of at least ten customers who were using SAML for authentication.

The SAML SA instance was fixed by the researcher by implementing the missing SAML authentication flow.

[1] Security Assertion Markup Language, an XML-based standard language for communicating security assertions, often used in single sign-on protocols.

4.3 Reactions from Developers

According to members of the support team, SA was an "effective" solution to reduce customers' frustration for the number of help-desk calls that customers received when end-users experienced network interruptions. SA addressed customers' requirements for a product that would deliver an easy-to-use and frustration-free experience for both network operators and end-users. However, we found that in general, developers were somewhat hesitant to talk about SA, either because they did not understand how that part of the software worked or it was code they were not proud of.

Attempts to explain the security implications of SA to the developers and get the problem fixed was not a straightforward process. A common perception is that security vulnerabilities are introduced into code because developers are not aware of the security issues involved, and explicit exposure to the issues would allow them to understand and take immediate corrective actions to remediate their software. This was not our experience in the case of SA. First, for three out of the four SA instances described above, developers were aware of them. And for all the SA instances, they acknowledged that it was problematic. However, even after we brought these issues to their attention, the developers still did not implement the expected security fixes. Moreover, reactions after being exposed to the insecure code were not always consistent with subsequent behaviors, which suggested that there were other reasons why the SA problem persisted in the code. Some examples are explained below.

- Blaming misconfiguration and broken integration issues on customers. Developers complained that customers' limited understanding of their product and networking was the main reason why they weren't able to configure the product within their networks correctly. However, sometimes even the most senior developers spent days to get the product and network configured right.

- Writing limited documentation about authentication flows, which made them difficult to understand for anyone other than the code's original authors. While initial explanations were that there was not enough time to write documentation, months later some participants admitted that some areas were not documented because they believed their implementation could be wrong.

- Not practicing what they preached with respect to testing practices. Everyone in the development team said things like *"we should all do more testing"* and some asked the researchers to write very detailed test cases. But often they did not hold themselves to the same standards and wrote very few tests for the code they wrote. There were also many trivial tests in the codebase from the time when they were using code coverage tools. Some developers admitted that at times they would write trivial tests and minimize code changes just to get the code coverage numbers required for a release.

- Blaming previous developers who were no longer working for the company. Developers often talked about what previous developers did was wrong, but it was not clear whether anyone attempted to correct the problematic code prior to being released.

In summary, the attempts of the researcher's intervention in the SA case were mostly unsuccessful. The developers seemed to be aware that this security flaw could be a major issue for the software and company in the future. However, of all the SA instances identified by the researcher, only the SAML flow was fixed (by the researcher). For the other instances, developers did not see great value in fixing them. They also thought that if they had to fix all security bugs the company would go out of business.

4.4 Lessons Learned

Incomplete stories told by developers, attitude-behavior inconsistencies, and poor documentation were good reasons to believe that the introduction of the different instances of SA deserved deeper analysis. In addition to the evidence directly related to SA, the embedded researchers collected information about interactions among participants during design meetings, prioritization meetings, and informal discussions. Because the company's organizational structure was organic, they were able to interact not only with developers but also with support techs, network engineers, and managers. These interactions provided valuable information for analyzing the deeper motives behind the introduction of the different SA instances. Moreover, the research team identified relationships that connected observed behaviors of the participants with the evidence found on the ticketing systems, the code, the internal wiki, and release notes. The main insights learned from the analysis are described next.

1. Vulnerabilities are sometimes introduced to make errors unnoticeable in an attempt to reduce the number of customer support tickets.

One of the effects of SA was that it would make runtime errors unnoticeable by customers. This reduced the chances that customers would complain that the AUS was not working correctly. When developers talked about these complaints, they implied that customers would blame the company if the AUS could not communicate with other servers, even if the problem was extraneous to the AUS. As one developer stated:

> *"If the system breaks because we followed the specification and the system cannot talk to another server because they are not following the spec, we are probably going to lose money. So we need to code to prevent that."*

Moreover, preventing those errors from bothering customers would imply that less integration-support tickets would be generated, reducing the chances that some of those tickets would be escalated to the development team. Because developers often complained about how much time they spent debugging issues reported by customers (documented on integration-support tickets), it is likely that developers introduced SA in part to alleviate their job stress.

Integration-support tickets were created by support techs assisting customers in integrating AUS with their other network products (e.g., routers and switches). Whenever support techs were unable to resolve tickets in this category, they would escalate them to the development team for further investigation and potential development of bug fixes or custom integration code. Although developers understood that assisting support techs was necessary, they appeared to be more interested in developing new features than fixing bugs or writing custom integration code (which would later require more effort to be maintained).

Further, debugging the issues described on the tickets was challenging because it often required setting up environments that were similar to their customers', which was difficult because of the diversity of network device vendors. In this context, SA reduced the number of integration-support tickets, so developers would spend less time debugging integration problems, and have more time to develop new features.

2. Managers and developers prioritize tasks by doing a heuristic cost-benefit analysis, ranking tasks by urgency and effort required, and security improvements rank low and are usually not fixed because they are not considered urgent or easy to implement in practice.

In general, tasks were prioritized based on heuristic urgency and effort estimations for cost-benefit analysis. Task urgency was often estimated by development managers who were also part of sales and support-prioritization meetings. Urgency estimations tried to measure how much positive or negative impact some new feature or bug fix would have on the business. For example, the development of a new feature could help win or lose a new deal, or a bug fix could help retain or let go of an existing customer. Task effort estimations were often briefly discussed at scrum meetings and more in-depth during "scrum poker" and prioritization meetings. In scrum poker meetings developers estimated the effort required to resolve certain development tasks by assigning them numbers in a scale from 1 to 100. Every week, task priorities were re-evaluated and reassigned or put on hold if necessary.

Of the four instances of SA, at the time of this writing our field researcher was only assigned to fix one, the unimplemented SAML flow, which was considered to be more urgent. Managers and developers believed that fixing the other SA instances would have no positive impact on the business, so they considered them not urgent, low priority, and thus did not address them.

3. Some security vulnerabilities were introduced by leaving deprecated features in production code, and this could avoid breaking existing implementations.

Another possible reason why SA was still in production code was that developers wanted to avoid introducing potential issues that could be caused by fixing the problem. For instance, when asked why misconfigurations SA (Section 4.2.2) was still there, developers' answer was that nobody took the time to remove it. In fact, removing the SA implied a risk, i.e., some other part of the software could break on production systems. Thus developers likely would rather not to take this risk, especially because there was no need for it (customers were not demanding it). In summary, there were just not enough incentives to remove it.

4. Some security vulnerabilities are not detected during development partly because testing is not always embraced by developers.

Any of the SA instances could have been detected if the appropriate tests were executed. However, because developers didn't necessarily like testing and could just write tests to pass the minimum testing requirements, the tests were ineffective and thus would not detect the SA vulnerabilities.

5 Live Discovery through Pen-testing during Ethnography

This section describes how the research combined vulnerability discovery with participant observation of developers' behaviors and reactions. This technical-ethnographic combination is similar to the method used in studying the SA issues, but is also unique. Unlike in the SA case, none of the pen-testing discovered vulnerabilities were intentionally introduced by developers, and as such the researchers had the opportunity to observe the unfolding of developers' reactions with the discovery of a totally unknown problem. It also provided the opportunity for researchers to intervene in a way that resulted in a co-creation model, in which security experts work jointly with developers to improve code security.

5.1 Specific Methods Adopted in This Study

Penetration testing, also known as pen-testing or ethical hacking, is an authorized simulated cyber-attack process against a computer system to reveal security flaws. The goal is to identify weaknesses which might provide a passage for unauthorized users to gain access and alter the integrity of the computer system. There are multiple software pen-testing methodologies, including the Open Web Application Security Project (OWASP) [18], Open Source Security Testing Methodology Manual (OSSTMM) [19], NIST SP 800-115 [20], Penetration Testing Execution Standard (PTES) [21], and Information System Security Assessment Framework (ISSAF) [22]. Results vary based on the way the process is performed.

Two of the company's products were picked by the researcher for further study. By acquiring some information and insights about the applications, the researcher started to apply a customized penetration testing methodology. The basic information such as the software workflow, authentication information, and so on was captured by talking with developers and the support team. Both products were designed to work on a web platform, so OWASP's top ten security vulnerabilities [18] were chosen for the testing.

At the same time as developing the pen-testing process, the researcher worked to gain the developers' trust. This process required building rapport, an understanding with research informants, by participating in daily tasks and getting to know individuals who worked there. The role of a security pen-tester also needed accurate planning and time management. Code injection was selected for the first vulnerability to be tested. It is one of the well-known vulnerabilities which allows attackers to inject malicious codes into a computer system and change the course of execution. The result of a successful injection can potentially be catastrophic. Before we describe the pen-testing findings and developers' reactions, we first briefly introduce the three types of vulnerabilities found.

5.2 Vulnerabilities Found

- Cross-site Scripting (XSS): generally found on web platforms. Attackers typically use web applications to inject malicious codes into the application which can be viewed by other users.
- HTML Injection: similar to XSS. However, instead of inserting malicious scripts, the attacker can inject valid HTML tags and modify the content of the target website.
- Shellcode Injection: a type of vulnerability that allows an attacker to inject malicious code into a system and provide the attacker a shell on the system.

5.3 Behaviors and Reactions from Developers

During the first day of pen-testing, an XSS was found in the AUS by the researcher. The vulnerability was brought up to the developer team, and a proof of concept was provided for why it was significant. While they showed interest in the finding, since the vulnerability was in a 3rd-party application integrated with the AUS, their first reaction was to hope the problem had been fixed by the 3rd-party. One participant said:

> *"This vulnerability belongs to our 3rd-party application, and we did not develop this part. It is better to upgrade the software and see if we will still have the issue"*

They also mentioned that it would be more interesting if the researcher could find any vulnerability inside part of the company's code. Thus, they expressed interest in security, but saw solving this problem as the responsibility of an outside group even though the application formed part of the company's software.

In the next round of testing, the researchers tried other parts of the software to see if there were any other vulnerabilities. Multiple XSS vulnerabilities were found. Developers were both excited and concerned about the findings. They said things like,

> *"If they want to test more, it seems that they will find more things inside our software"*

and

> *"We tried to minimize our bugs, but it seems something is wrong."*

Once again,the third-party issue came up:

> *"We are using Angular, and I thought we shouldn't have the XSS. Angular should take care of this issue."*

On the same day, the researcher found another vulnerability in the AUS; this time it was shellcode injection. The vulnerability allowed attackers to inject their customized shellcode into a valid file and upload it into the server and get backdoor access with a powerful user's account on the server. The attacker must be someone who already had a regular account inside the AUS web platform. The finding was interesting to the researchers and the developers for different reasons. For the research team, this was a critical vulnerability – customers should never have escalated access to the server. They should only be able to perform some limited commands on the OS such as changing the network IP address. This essentially allows their customers to jail-break out of the sandbox set up on the server. Developers on the other hand were more interested in understanding how access had been gained. They said things like,

> *"Interesting! Could you show us how you got the access?"*

and

> *"What is your user's privilege?"*

At the same time, they discussed the risk in the context of the product,

> *"because we already ship the OS to the clients with everything inside it, it's kind of okay! They have the box already. ...We do not have any important information on it."*

During the subsequent discussion with the developers, one researcher asked, *"Do you have any hard-coded password or credentials?"* The answer was *"yes."* This hard-coded

credential would allow a customer who successfully exploited the shellcode injection vulnerability to see other customers' information. Faced with this fact, the developers indicated that this vulnerability was bigger than they originally thought, and that they should take imminent action on it. However, that did not come to fruition at the next group meeting, where they continued to talk about these vulnerabilities. Based on this discussion, the research team inferred that there were other factors that affected the developers' actions. In fact they said *"fixing the vulnerability has additional impacts and may cause some problems for other parts of the application or customers."* They also downplayed the significance of the shellcode injection vulnerability, and said that the team should focus on developing new features. They added,

> *"If we want to fix every bug in our system, we will be out of business very soon."*

Contrary to the researchers' initial hope, our intervention effort to fix the discovered security vulnerabilities proved ineffective in the context of the company's overall functioning. This moment helped the researchers realize that they needed to rethink how security researchers engage developers to create positive change.

It started by simply offering to work on the issue and building the tools and libraries for them to prevent XSS. The bug was then fixed by one of the researchers. One challenge the researcher faced was that the AUS was uniquely designed and would only accept specific types of input entry for the various fields. As a result, it was not possible for the researcher to utilize standard input sanitization solutions, e.g., one that removes all special characters, because that would break the application. Working at the company and interacting with the developers helped the researcher understand this uniqueness and come up with a customized solution. It was a number of specially designed regular expressions that enforce the proper formats for the various types of fields. The researcher included the application of these regular expressions in a standalone Javascript file that can be invoked at the front-end pages. It turned out that the company's existing code already contained a similar mechanism for checking other properties of front-end input fields, e.g., if a field is empty. The researcher only needed to extend this mechanism to include the regular expression checks he designed for preventing XSS. Developers could then simply invoke these checks in the same manner they had been doing for the other types of checks. This allowed for easy integration of the security check into existing code with minimum change, and was readily accepted by the development team. For back-end input sanitization, the researcher first tried to apply standard OWASP sanitization functions for Java, which was the language the back-end was written in. However, due to the uniqueness of the formatting requirement of the AUS, those standard checks were blocking some legitimate inputs. Thus the researcher needed to customize those OWASP functions to work properly with the AUS's requirements. During the fixing process, another XSS was found in the AUS. When the researcher brought up the problem to the development team this time, they accepted it very fast. A new ticket was created and the researcher was asked to fix the issue as soon as possible: *"...Go ahead and fix this bug as well."*

This example highlights the importance of "being there" for security experts to drive positive change for secure coding. The researcher was able to accomplish this in this case due to two factors: 1) he understood the company's existing code and designed an effective security check that minimized disruption; 2) he provided the needed security expertise in designing the proper checks using regular expressions and the customization of the OWASP functions, and this expertise was delivered through code artifacts that were readily applicable within the existing software workflow. Both factors were important for this success.

The researcher tried later to bring up the shellcode injection vulnerability once more in a discussion and tried to convince them to start fixing the issue, but the suggestion was turned down. One of the developers responded,

> *"It's somewhere in our backlog. We didn't do anything about it, and no one has found that exploit so far. So we are safe."*

This comment matched similar instances where developers reacted as though if there were no problems at present, the vulnerabilities might not be an issue that needed urgent attention. The research team considered a possible explanation why the shellcode injection vulnerability was not treated as urgently as the XSS. Exploiting the shellcode injection vulnerability would require a rogue player that can be held accountable (a customer's IT staff member who possessed the regular account access to the server). This may have alleviated the concern on the company's liability resulting from this vulnerability.

Later on in the research, an HTML injection was found inside a newly developed part of the code. Like the XSS, the issue was brought up to the developers. Initially, developers mentioned,

> *"Angular should cover it and not allow the HTML tag in the code! It seems it does not."*

During the next group meeting, they recognized that they had omitted security issues previously. They said,

> *"When we discussed the development of the page, we talked about everything except security and XSS problems. They didn't come to our mind."*

At this time, the researcher thought that the developers would ask him to fix this issue like in the XSS case, but they started to fix it by themselves and did not ask the researcher for any help. Most interestingly, the developers created correct

solutions to fix the HTML injection vulnerabilities, based on the way the researcher solved the XSS problem. This showed that the developers learned from the researcher how to create security fixes within their code base, without being explicitly taught so. They learned by simply observing the code artifacts created by our researcher.

This is an example that illustrated the importance for security experts to be in the development environment and "co-create" security solutions with the developers. The difference in the developers' reactions in this case, compared to earlier ones, pushed the researchers to realize that co-creation happens more in the moment, rather than trying to retroactively fix things. Our earlier interventions mainly focused on fixing vulnerabilities found in code written in the past. We had success in getting some fixed (by the researchers). Whereas in this case, the developers took their own actions and fixed the bugs using the knowledge and tools provided to them by the researchers. This shows that if security professionals are present and part of the team when a product is in the process of being designed and implemented, their views are more likely to be taken into account when decisions about what to do are being made. It was also the researcher's feeling that the quickness with which the development team accepted his suggestion to fix this issue was related to the increased level of trust he enjoyed from the development team at this point in the research progress.

5.4 Analyzing the Findings

After initially finding the first bug in the pen-testing process, the researcher assumed that developers did not know about these security problems, and lack of security knowledge led them to write code with the vulnerabilities. After working with them on various tasks, he realized that they actually possessed quite a bit of security knowledge. As our research progressed, group discussions and analysis of field notes highlighted some non-intuitive reasons for developers' behaviors. This indicates that there were other significant factors in causing these vulnerabilities. We outline these factors in the rest of this subsection.

5.4.1 Developers Should Not Totally Trust Programming Languages and Frameworks

One of the important conversations that the research team had with the developers was that the developers believed that the programming language/framework should take care of some vulnerabilities by default.

"...Angular should take care of this vulnerability..."

In this case and according to Angular documents [23], the Angular engine could handle most of the XSS and HTML injection attack scenarios by sanitizing the input fields. Angular documents also mentioned that developers need to take care of backend servers to make sure injection vulnerabilities are not introduced there. After analyzing our field notes carefully, we found that the developers believed (incorrectly) that Angular could handle all XSS and HTML injection vulnerabilities.

In the past decade, programming languages and frameworks have been doing a great job in creating built-in security measures to prevent accidental mistakes by developers, but they still do not offer a comprehensive security solution. How can developers know accurately where they can rely on language/framework and where they must rely on their own code to achieve a security property? Can this be communicated in a way that does not require sophisticated knowledge on all possible ways attack could happen?

5.4.2 Outsider vs. Insider, and from Deficit Model to Co-Creation Model

One of the embedded researchers in the past worked as a pen-tester for four security consulting companies in three countries for four years. In his experience, the pen-testers were not incorporated into the development team. The developers might only receive a document with discovered vulnerabilities and statements about what they should or should not do. This appeared to be a common industry approach to software security pen-testing [24]. The problem was that the security pen-testers did not understand how much workload the developers had, nor the actual reasons for the vulnerabilities. As a result, this approach did not often lead to the desired changes in the development process, but set up an outsider/insider dynamic, where developers felt the need to defend what they had done and/or minimize the security issues. The developers would say that the report came from an outside group who did not really understand how software development was done, and the security pen-testers would say that the developers wrote defective code in the first place and did not appreciate security, otherwise they would have done something to fix all those problems. Having these past experiences, in contrast with what he experienced in this research where he worked *inside* the development team as a software pen-tester, helped the research team to understand the impact the outsider/insider dynamic had on effectuating changes in software development processes.

From our fieldwork experience, we clearly see how this outsider/insider dynamic can play out. When we first found the vulnerabilities about SA, XSS, or code injections, the researchers' initial thought was that the developers did not know about these security issues. However, after researchers explained to developers and developers had clearly understood the technical details, still some vulnerabilities were not fixed. It was only after further communication with the developers, reflecting on other relevant observations made by the researchers, and brain-storming among the larger research team, that we better recognized why some vulnerabilities were not prioritized to be fixed. Most importantly, *it is when we had*

this understanding, and produced an easy-to-apply solution that fit into the company's development workflow, that our intervention was the most successful.

The point of view that if developers know better and work harder, they should be able to write software without any security flaws, can be characterized as the so-called "deficit model," where the problem of software insecurity is attributed to the developers' lack of knowledge or efforts. The solution driven by this deficit model would mainly involve experts explaining to developers the various software security issues and how to prevent them, and hoping this would drive the needed changes. Research in fields such as education, anthropology, and science communication have examined how using such a deficit model does not prove as useful as imagined because it localizes the problem inside the person and assumes that simply fixing that internal lack will also successfully address larger concerns such as successful learning, cross-cultural understanding, and the application of science to at times controversial topics [25–27]. One analytic concept of note that emerged through the research was our own use of a "deficit model" to initially interpret why people in the company did not respond to security concerns. We assumed that they might not have the knowledge or awareness to understand security risks and recognize how and why particular aspects of the software might increase those risks. Our research found that this deficit model-driven approach was not working well. Simply communicating security issues found and presenting solutions for fixing them did not lead to the anticipated fixes.

Overcoming this "deficit model" in our own thinking helped us to better interpret why participants responded or not to security issues and to recognize how security concerns existed alongside other factors that shaped their work. We then developed a co-creation model, where developers and security experts collaborate together. Co-creation is a form of collaboration in which ideas and processes are shared and improved together rather than kept to only one-party side. By having a co-creation model, security auditors have the chance to jump into the development process and provide the knowledge and tools that developers can readily apply to prevent vulnerabilities. Part of this co-creation model meant that our embedded researchers did not work exclusively on security but dealt with different tickets. This showed the developers that the researchers knew how to program, and could do so as part of a team, while also having expertise in security that they could draw on if needed.

It appeared to us that developers prefer to trust a person inside their team rather than an outsider. Moreover, our field notes showed that a security person inside the developer team can provide more in-depth knowledge than outside resources such as pen-testing reports, internet, and so on. For example, after the XSS got fixed on the AUS, when developers faced the HTML injection they said: *"Is this HTML Injection going to be easy to fix? It should be very easy to fix,"* and without asking the researchers to provide the solution for them, they fixed the issue.

5.4.3 Thinking as an Attacker, Thinking as a Developer

It has almost become a platitude in the security field that one must "think as an attacker." Applying this to software development, the developers can put themselves in the attackers' shoes and understand how software may be misused. It is an interesting question as to *how much* developers need to think as an attacker. These days, understanding the mechanisms of all types of cyber attacks can be overwhelming even to a security expert. Our data implied that the developers and the company were aware of some of those threats that they may face, but just knowing them was not enough. The problem is not necessarily about the lack of understanding the attackers; it is more about not being able to implement security features correctly into software, which unfortunately requires some non-trivial amount of security knowledge. Is it realistic to expect all software developers to become security experts? How much time should developers spend on thinking about how their code may be attacked, among all the other competing demands they face? Security professionals can help bridge this gap by starting to think like developers, just like how we ask developers to think like attackers. Security professionals need to better understand how developers have to negotiate many competing interests, not just a sole focus on security. This could help in providing security knowledge and information at the right level of abstraction that can be easily integrated into the software development process. The co-creation model we used as part of this research allowed the security researchers to think like a developer, and to create some positive impacts in the software development process.

6 Discussion

Our ethnographic study found that software security or lack thereof emerges from a network of technologies and humans, rather than happening solely because of deficits in the software and/or in developers' knowledge or efforts. The people in this company were trained professionals acting in good faith to create successful products. They faced dilemmas that can be common in the software industry and aimed to resolve them in ways that produced a viable product, established good relationships with customers, maintained profitability, and enhanced usability and security. Security issues were not directly attributable to deficits in knowledge, but rather both emerged and could be resolved in terms of the dynamics that shaped how developers dealt with both technical and human demands and balanced their primary aim to develop successful features for the software with their understanding that security is an important part of software integrity and functioning. The co-creation model emerged by attending to these dynamics in the workplace, and responding to what worked and did not work

to address specific vulnerabilities during the daily work of development. This co-creation model could form an integral part of a workable solution to improve software security. In fact, some success in applying secure development life cycle (SDL) in industry echoes our findings [28]. Whether the co-creation model could work under financial realities, and what other ingredients need to be there to form a working solution, remains an intriguing question for future research.

7 Related Work

Assal and Chiasson [1, 2] utilized interviews and surveys to explore the interplay between developers and software security processes. Their research found that developers were motivated to develop secure code, but were often hindered by a mismanaged organizational process. The authors advocated looking beyond developers and examining broader organizational factors that may impact the security of the developed software. Our work is one such attempt, utilizing an extensive ethnographic study in a software company. Many of our findings confirmed the analysis results from Assal and Chiasson's work. Our work also revealed some deeper insights into the reason of software (in)security, as well as a co-creation model that can help address them.

Ruef et al. [3] and Votipka, et al. [4] conducted a series of studies based on data collected from the Build It, Break It, Fix It (BIBIFI) contests. A number of patterns of developer mistakes leading to vulnerabilities were analyzed. Our work examined the software development process in a real company. Our in-depth ethnographic study is complementary to the analysis based on large-scale competition data. One possible cross-over between the two types of studies is that one can use the insights from one to drive the analysis in the other. For example, an observed real-world phenomenon that has significant security impact could be replicated in the BIBIFI contest to further examine a hypothesis on a much larger and more diverse population.

Oorschot and Wurster [5] posited that developers have different skills which often do not include security and suggest that the focus should be on those who design APIs, because it is unrealistic to expect all developers be taught sufficient security. We raise a similar question in our paper from our ethnographic data, regarding how much security knowledge developers can realistically master, and whether a co-creation model where security experts and developers closely collaborate would be a more effective approach.

Green and Smith [6] discussed that developers are not the problem for insecure code. The focus should be on creating more developer-friendly and developer-centric approaches and supporting them when they are dealing with the security tasks. Our ethnographic data supports this conclusion. Moreover, our fieldwork resulted in a co-creation model that could be part of a solution to provide the needed support to developers for writing more secure code.

In addition to the works mentioned above, the research community has explored this area through a number of angles. Oliveira et al. [7] conducted surveys to understand developers' attitudes toward security which leads to understanding that APIs and tools can be improved significantly. Votipka et al. [8] performed semi-structured interviews to compare how hackers and testers find vulnerabilities. Acar et al. [9] studied whether different documentation resources influence the security of programmers' code. Naiakshina et al. [29] conducted a qualitative study with 20 computer science students and investigated how and why they failed with regards to secure password storage. Gorski et al. [30] designed a controlled online experiment with 53 participants to study the effectiveness of API-integrated security advice. Acar et al. [10] conducted an online study and evaluated five cryptographic APIs with GitHub Python developers about the usability of the crypto APIs. There has also been research that studied and characterized different aspects of software bugs [31–33]. These studies focused on the quality of bug reports and found that important information was often missing in bug reports which made it harder to reproduce and fix them.

To the best of our knowledge, our work is the first in using participant observation and long-term ethnography to study secure software development in a real company. It allowed us to observe and reflect upon all contextual factors that have an impact on secure development processes in a real company. Our data and findings serve to complement the efforts discussed above, and often times reveal deep insights not obtainable through other approaches.

Going beyond secure software development, research into other aspects of usable security has also revealed the importance of incorporating broader stake holders' perspectives in thinking about security solutions [34, 35]. Haney et al. studied the role of cybersecurity advocates within organizations [36–39]. Much of the findings in that line of research echoes ours, in particular the importance of co-creating security solutions with relevant stake holders.

8 Conclusion

This research shows how security intersects with software development on the ground, based on two embedded researchers with one and half years' data. There remains a considerable gap between security and developers. Our research shows that security professionals can better bridge the gap by understanding how (in)security emerges from the interacting technological and human factors in the development process. Our ethnographic study provided a way to understand this complicated phenomenon, both by better understanding the competing demands under which developers work, and by demonstrating how security can successfully be integrated into software development through a co-creation model.

Acknowledgments

We thank Raj Rajagopalan for bringing us the idea of using anthropological methods to study secure software development in its native environment, and continuous support for this research. We thank Michael Hicks and Michelle Mazurek for insightful discussions throughout this ethnographic study. The authors of this paper owe all the above, and the anonymous reviewers, for the many valuable comments that helped tremendously in improving the paper. We owe gratitude to the company, and its employees who participated in the study. This research is supported by the National Science Foundation under Grant No. 1801633 and 1801545. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Hala Assal and Sonia Chiasson. Security in the software development lifecycle. In *14th Symposium on Usable Privacy and Security*, Baltimore, MD, USA, 2018.

[2] Hala Assal and Sonia Chiasson. Think secure from the beginning: A survey with software developers. In *CHI Conference on Human Factors in Computing Systems*, Glasgow, UK, 2019.

[3] Andrew Ruef, Michael Hicks, James Parker, Dave Levin, Michelle L Mazurek, and Piotr Mardziel. Build it, Break it, Fix it: Contesting secure development. In *2016 ACM SIGSAC Conference on Computer and Communications Security*, Vienna, Austria, 2016.

[4] Daniel Votipka, Kelsey Fulton, James Parker, Matthew Hou, Michelle L. Mazurek, and Michael Hicks. Understanding security mistakes developers make: Qualitative analysis from Build It, Break It, Fix It. In *29th USENIX Security Symposium*, Boston, MA, USA, 2020.

[5] Glenn Wurster and Paul C Van Oorschot. The developer is the enemy. In *New Security Paradigms Workshop*, Lake Tahoe, CA, USA, 2008.

[6] Matthew Green and Matthew Smith. Developers are not the enemy!: The need for usable security APIs. *IEEE Security & Privacy*, 14(5):40–46, 2016.

[7] Daniela Oliveira, Marissa Rosenthal, Nicole Morin, Kuo-Chuan Yeh, Justin Cappos, and Yanyan Zhuang. It's the psychology stupid: how heuristics explain software vulnerabilities and how priming can illuminate developer's blind spots. In *30th Annual Computer Security Applications Conference*, New Orleans, LA, USA, 2014.

[8] Daniel Votipka, Rock Stevens, Elissa Redmiles, Jeremy Hu, and Michelle Mazurek. Hackers vs. testers: A comparison of software vulnerability discovery processes. In *2018 IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, 2018.

[9] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. You get where you're looking for: The impact of information sources on code security. In *IEEE Symposium on Security and Privacy*, San Jose, CA, USA, 2016.

[10] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L Mazurek, and Christian Stransky. Comparing the usability of cryptographic APIs. In *IEEE Symposium on Security and Privacy*, San Jose, CA, USA, 2017.

[11] Ross Anderson. Why information security is hard – an economic perspective. In *17th Annual Computer Security Applications Conference*, New Orleans, LA, USA, 2001.

[12] Sathya Chandran Sundaramurthy, Alexandru G. Bardas, Jacob Case, Xinming Ou, Michael Wesch, John McHugh, and S. Raj Rajagopalan. A human capital model for mitigating security analyst burnout. In *11th Symposium On Usable Privacy and Security*, Ottawa, ON, Canada, 2015.

[13] Sathya Chandran Sundaramurthy, John McHugh, Xinming Ou, Michael Wesch, Alexandru G. Bardas, and S. Raj Rajagopalan. Turning contradictions into innovations or: How we learned to stop whining and improve security operations. In *12th Symposium on Usable Privacy and Security*, Denver, CO, USA, 2016.

[14] James P. Spradley. *Participant Observation*. Holt, Rinehart, and Winston, 1980.

[15] Kathleen M. DeWalt and Billie R. DeWalt. *Participant Observation: A Guide for Fieldworkers*. Lanham: AltaMira Press, second edition, 2011.

[16] David R Thomas. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, 27(2):237–246, 2006.

[17] H Russell Bernard, Amber Wutich, and Gery W Ryan. *Analyzing qualitative data: Systematic approaches*. SAGE publications, 2016.

[18] OWASP top 10 most critical web application security risks, 2017. https://www.owasp.org/images/7/72/OWASP_Top_10-2017_(en).pdf.pdf.

[19] The open source security testing methodology manual. https://www.isecom.org/OSSTMM.3.pdf.

[20] Karen Scarfone, Murugiah Souppaya, Amanda Cody, and Angela Orebaugh. Technical guide to information security testing and assessment. NIST Special Publication, 2008.

[21] Penetration testing execution standard (PTES), 2014. http://www.penteststandard.org/.

[22] Information system security assessment framework (ISSAF), 2019. https://untrustednetwork.net/files/issaf0.2.1.pdf.

[23] *Agular security*. https://angular.io/guide/security#xss.

[24] Michael Felderer, Matthias Búchler, Martin Johns, Achim D Brucker, Ruth Breu, and Alexander Pretschner. Security testing: A survey. *Advances in Computers*, 101:1–51, 2016.

[25] Hee-Je Bak. Education and public attitudes toward science: Implications for the "deficit model" of education and support for science and technology. *Social Science Quarterly*, 82(4):779–795, 2001.

[26] Susan D Blum. Why don't anthropologists care about learning (or education or school)? an immodest proposal for an integrative anthropology of learning whose time has finally come. *American Anthropologist*, 121(3):641–654, 2019.

[27] Tom Humphries. Schooling in American sign language: A paradigm shift from a deficit model to a bilingual model in deaf education. *Berkeley Review of Education*, 4(1):7–33, 2013.

[28] Steve Lipner. Lessons learned through 15 years of SDL at work. Online Opinion Piece: https://www.csoonline.com/article/3440120/lessons-learned-through-15-years-of-sdl-at-work.html, September 2019.

[29] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why do developers get password storage wrong? a qualitative usability study. In *ACM SIGSAC Conference on Computer and Communications Security*, Dallas, Tex, USA, 2017.

[30] Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke, Christian Stransky, Sebastian Möller, Yasemin Acar, and Sascha Fahl. Developers deserve security warnings, too: On the effect of integrated security advice on cryptographic API misuse. In *14th Symposium on Usable Privacy and Security*, Baltimore, MD, USA, 2018.

[31] Jorge Aranda and Gina Venolia. The secret life of bugs: Going past the errors and omissions in software repositories. In *IEEE 31st International Conference on Software Engineering*, Vancouver, Canada, 2009.

[32] Philip J Guo, Thomas Zimmermann, Nachiappan Nagappan, and Brendan Murphy. Characterizing and predicting which bugs get fixed: an empirical study of microsoft windows. In *32nd ACM/IEEE International Conference on Software Engineering*, Cape Town, South Africa, 2010.

[33] Silvia Breu, Rahul Premraj, Jonathan Sillito, and Thomas Zimmermann. Information needs in bug reports: improving cooperation between developers and users. In *22nd ACM conference on Computer Supported Cooperative Work*, Savannah, GA, USA, 2010.

[34] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel von Zezschwitz. "If HTTPS were secure, I wouldn't need 2FA"- End user and administrator mental models of HTTPS. In *IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, 2019.

[35] Constanze Dietrich, Katharina Krombholz, Kevin Borgolte, and Tobias Fiebig. Investigating system operators' perspective on security misconfigurations. In *25th ACM SIGSAC Conference on Computer and Communications Security*, Toronto, Canada, 2018.

[36] Julie M. Haney and Wayne G. Lutters. "It's scary... it's confusing... it's dull": How cybersecurity advocates overcome negative perceptions of security. In *14th Symposium on Usable Privacy and Security*, Baltimore, MD, USA, 2018.

[37] Julie M. Haney, Mary Theofanos, Yasemin Acar, and Sandra Spickard Prettyman. "We make it a big deal in the company": Security mindsets in organizations that develop cryptographic products. In *14th Symposium on Usable Privacy and Security*, Baltimore, MD, USA, 2018.

[38] Julie M. Haney and Wayne G. Lutters. Motivating cybersecurity advocates: Implications for recruitment and retention. In *Computers and People Research Conference*, Nashville, TN, USA, 2019. Association for Computing Machinery.

[39] Julie M. Haney and Wayne Lutters. Security awareness in action: A case study. In *Workshop on Security Information Workers, USENIX Symposium on Usable Privacy and Security*, Santa Clara, CA, USA, 2019.

Appendix: Coding

Coding – including the development of specific codes and of a codebook – proceeded in an iterative fashion. Weekly meetings facilitated the discussion of emerging research themes and specific examples. Codes emerged from these discussions, where the researchers built consensus on specific analyses. The researchers relied on a general inductive approach [16], as well as techniques derived from grounded theory and related approaches for doing qualitative analysis [17]. Overall, the development of codes initially focused on the Silently Allow example, then on emerging results from penetration testing, the development of a specific coding system for each researcher for their field notes, and a final collaborative phase to find commonalities in codes for both the specific examples and overall corpus of field notes.

Initial Coding of Silently Allow

The coding for Silently Allow grew from weekly meetings and initially involved using flash cards to organize information. These flashcards then became organized into different types of Silently Allow, which were written up and shared with the research group for further input. Here are the written notes on the first two categories:

Type 1

Device allowed network access. Device not authenticated but needs network access to function. Silently Allow an "internal state" that is used to get devices onto customers' network.

This is the original "Silently Allow" (SA)

In this case, not a problem with the actual code. They programmed the code to do this. Some other SAs are because of problems with the code itself.

**Need vignette – use field notes and participant observation to describe the situation.*

Potential category: IoT devices
Initial problem: Customer

Type 2

Saml – authorization protocol, a way to give access to the network. This is a protocol problem, because hadn't implemented the necessary checks

This is an issue with users who input credentials and/or third party authentication server.
If user tried to use Saml, went into SA automatically. And thus gets network access.
This problem doesn't affect all users; affects users who are set up with Saml authentication. Users had a single login; with SA, once clicked login, would be able to login, no matter what. Didn't implement security checks. Whatever result, grant access.

Potential category: Authentication
Initial problem: Internal/program

Full Set of Codes

Subsequent research focused on developing a full set of codes by each of the student researchers. Because the two embedded researchers often worked on different projects and at different times, each wrote their own field notes and then subsequently engaged in coding of their own notes. This process permitted inductive analysis from their own data, which could then be shared in research meetings to produce consensus. Below are the sets of codes developed by the two embedded researchers:

Codes Developed by Researcher One

- Ad-hoc-development
- Customer-driven-development
- Debugging-infrastructure-vs-security
- Developers-not-trained-in-security
- Difficult-to-fix
- Inconsistent-narratives
- Ineffective-peer-reviews
- Insecure-defaults
- Insecurity-to-avoid-breaking-implementations
- Lack-of-documentation
- Lack-of-security-audits
- Lack-of-security-awareness
- Learn-by-doing
- Legacy-code
- Limited-testing-practices
- Misconfiguration
- Non-urgent-issues
- Outdated-libraries
- Positive-bias
- Rapid-prototyping
- Reactive-security
- Revenue-streams
- Security-vulnerabilities-documented-as-bugs
- Selling-unimplemented-features
- Social-frictions
- Technically-challenging-problems
- Time-spent-debugging
- Time-to-market-vs-security
- Trivial-tests
- Underestimating-effort
- Unmanageable-code
- Urgencies
- Usability-over-security

Codes Developed by Researcher Two

- Caring-about-subject
- Changing-attitude
- Development-process
- Documents-not-updating-frequently
- Fix-first-update-later
- Joking-about-intern-work
- Knowing-bug-do-nothing
- Lack-of-knowledge
- Learning-process-with-company
- Looking-for-new-idea
- New-idea-vs-tasks
- Not-caring-about-subject
- Not-trusting-other-developer-or-intern
- Performance-reaction
- Protective-about-subject
- Say-something-do-something-else
- Security-vs-performance
- Security-vulnerabilities-blocked
- Security-vulnerabilities-concern
- Security-vulnerabilities-denying
- Security-vulnerabilities-execution
- Security-vulnerabilities-fixing
- Security-vulnerabilities-interested
- Security-vulnerabilities-process
- Security-vulnerabilities-reaction
- Security-vulnerabilities-thinking
- Security-vulnerabilities-upgrading

Final Collaborative Phase

In the final collaborative phase, researchers often worked on a whiteboard to find the overlap between different types of data, specific examples, and inductive insights. For example, "co-creation" emerged as an overarching conclusion that came out of working through the data to find commonalities in both field notes and in researcher experience during participant observation. Reviewing similarities, via the coding and then the notes, also led to data-driven conclusions about what was successful and what proved to be bottlenecks or limitations in cybersecurity during the months of embedded research.

Why Can't Johnny Fix Vulnerabilities: A Usability Evaluation of Static Analysis Tools for Security

Justin Smith
Lafayette College
smithjus@lafayette.edu

Lisa Nguyen Quang Do
Google
lisanqd@google.com

Emerson Murphy-Hill
Google
emersonm@google.com

Abstract

Static analysis tools can help prevent security incidents, but to do so, they must enable developers to resolve the defects they detect. Unfortunately, developers often struggle to interact with the interfaces of these tools, leading to tool abandonment, and consequently the proliferation of preventable vulnerabilities. Simply put, the usability of static analysis tools is crucial. The usable security community has successfully identified and remedied usability issues in end user security applications, like PGP and Tor browsers, by conducting usability evaluations. Inspired by the success of these studies, we conducted a heuristic walkthrough evaluation and user study focused on four security-oriented static analysis tools. Through the lens of these evaluations, we identify several issues that detract from the usability of static analysis tools. The issues we identified range from workflows that do not support developers to interface features that do not scale. We make these findings actionable by outlining how our results can be used to improve the state-of-the-art in static analysis tool interfaces.

1 Introduction

Security-oriented static analysis tools, like Spotbugs [12], Checkmarx [2], and CodeSonar [3] enable developers to detect issues early in the development process. Among several types of code quality issues, developers rank security issues as the highest priority for these tools to detect [22].

Evaluating the efficacy of these security-oriented static analysis tools has been a popular topic for researchers [17, 29, 51, 63]. However, prior work has largely overlooked the usability of these tools, instead focusing on functional properties such as the types of vulnerabilities tools detect (or fail to detect), false alarm rates, and performance. Chess and McGraw argue that usability is essential to actually make software more secure: "Good static analysis tools must be easy to use, even for non-security people. This means that their results must be understandable to normal developers who might not know much about security and that they educate their users about good programming practice" [21].

Unfortunately, developers continue to make mistakes and need help resolving security vulnerabilities due to the poor usability of security tools [33]. Recently, Acar and colleagues set forth a research agenda for remedying usability issues in developer security tools, explaining that "Usable security for developers has been a critically under-investigated area" [14]. As part of that research agenda, they call for usability evaluations of developer security tools. This approach has been successfully applied in the adjacent field of end-user security tools [25, 30, 31, 58, 60]. For instance, Whitten and Tygar conducted cognitive walkthroughs to identify usability issues in PGP, an end user security tool [60]. We use a similar evaluation technique, namely heuristic walkthroughs [55] in combination with a user study, to identify usability issues in developers' security-oriented static analysis tools.

We evaluated four security-oriented static analysis tools, Find Security Bugs [13], RIPS [10], Flawfinder [7], and a commercial tool. (Our license agreement with the tool vendor stipulates that we anonymize the commercial tool; we refer to it as CTool throughout the paper.) To our knowledge, this study is the first to identify usability issues across multiple developer security tools using heuristic walkthroughs. As a result of this study, we identified several usability issues, ranging from missing affordances to interfaces that scale poorly. Each of these usability issues represents an opportunity to improve developers' security tools. Alongside these usability issues, we contribute our visions for how toolsmiths and security researchers can improve the usability of static analysis tools for security. Ultimately, by improving the usability of these

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

tools, we can enable developers to create secure software by resolving vulnerabilities more accurately and efficiently. To support replication, we make our study setting available in a virtual machine [6], along with the study protocol [5] and the detailed list of usability issues we identified for the four tools [8].

The contributions of this paper are:

- A heuristic walkthrough and a user study evaluating the usability of the interfaces of four static analysis tools.
- A categorization of usability issues that serves both as a list of known pitfalls and as a list of opportunities to improve existing tools.
- Design guidelines and discussions that illustrate the actionability of these issues.
- Specifications for a low-cost heuristic walkthrough approach that researchers and practitioners can use to improve additional tools.

2 Related Work

We have organized the related work into two categories. First, we will discuss relevant work concerning usability evaluations of end user security tools. Next, we will discuss prior evaluations of developer security tools.

2.1 Usability Testing End-User Security Tools

Several studies have evaluated the usability of end user security tools. Through these evaluations, researchers identified usability issues in various end user security tools, ranging from encryption tools [49] to Tor browsers [25]. Collectively, these studies have improved the usability of end user security tools by contributing a better understanding of how users interact with these tools. In their foundational work, Whitten and Tygar studied the usability of PGP, using a combination of a cognitive walkthrough and a laboratory user test to identify aspects of the tool's interface that failed to meet a usability standard [60]. Their study revealed issues such as irreversible actions and inconsistent terminology.

Since the Whitten and Tygar's study, others have successfully applied similar approaches to study the usability of additional end user tools. For instance, Good and Krekelberg studied the usability of the Kazaa P2P file sharing system [31]. Their findings suggest that usability issues led to confusion and privacy concerns for users. Similarly, Gerd tom Markotten studied the usability of an identity management tool using a heuristic evaluation and cognitive walkthrough [58]. Reynolds and colleagues conducted two studies to understand multiple aspects of YubiKey usability [52].

Clark and colleagues conducted cognitive walkthroughs to examine the usability of four methods for deploying Tor clients [25]. Based on their evaluation, they make recommendations for facilitating the configuration of these systems. Also studying the usability of Tor systems —through a user study instead of a cognitive walkthrough—Gallagher and colleagues conducted a drawing study to elicit users' understanding, and misunderstandings, of the underlying system [30].

Like these previous studies, we are concerned with the usability of security tools and strive to better understand usability issues by conducting an empirical evaluation. We are encouraged by these studies' successful applications of evaluation techniques like cognitive walkthroughs and heuristic evaluations to end user tools. In contrast to these prior studies, we evaluate static analysis tools to identify usability issues in the domain of *developer* security tools.

2.2 Evaluating Developer Security Tools

Several studies have conducted comparative evaluations of developer security tools. For instance, Zitser and colleagues evaluated five static analysis tools that detect buffer overflow vulnerabilities, comparing them based on their vulnerability detection and false alarm rates [63]. Comparing a wide range of Android security tools, Reaves and colleagues categorize tools according to the vulnerabilities they detect and techniques they use [51]. Their results do also include some usability experiences, such as how long it took evaluators to configure the tool and whether output was human-readable. Austin and colleagues compare four vulnerability detection techniques, including static analysis, with respect to the number of vulnerabilities found, false positive rates, and technique efficiency [17]. They conclude that multiple techniques should be combined to achieve the best performance. Emanuelsson and Nilsson compare three static analysis tools, Coverity Prevent, Klocwork K7, and PolySpace Verifier, in an industrial setting [29].

There have been a limited number of studies that account for usability in their evaluations of developer security tools. Imtiaz and colleagues [38] study developer actions on Coverity warnings to determine how it helps fix bugs. They show that despite the quick fixes and the low complexity of the warnings, developers still take a disproportionately large amount of time to fix them. Assal and colleagues conducted a cognitive walkthrough evaluation of Findbugs to determine how well it helped developers determine the number of vulnerabilities in a codebase [16]. Based on the usability issues identified in this study, the authors created a tool, Cesar, designed to be more usable. Gorski and colleagues conducted a participatory design study of security warnings generated for cryptographic APIs [32]. They find that design guidelines for end-user warnings are insufficient in this context. Nguyen and colleagues describe some usability issues that affect Android lint tools to motivate the design of their tool, FixDroid, which uses data flow analysis to help secure Android apps [44]. However, the descriptions of Lint's usability issues are not based on a formal evaluation. Smith and colleagues conducted a user study which identified 17 categories of developers' information needs while using a security-oriented static analysis

tool [56]. Thomas and colleagues leveraged Smith and colleagues' framework to evaluate the usability of ASIDE, an interactive static analysis tool [57]. Our work differs from these prior studies because we study the usability (rather than the technical capability) of static analysis tools.

The studies closest to our own focus on the usage and usability of different analysis tools. Sadowski and colleagues describe the Tricorder static analysis ecosystem at Google [54]. They also provide guiding principles based on their experience with Tricorder. Some of these guidelines emphasize the importance of usable static analysis. For instance, they argue that analysis tools should fix bugs, not just find them. Johnson and colleagues [40] interview 20 developers on their experience with the static analysis tools they use at work. Christakis and colleagues [23] survey the developers at Microsoft about their usage of the tools, and report on live-site incidents to complete the survey. Lewis and colleagues [43] interview developers on two analysis tools at Google. Nguyen Quang Do [45] surveys 87 developers and analyzes the logs of static analysis tools at Software AG. Through those studies, the authors find common usability issues such as workflow integration, waiting times, bad warning explainability, and bad tool design. This is related to the findings of Imtiaz and colleagues. [37], who mined StackOverflow posts to discover that filtering and verifying false positives are major concerns of developers when using static analysis tools. Similarly, studies of static analysis tools, such as Parfait [24], focus on scalability and developer workflow and not on the user interface. While those studies report on general usability issues, we focus on tool design, and in particular, on the tool's user interface (which includes the Graphical User Interface, but also all functionalities provided to the user, e.g., generating reports).

3 Methodology

In this section we first justify our choice of tools and then describe the interfaces of those tools. Next, we describe the study environment, including the projects that each of the tools scanned. We also outline our approach toward conducting the heuristic walkthroughs and the user study. Finally, we provide a replication package.

3.1 Tools

We chose to examine four security-oriented static analysis tools, Find Security Bugs (FSB), RIPS, Flawfinder, and CTool. In this section, we justify our choice of those tools and describe their interfaces, focusing particularly on how they present information to developers.

We considered 61 candidate tools from lists of static analysis tools compiled by organizations and reserchers [64–67]. To narrow the selection of tools to use for our evaluation, we followed two criteria.

Table 1: Evaluated tool interface dimensions

Dimension	FSB	RIPS	Flaw-finder	CTool
Remediation Information[1]	✓	✓	✓	✓
Trace Navigation[2]	✓	✓		✓
Quick Fixes[3]	✓			
Graphical Rep. of Traces[4]		✓		✓
Command-Line Interface		✓	✓	
IDE Integration	✓			
Standalone Interface		✓		✓

[1] Information that helps developers fix a vulnerability;
[2] Affordances that allow developers to trace dataflow;
[3] Features for applying patches automatically;
[4] Graphical representations of dataflow traces;

The first one was *availability*. We only considered tools we could access and run. This criteria limited our selection of commercial tools, because their license agreements often explicitly forbid the publication of evaluations or comparisons. (For example, Coverity's license agreement states, "Customer will not disclose to any third party any comparison of the results of operation of Synopsys' Licensed Products with other products."[1]) Although most tool vendors we contacted were interested in the study in principle, only one agreed to participate in our study, under the condition that we anonymize the tool by not using screenshots, names, trademarks, or other distinguishing features in any publication. We thus refer to this tool as CTool (for Commercial Tool).

Second, to increase the *generalizability* of our results, we chose four tools that cover different aspects of the tool interface design space. This primarily translates to selecting tools with four different modes of interaction: command-line interface, IDE integration, and two standalone tools. Table 1 summarizes how FSB, RIPS, Flawfinder, and CTool vary along some of the interface dimensions we considered. Note that Table 1 reflects the interfaces of the versions of the tools we chose to evaluate. For instance, FSB can also be run as a command-line tool. The full table of tools and interface dimensions is available online [11].

Though definitive usage statistics are hard to find for these tools, it is fair to characterize all four of these tools as widely-used. According statistics published by Sourceforge,[2] FindBugs,[3] RIPS, and Flawfinder have approximate download counts of: 1,410,000, 143,000, and 19,000 respectively. CTool has been adopted by government agencies and hundreds of companies across different industries.

Due to the *availability* constraint, our sample only includes one commercial tool. This introduces a potential threat that

[1] https://www.synopsys.com/company/legal/software-integrity/coverity-product-license-agreement.html
[2] https://sourceforge.net/
[3] Download statistics are not available specifically for FSB, which is a plugin for FindBugs

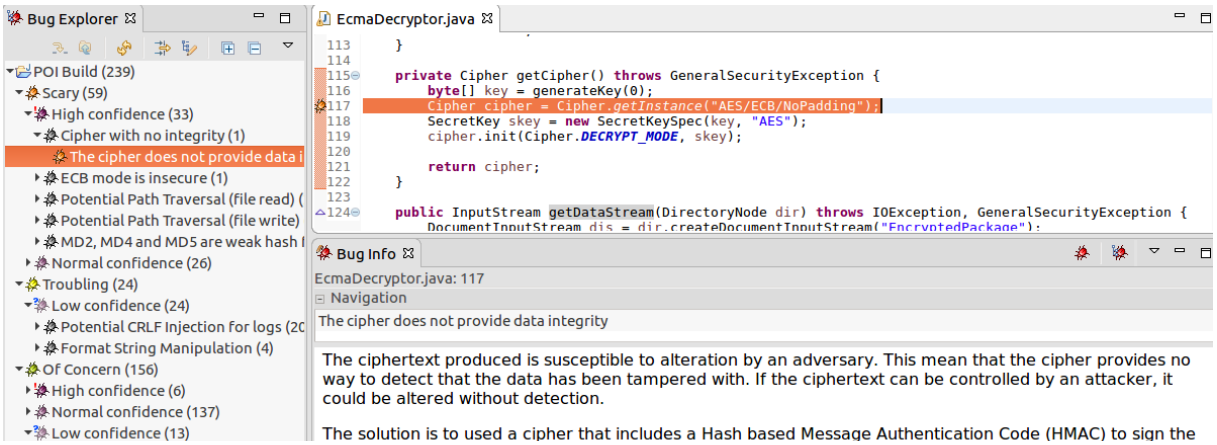


Figure 1: The Graphical User Interface of Find Security Bugs.

our sample may not represent tools used most frequently in the real-world. However, an empirical study examined the 20 most-popular Java projects on GitHub that use static analysis as part of their continuous integration pipeline [62]. They report that those projects commonly use open-source tools, like CheckStyle, FindBugs, and PMD. Reportedly, none use commercial analysis tools. Vassallo and colleagues' findings support this general trend in different contexts [59].

3.1.1 Find Security Bugs

Find Security Bugs (FSB) is an extension of the FindBugs/SpotBugs [12] static analysis tool for Java programs [13]. FSB detects 125 types of security vulnerabilities. We used the open-source Eclipse [4] plugin of the tool, version 1.7.1.

Figure 1 depicts FSB's GUI. The "Bug Explorer" pane on the left lists the potential vulnerabilities found by the most recent scan. Categories indicate the severity of the errors. In each category, errors are grouped according to FSB's certainty. Finally, vulnerabilities are grouped by type (e.g., "Cipher with no integrity"). When a user double clicks an error, the tool highlights the relevant lines of code in the editor. Tooltips for those icons provide information on all errors occurring at that particular line. In addition, the "Bug info" pane provides more information about the vulnerability that is being examined. It contains a bug description, examples of similarly vulnerable code and how to fix it, links to useful information on this vulnerability, and tool-specific information. It also provides a "Navigation" panel that contains a trace of the vulnerability.

FSB also allows users to customize how the list of results is shown in the left view. Bug patterns and categories can be toggled in and out, in which case, errors matching the category will no longer show in the list. The different types of vulnerabilities can also be reclassified in different severity categories. It is also possible to exclude particular source files from the scan, and to choose which analyses to run.

3.1.2 RIPS

RIPS [10] is a security static analysis tool for PHP code that detects more than 80 vulnerabilities. It provides a standalone web interface from which the user can configure and launch scans, and consult the results. We used version 0.55 of RIPS.

Figure 2 presents the main screen of RIPS. RIPS summarizes its results in the "Result" popup. Vulnerabilities are grouped by files and ordered by vulnerability type. RIPS provides a short description of each vulnerability alongside the problematic code. An icon on the left (not pictured) opens a non-editable version of the file containing the error. Sometimes, a "help" icon and a "generate exploit" icon are also shown on the left. When available, the help view explains vulnerabilities in more detail and sometimes suggest fixes. The generate exploit icon opens a view in which the user can generate an example exploit for this vulnerability.

The top menu of the page gives access to additional views that include a summary of the scan, a list of the program's functions, and a call graph illustrating which functions call each other. The user can rearrange the layout of the graph.

3.1.3 Flawfinder

Flawfinder [7] is a command-line tool that detects uses of dangerous functions in C/C++. We used version 2.0.4 of this open-source tool.

Figure 3 depicts Flawfinder's HTML report. The report lists all files that were scanned and all the vulnerabilities Flawfinder found, ordered by severity. For each error, Flawfinder provides the location of the error, the severity score, the vulnerability type, the dangerous function, a short description of the vulnerability, a link to the CWE page of the vulnerability, and a proposed fix—which is often the name of a safe function that can be used instead of the vulnerable one. The bottom of the report shows the analysis summary, which contains statistical data about the scan, such as with

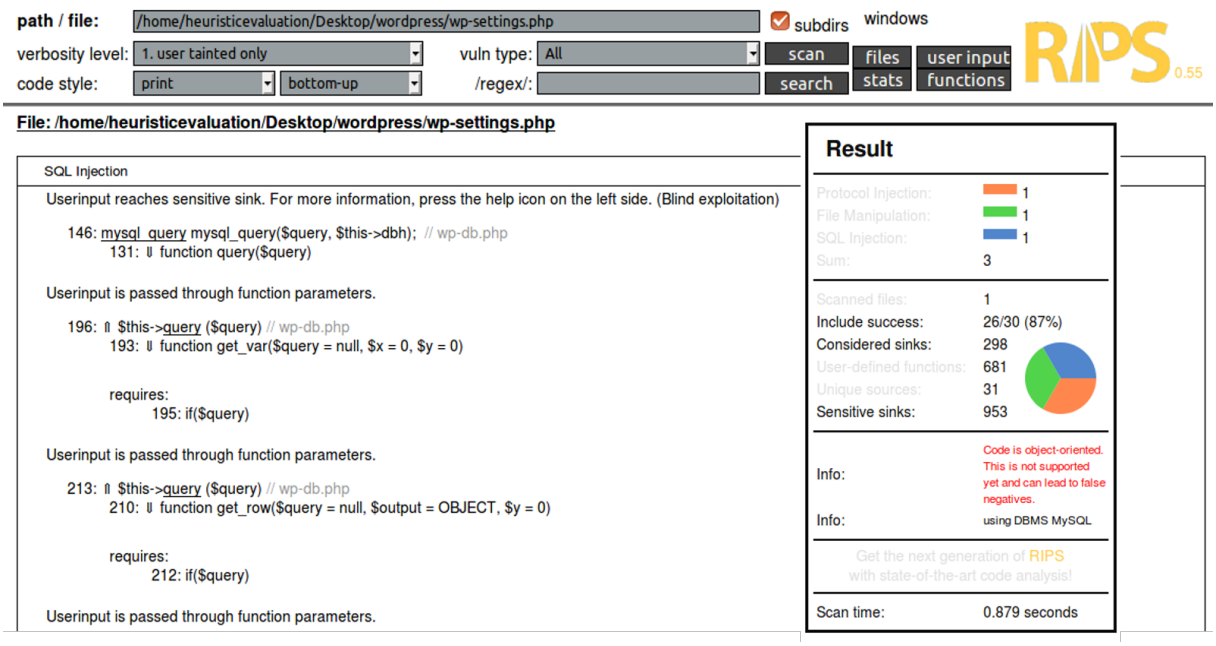


Figure 2: The Graphical User Interface of RIPS.

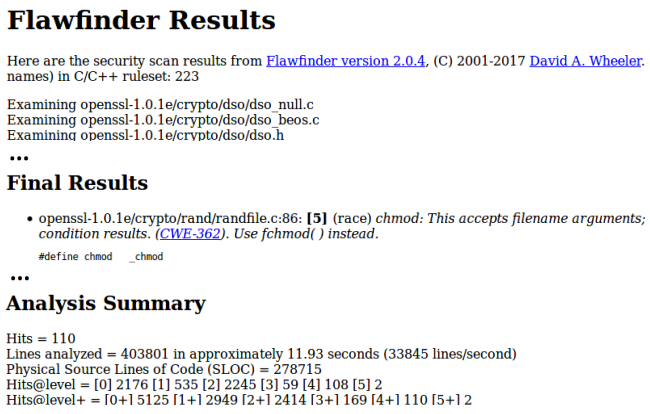


Figure 3: The Graphical User Interface of Flawfinder.

the number of files scanned, the number of errors reported, etc. The tool either prints its report in the command-line or produces reports in HTML or CSV format.

Flawfinder can be customized with command-line options to exclude files, or run on patchfiles, i.e. the diffs between two git commits. Errors can also be filtered out of the output with regex patterns.

3.1.4 CTool

CTool is a commercial tool that is largely used in industry to scan C, C++, and Java code and bytecode. It is able to detect a large range of software defects, from simple bugs to complex security vulnerabilities. The tool can be run from the command line or through a full application. It also provides different interfaces such as an IDE plugin, or a web page. The creators of CTool provided us with the default web interface, which we used to scan Apache POI version 3.9 [1].

The web interface of CTool has two main views, which we detail on a high level to keep the tool's anonymity. The first view is an overview of the warnings found in the project. Each warning is reported along with a priority score, the warning type, the code location, and information on its severity. The second view details the selected warning. It shows its details in the source code, and sometimes suggests fixes. In this view, the user can comment on the warning and manage it (e.g., edit its priority). The two main views of CTool are supported by a large number of visuals, in particular diverse charts and graphs and a complex navigation system. The GUI cannot be customized, since it is a web page, but the diversity of the visuals and the navigation capabilities cover a large number of potential use cases developers would run into.

The tool provides the ability to export a report in xml, html, or pdf format, and to customize the report. It also allows the users to customize the analysis by choosing which checkers to run, and set code annotations that guide the analysis at runtime. Users can also annotate warnings and track scores throughout the development lifecycle across different runs.

3.2 Analyzed Applications

To ensure the evaluators could exercise the tools in a variety of situations, we chose subject applications that contained multiple types of vulnerability patterns. Synthetic suites, like

the Juliet test suite [19], could have helped us ensure this coverage. However, those hand-crafted tests do not represent how vulnerabilities appear to developers in the wild.

We selected test suites from the "Applications" section of the Software Assurance Reference Dataset [9], originating from production code bases containing known vulnerabilities.

We selected three test suites: RIPS scanned WordPress version 2.0 (PHP); FSB and CTool scanned Apache POI version 3.9 (Java); and Flawfinder scanned OpenSSL version 1.0.1e (C). All three open-source tools and their associated applications were configured in a single virtual machine image for evaluation. The virtual machine is available online [6].

3.3 Heuristic Walkthroughs

We first identified usability issues by using a *heuristic walkthrough* [55], a two-phase method that combines the strengths of two usability evaluation techniques: cognitive walkthroughs [50] and heuristic evaluations [47]. In a cognitive walkthrough, evaluators simulate the tasks that real users would perform with a system. In a heuristic evaluation, evaluators systematically examine a system following a set of heuristics (as opposed to the task-driven approach in a cognitive walkthrough). In a heuristic walkthrough, evaluators first perform a cognitive walkthrough and then perform a heuristic evaluation. Combining the strengths of these two techniques in this way, heuristic walkthroughs have been shown to be more thorough than cognitive walkthroughs and more valid than heuristic evaluations on their own [55]. We chose this evaluation technique because heuristic walkthrough evaluations have successfully identified usability issues in other domains, such as electronic health records systems [28], music transcription tools [20], and lifelong learning systems [35]. Heuristic walkthroughs enable relatively few external evaluators [15] to identify usability issues by providing them with a structured process and requiring them to be double experts (experts in the application domain and usability principles). In compliance with these requirements, the two authors who conducted the heuristic walkthroughs for all four tools were familiar with security tools and usability principles. Our evaluation draws on our experience building advanced static analysis tools, conducting laboratory experiments, and graduate-level courses in security, software design, and user experience. Not including the time required to configure each tool, each evaluator spent approximately one workday with each tool. The study protocol [5] and the detailed list of usability issues [8] are available online.

3.3.1 Phase 1: Task-Oriented Evaluation

Phase 1 of a heuristic walkthrough resembles a cognitive walkthrough, where evaluators approach a system with a list of predefined tasks. The goals of this phase are for evaluators to familiarize themselves with the system and complete tasks similar to those actual users would try to complete. In Phase 1 of our study, we used the tools with a particular task in mind: fixing as many errors as possible in a limited time. To do so, we used the following guidelines:

- Choose a vulnerability to inspect first.
- Determine whether it is a true positive or a false positive.
- Propose a fix to the vulnerability.
- Assess the quality of the fix.

To help us think critically about each tool, we used Sears' list of guiding questions [55]. These questions ask evaluators to consider whether users will: know what to do next; notice the correct controls; know how to use the controls; see progress being made. During Phase 1, we recorded the vulnerabilities we inspected, our impressions of the tool, and any usability issues we encountered.

3.3.2 Phase 2: Free-Form Evaluation

Phase 2 of a heuristic walkthrough resembles a heuristic evaluation, where evaluators freely explore an entire system using a set of usability heuristics to identify issues.

Among the many sets of available heuristics, we chose to use Smith and colleagues' 17 information needs [56] as the basis for our heuristics. We made this choice because the information needs are specific to security-oriented static analysis tools, in contrast with Nielsen's ten heuristics [46] or the cognitive dimensions framework [34], which are both more generic. These 17 information needs pertain specifically to security-oriented static analysis tools and cover a range of relevant topics such as: vulnerabilities, attacks, and fixes; code and the application; individuals; and problem solving support. Prior studies have shown that domain-specific heuristics can be more effective [27, 39, 48] and these particular heuristics have been used previously to evaluate a security static analysis tool [57]. Table 4 summarizes these heuristics.

The two evaluators considered each of the 17 information needs and recorded any usability issues that related to those information needs. During this phase, the evaluators also recorded additional usability issues that did not precisely fit any of the provided heuristics.

Finally, because similar issues were identified across tools, heuristic categories, and evaluators, we performed an informal thematic analysis to group usability issues into themes and subthemes. This analysis is only intended to reduce repetition and clarify the presentation of the results. Section 4 is organized according to these themes; each subsection describes one theme.

3.4 User Study

In the second part of our evaluation, we conducted a user study on the four static analysis tools, with the goal of triangulating the observations made in the heuristic evaluation. To this end, we recruited 12 participants, who we refer to as **P01**–**P12**,

with various degrees of professional experience as software developers. Participants answered questions on a Likert scale from 1 (novice) to 5 (expert) about their experience. Table 3 reports on their responses. In summary, participants self-reported familiarity with software security (median 3/5), Java (median 4/5), C++ (median 2/5), and PHP (median 1/5).

We presented each participant with two of the four static analysis tools, and asked them to fix warnings reported by the tools, using the same code bases as for the heuristic evaluation. Participants thought aloud while working with each tool for approximately 20 minutes. Following each tool, we conducted semi-structured post-task interviews. At the end of the session we collected demographic data. During the study, we allowed participants to ignore warnings they were uncomfortable with. This choice helped to account for differences in programming language skill. It also simulated real-world developers' strong propensity to selectively ignore most warnings [53, 61]. For all tools, all participants managed to find warnings they were comfortable with. In the post-task interviews, participants described their experience with the tools, focusing on how the tool helped them understand and fix warnings. Appendix C lists the questions we used to guide this discussion.

To reduce fatigue, we only asked participants to interact with two tools. Still, the mean session duration was 52 minutes, 58 seconds. To avoid learning effects between the two tools, we applied a latin-square design [26]. We distributed the tools evenly between participants—each tool was evaluated by six participants.

3.4.1 Data Extraction

We captured screen, audio, and questionnaire responses. Afterwards, we asked two independent researchers to review the audio recordings and extract the usability issues encountered by the participants. This yielded a total of 562 individual usability incidents. We kept the intersection of both reviewers' reported incidents, thus reducing the number of total usability incidents to 140. Two authors then classified the incidents into distinct usability categories using the open card sort methodology [36]. The classification yielded a Cohen Kappa of $\kappa = 0.93$, indicating an almost perfect agreement [41]. Afterwards, the two raters discussed and agreed on a final classification, which we present in Section 4.

3.5 Replication Package

To support the replication and continuation of this work, we have made our materials available, including the virtual machine image used during our heuristic evaluation. It contains the static analysis tools (we exclude CTool for legal reasons) and the code bases they were used on [6], the study protocol [5] and the list of usability issues we detail in the following section [8].

The user study protocols and heuristic walkthrough guide are also available in the Appendix.

Because usability evaluations of security tools are beneficial beyond the scope of what was feasible during our study, we also provide the heuristic evaluation guide we developed. With this guide, a qualified evaluator with expertise in static analysis tools and usability principles could extend our work to any additional static analysis tool for security.

4 Results

Through our heuristic walkthrough evaluations and user study, we identified 194 and 140 usability issues, respectively. We do not intend for the presence or quantity of these issues to be a statement about the overall quality of the tools we evaluated. Instead, each of these issues represents a potential opportunity for tools to be improved. For completeness, we provide the full list of usability issues in the supplemental materials [8].

In each section, we will give a general description of the usability issues relating to that theme, explain how instances of those issues impact developers, and sketch how our insights could be used by tool designers and researchers to improve security-oriented static analysis tools. Next to the title of each theme, we report the number of usability issues in parenthesis (X) that we identified during the heuristic walkthrough phase. This number simply characterizes our findings and should not be interpreted as the ranking or severity of the issues in that theme. Also note that these counts sum to slightly more than 194, because some usability issues span multiple themes.

To further organize the results, we have bolded short titles that describe subthemes of issues within each theme. Next to each subtheme title are the tools, in {braces}, that issues in that subtheme apply to. For instance, "**Immutable Code {RIPS, Flawfinder, CTool}**" denotes that Immutable Code usability issues apply to RIPS, Flawfinder, and CTool, but not FSB. In addition, Table 2 provides an overview of the themes and subthemes.

4.1 Missing Affordances (39)

Beyond presenting static information about code defects, analysis tools include affordances for performing actions, such as navigating code, organizing results, and applying fixes. Issues in this category arose when tools failed to provide affordances.

Managing Vulnerabilities {FSB, RIPS, Flawfinder}: After scanning the source code, tools must report the identified vulnerabilities to developers. We found that FSB, RIPS, and Flawfinder did not provide adequate affordances for helping developers navigate and manage the list of reported vulnerabilities. Managing the list of reported vulnerabilities is important, because it allows developers to quickly find the vulnerabilities they would like to inspect and fix. For instance, some developers might only be interested in a subset of the

Table 2: Usability issues, grouped by theme

Theme	Subtheme
4.1 Missing Affordances	Managing Vulnerabilities Applying Fixes
4.2 Missing or Buried Information	Vulnerability Prioritization Fix Information
4.3 Scalability of Interface	Vulnerability Sorting Overlapping Vulnerabilities Scalable Visualizations
4.4 Inaccuracy of Analysis	
4.5 Code Disconnect	Mismatched Examples Immutable Code
4.6 Workflow Continuity	Tracking Progress Batch Processing

code, or have expertise fixing particular types of vulnerabilities. These three tools simply show a list of all vulnerabilities they found; the options to manage the list of all potential vulnerabilities were limited.

Flawfinder, for example can generate a single text file, csv, or HTML page containing all the scan results. As Figure 3 depicts, in Flawfinder these "Final Results" are presented as a list that cannot be reorganized or sorted. To find a vulnerability to inspect in detail, a developer must linearly search through the full list of results. Consequently, it is difficult for developers to quickly find and fix vulnerabilities they are interested in.

User Study: Most issues in this category impacted Flawfinder, where participants found the presentation of vulnerabilities (3) "irritating" (**P09**) and "poor" (**P11**). **P01** and **P12** were also confused by RIPS's presentation of vulnerabilities—**P12** suggested that it would be easier to make sense of the error messages if they were presented in a table.

Applying Fixes {FSB, RIPS, Flawfinder, CTool}: The tools we evaluated did not fully support *quick-fixes* (semi-automated code patches for common vulnerabilities) and did not otherwise offer assistance applying changes. Instead, developers must manually fix the issues reported by these tools. Only FSB included some quick-fixes, but this feature was available for just three out of the 21 defect patterns present in our test suite. Without these affordances for applying fixes, developers must exert extra effort to resolve the defects presented by their tools.

User Study: We did not identify any participants who faced these issues during our user study. One explanation for this may be that issues in this category only become apparent after working with a tool for an extended period of time, trying to apply several fixes.

Discussion: Many of the affordances that we noted as missing from these tools do not represent revolutionary breakthroughs in user interface design. In fact, features like sorting and filtering lists are commonplace in many applications. Integrating these well-known affordances into static analysis tools for security could be one low-cost way to improve the usability of these tools. On the other hand, some affordances will require more effort to incorporate into analysis tools. For example, affording developers the ability to accurately apply automated patches remains an open research area. We are encouraged by systems like FixBugs [18], which assists developers in applying quick-fixes with FindBugs. Our results suggest that security-oriented static analysis tools would benefit from advances in this area.

4.2 Missing or Buried Information (96)

Static analysis tools can provide developers with a wide range of information about the defects they detect. For example, all four tools we studied give information about the location and defect-type of the vulnerabilities detected. The issues in this theme correspond to instances where tools failed to provide information that would be used to resolve defects. In this theme we discuss both missing information and buried information. These two issues are intertwined, because buried information that a developer never unearths is effectively missing.

Vulnerability Prioritization {FSB, RIPS, Flawfinder, CTool}: Since tools can generate many alerts from a single scan, before fixing a vulnerability, developers must decide which alert to inspect first. To varying extents, all four tools failed to provide information that, had the information been present, would have helped developers decide which vulnerabilities to inspect. Many of these issues arose as we considered the "Vulnerability Severity and Rank" heuristic during Phase 2. We noted several different types of missing information, such as information about: which files contained clusters of vulnerabilities (Flawfinder); a vulnerability's severity (RIPS); and how to interpret severity scales (FSB, Flawfinder CTool). For example, unlike RIPS, FSB provides information about the severity of each vulnerability, typically in the following form:

Rank: Of Concern (18), **confidence**: Normal

However, even FSB does not provide information about how to interpret this report. A developer might be left wondering whether 18 is high or low, or what other confidence values are possible. This issue may disproportionately affect users who are using a tool for the first time and still learning to interpret the scales. Nonetheless, lacking information about how to prioritize vulnerabilities, developers might misallocate their limited time by fixing low-severity vulnerabilities.

User Study: Participants in the user study encountered similar issues to those we identified. **P04** complained that RIPS did not provide any severity or priority scores. Further, several participants (**P01**, **P04**, **P05**, **P08**, and **P10**) were confused by FSB and CTool's scales.

Fix Information {FSB, RIPS, Flawfinder, CTool}: The tools we evaluated also failed to provide some information that developers would need to accurately fix vulnerabilities. The types of missing information spanned many different categories. To name a few, the tools were missing code examples, fix suggestions, definitions of unfamiliar terms, and explanations of how vulnerabilities could be exploited. Furthermore, some types of information that were present were not detailed enough, such as when the tools provided terse issue descriptions or when the tools listed possible fixes, but did not articulate the tradeoffs between those solutions.

User Study: Nearly all participants experienced issues as a result of insufficient fix information across all four tools. For instance, participants described the information that was provided as, "not helpful and too complicated" (**P05**), "generic" (**P06**), "unclear and very irritating" (**P09**), and "short" (**P11**).

Discussion: One solution to these types of issues would be to simply add more information to tool notifications. This simple solution would ensure all the information needed to select and fix vulnerabilities is present for the developer. However, overstuffing notifications with too much information might bury the most pertinent information at a given time. Instead, the challenge for static analysis tools is to discern when developers need a particular piece of information and deliver that information.

4.3 Scalability of Interface (11)

As static analysis tools scale to find more defects in larger codebases, so too must their interfaces for presenting those defects. The issues in this section arose when tools struggled to present large amounts of information about vulnerabilities. Here we distinguish between scalable interfaces and scalable tools because we are interested in the usability of these tools' interfaces, not their technical capabilities, which have already been explored elsewhere [17, 51, 63]. Each of the four tools we examined exhibited an interface scalability issue.

Vulnerability Sorting {Flawfinder}: As we previously discussed in Section 4.1, Flawfinder does not provide affordances for managing the list of vulnerabilities it detects. This issue is magnified as Flawfinder scales to identify more vulnerabilities in a project. Lacking the ability to manage this list, developers must resort to sub-optimal task selection strategies, such as searching linearly through the list for a vulnerability they would like to inspect.

User Study: Unsurprisingly, participants requested the ability to "sort the warnings by type, class, and significance."—**P11**

Overlapping Vulnerabilities {FSB, CTool}: Like the other tools we evaluated, FSB and CTool can detect multiple different patterns of vulnerabilities. When multiple vulnerability patterns are detected on the same line, these tools do not provide clear indications that multiple problems are occurring in the same location. FSB, for example draws multiple bug icons directly on top of each other, which appears just the same as a single bug icon. In fact, Line 117 in Figure 1 contains multiple overlapping vulnerabilities, however this is not perceptible without hovering over the bug icon (Figure 4). CTool includes a feature for displaying overlapping vulnerabilities, but this feature is turned off by default and is located in a somewhat hidden location.

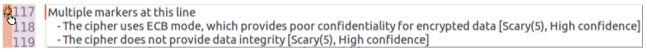


Figure 4: Instance of **Overlapping Vulnerabilities**

Scaling up the number of defect detectors increases the likelihood that these instances of overlap will occur. The decision to display vulnerability markers in this way deprives developers of information and forces FSB users to manually inspect each line for duplicates. As a consequence, developers might overlook severe vulnerabilities if their indicators are hidden beneath minor vulnerabilities.

User Study: Participants did not explicitly mention this issue during the user study, either because they did not inspect any overlapping notifications, or because they didn't notice that they were overlapping.

Scalable Visualizations {RIPS}: Finally, the clearest scalability issue we encountered was the call graph visualization in RIPS. An example of this visualization is depicted in Figure 5. The graph is intended to show the flow of tainted data from sources to sensitive sinks. However, when functions along the flow are called from many sites, all the call sites are added to the graph, resulting in a crowded visualization. Furthermore, when these call sites span more than 50 files, RIPS will not generate any visualization (Figure 6).

User Study: Only one participant (**P04**) experienced this issue with RIPS during the study, suggesting that the graph representation, "could be improved."

Discussion: We propose two potential design changes that would improve FSB and CTool's overlapping vulnerabilities issues. One possibility is that these tools could use the stack metaphor to present the co-located vulnerabilities. Rather than draw the icons directly on top of each other, the tool could offset each subsequent bug by a few pixels. Alternatively, the tools could annotate a single bug icon with the number of vulnerabilities present on that line (e.g., ).

RIPS provides a call graph visualization, whereas the two open-source tools we evaluated provided no similar feature. Such a feature could help visually oriented developers trace vulnerabilities and reason about the flow of tainted data through their system. However, if tools are to implement successful graph visualizations to support developers, the scalability of the visualization must be considered. Tool designers could consider implementing features such as those in Reacher [42], which enable developers to expand and highlight only the relevant paths through the graph.

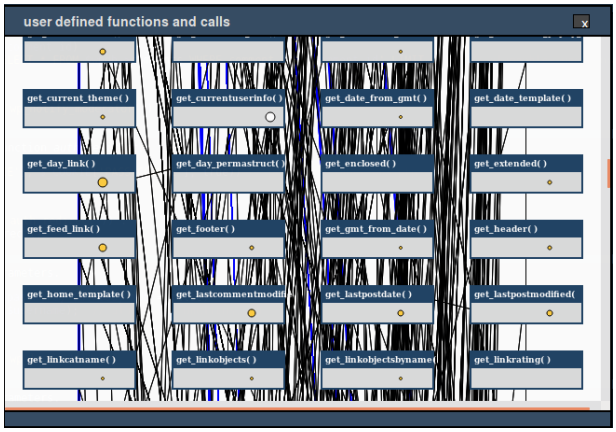


Figure 5: Scalability of RIPS' function view.

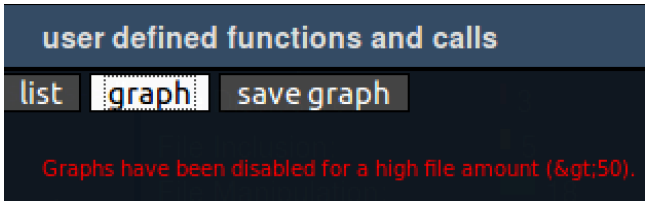


Figure 6: RIPS call graph visualization for more than 50 files.

4.4 Inaccuracy of Analysis (17)

The issues in this category arose when we encountered implementation bugs or unexpected behaviors. These issues do not necessarily represent deep design flaws of FSB, RIPS, Flawfinder, and CTool. However, these bugs do have an impact on usability, because implementation bugs may affect developers' confidence in tools and their abilities to complete tasks. We encountered several issues in this category spanning all four tools; to illustrate the types of issues in this category, here we describe in detail one of these issues affecting FSB.

One set of issues in this category affected FSB, specifically its code navigation features. For instance, when we used FSB's quick-fix feature, the IDE unexpectedly navigated away from the current file. This behavior was disorienting and could easily cause a developer to lose track of the code they were trying to fix. We also observed issues with FSB's navigation pane in the bug info window. This pane often either contained duplicated entries, was missing entries, or contained entries that, when clicked, didn't navigate the user anywhere. Figure 7 depicts an instance of the duplicated entries issue—both entries labeled "Sink method java/io/File.<init>(Ljava/lang/String;)V" refer to the same location in the code.

User Study: While using RIPS, participants (**P01**, **P06**, **P12**) encountered a similar unexpected behavior. RIPS summarizes results in a popup window (Figure 2). Frustratingly, this popup window obscures other information about the results and cannot be closed.

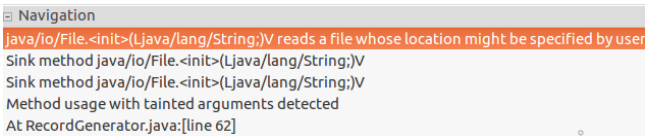


Figure 7: Duplicate entries in FSB's navigation feature

4.5 Code Disconnect (14)

Static analysis tools generate reports based on the code they scan. However, we identified usability issues when the content of those reports were disconnected from the source code.

Mismatched Examples {FSB, RIPS, Flawfinder, CTool}: The first issue in this category relates to the code examples used by all four tools. Many FSB notifications, for instance, contain hard-coded examples of vulnerable code and also examples of suggested code patches. Providing any code example is certainly more helpful than giving no information. Nonetheless, because the examples are hard-coded for each pattern, the burden of figuring out how to adapt and apply that example to the current context falls on the developer. Even if the example is written in the same programming language, this translation can be non-trivial, especially if the example is drawn from a source using different libraries or frameworks. Figure 8 depicts one instance where FSB's examples are mismatched with the vulnerable code. In this case, FSB's "solution" example (Figure 8b) differs substantially from the original problematic code: the variable names are different; the ciphers are initialized in different modes (encrypt vs. decrypt); and the ciphers are using different encryption algorithms (ECB vs. GCM).

```
115 private Cipher getCipher() throws GeneralSecurityException{
116   byte[] key = generateKey(0);
117   Cipher cipher = Cipher.getInstance("AES/ECB/NoPadding");
118   SecretKey skey = new SecretKeySpec(key, "AES");
119   cipher.init(Cipher.DECRYPT_MODE, skey);
120   return cipher;
```

(a) Vulnerable code

Solution:

```
Cipher c = Cipher.getInstance("AES/GCM/NoPadding");
c.init(Cipher.ENCRYPT_MODE, k, iv);
Byte[] cipherText = c.doFinal(plainText);
```

(b) FSB's example "solution"

Figure 8: Instance of **Mismatched Examples**

User Study: Several participants encountered similar issues with mismatched examples while using RIPS (**P04**, **P06**, **P10**, **P12**). Figure 9 illustrates a common issue. Here RIPS suggests that functions-compat.php contains an error on line 304, however, this file is only 155 lines long. The mismatched line numbers confused participants: "I am confused because the line number doesn't correspond to line in file."—**P10**

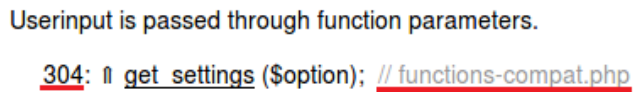


Figure 9: RIPS line numbers mismatched with code.

Immutable Code {RIPS, Flawfinder, CTool}: We encountered usability issues while trying to apply changes. RIPS, Flawfinder, and CTool's web UI do not allow developers to directly modify the code they scan. These three tools display a projects' source code (CTool) or snippets (RIPS, Flawfinder) containing potential vulnerabilities, but do not enable developers to directly make those changes. Instead, developers must view scan results in one window and edit the code using a separate text editor. This disconnect forces developers to transfer their findings from where the tool presents its results to where they need to make changes. Furthermore, without being able to access the complete original source code while using a static analysis tool, as is the case with RIPS and Flawfinder, developers cannot use navigation tools to browse code related to a vulnerability. This workflow is problematic, because developers are burdened with manually maintaining a mapping between the tool's findings and the code editor. For example, developers must mentally track details about the vulnerability—like which line, module, and version it is contained in—so that they can fix the appropriate line.

User Study: Several participants in the user study verified this issue. They described switching between a code editor and the static analysis tools as, "disorienting" (**P10**), "cumbersome" (**P03**), and "time-consuming" (**P09**).

Discussion: Presenting results within an IDE, like FSB and other versions of CTool do, helps developers maintain connections between a vulnerability and the code surrounding that vulnerability. Developers also have the added benefit of being able to use code navigation tools while inspecting a vulnerability within this environment. However, our findings reveal opportunities to establish deeper connections between vulnerable code and how tools present those vulnerabilities to developers. Considering the mismatched examples usability issue, we imagine that analysis tools' code examples could be parameterized to use similar variable names and methods to those in the scanned code. Achieving a closer alignment between source code and examples will hopefully reduce developers' *cognitive load* while translating between the two, freeing them up to concentrate on resolving the vulnerability.

4.6 Workflow Continuity (29)

Developers do not use static analysis tools in isolation. Tools must synergize with other tasks in a developer's workflow, such as editing code, testing changes, and reviewing patches. These issues arose when tools dictated workflows that were not compatible with a developer's natural way of working.

Tracking Progress {FSB, RIPS, Flawfinder}: Developers work with static analysis to reduce the number of vulnerabilities in their code, making code changes to progressively resolve warnings. However, some tools present their results in such a way that does not allow developers to track their progress. Instead of reporting which vulnerabilities were added and removed between scans, tools only provide snapshots of all the current vulnerabilities in a project at a given time. This is only somewhat problematic when a developer wants to consider the effects of their changes on a single vulnerability notification. For example, if the tool at first reports 450 vulnerabilities, and then reports 449 after the developer applies a patch, then they can assume their patch fixed the one vulnerability. However, when a developer or his/her team makes sweeping changes to address many several vulnerabilities simultaneously, it becomes much more difficult to determine which issues were added and removed between scans based only on two snapshots.

User Study: **P01** encountered this issue while using FSB; they were unable to verify that their change fixed the bug, "How do you check whether it has been fixed?" **P12** echoed **P01**'s frustration while using Flawfinder, "It would be nice if the user interface could track which errors have been added and removed. Now, I need to do all this work manually."

Batch Processing {FSB, RIPS, Flawfinder, CTool}: Secondly, all four tools we evaluated dictate that developers address notifications individually. This is problematic because projects can contain many occurrences of the same vulnerability patterns. For instance, CTool detected 30 occurrences of a package protection vulnerability in POI. Serially fixing these vulnerabilities is error-prone—developers must consistently apply the right fix to each occurrence. Since tools are technically capable of scanning many more files than developers could manually scan, they must also enable developers to fix similar vulnerabilities in batches. Otherwise, the tool far outpaces the developers, adding vulnerabilities to their work queues much faster than they dismiss them.

User Study: **P09** described this issue during the user study, "If the bugs are similar, you don't have option to fix them at the same time... I would like to fix them all at once."

Discussion: Static analysis tools for security can be improved to better support developers' workflows. By keeping track of the vulnerabilities added and removed in each scan, tools provide developers with scan diffs. These would help developers identify changes that add many potential vulnerabilities to the code as well as draw developers' attention to changes that remove vulnerabilities.

Tools could also support developers' workflows by enabling them to process similar vulnerabilities in batches. One way tools could accomplish this is by integrating with automated refactoring tools. Rather than fixing the individual occurrences of an issue one at a time, developers could describe to the static analysis tool how they would like to fix the

issue, then the static analysis tool would apply a refactoring to fix all occurrences of the issue.

5 Design Guidelines

To help make our findings more actionable for toolsmiths, we provide a set of design guidelines:

- **Communicate what and how to fix.** Just locating potential vulnerabilities is insufficient. Static analysis tools should enable developers to diagnose (and, when it's technically feasible, automatically fix) the vulnerabilities they detect. Tools should either provide semi-automated quick-fixes or provide enough information for developers to manually fix detected vulnerabilities. We are inspired by FixBugs [18], which shows how human-in-the-loop fixes can improve automation beyond what's possible with quick-fixes. For even more complex problems, better explanations would be preferable. (Section 4.2)
- **Situate alerts within editable code.** Switching between a code editor and a tool's results page is disorienting and time-consuming. Tools should present their alerts where developers can directly modify problematic code. (Section 4.5)
- **Integrate with existing workflows.** Tools should integrate with developers' workflows. After developers fix potential vulnerabilities, tools should clearly communicate which problems have been fixed and which new problems have been introduced. (Section 4.6)
- **Generate contextualized notifications.** Tools should infer contextual information about the code they scan and use that contextual information to generate alerts and examples that match the current context. (Section 4.5) For instance, FSB's example in Figure 8 could be contextualized by parsing the vulnerable code and generating an alert with matching variable names and cyphers.

6 Limitations

First, we only evaluated four tools. To mitigate this threat to generalizability, we selected analysis tools that are each representative of distinct interface features (Table 1) and we included both a commercial tool and open-source tools. We do not claim that all of the usability issues we identified necessarily generalize to other analysis tools. For example, not all tools use ambiguous vulnerability severity scales like FSB does (Section 4.2). However, the themes and subthemes we report (Table 2) are more generalizable—all but two subthemes describe usability issues that span multiple tools. We expect that these *types* of issues detract from the usability of other tools. To further mitigate this threat, we also made our evaluation guide available. If researchers or toolsmiths are interested in understanding how these categories manifest in a particular tool, our approach can easily be applied to additional tools.

We also acknowledge that the evaluators' individual tool usage styles might have influenced the issues they identified. To bolster the ecological validity of our study, we selected real-world applications and instructed evaluators to perform realistic tasks. Additionally, we triangulate our findings through a user study with professional developers.

Another limitation of our study is that we examined datasets containing known vulnerabilities (Section 3.2). This choice guaranteed that scans would yield results, thus enabling us to examine how each tool presented vulnerabilities. However, static analysis users may face additional usability issues when searching for unknown vulnerability patterns.

Our choice of usability evaluation technique also influences our results. Compared with user studies, where qualified participants might spend less than an hour using a tool, evaluators conducting heuristic walkthroughs have more time to find deeper issues. On the other hand, our choice of heuristics might have influenced the usability issues we identified. We chose to use these heuristics because prior studies have shown that domain-specific heuristics can be more effective [27, 39, 48]. To mitigate this issue, we performed a user study with independent participants and independent reviewers, resulting in a classification that is different from the heuristics we chose. The results of the user study confirm most of our observations from the heuristic walkthrough. The user study was performed in a closed setting, within time limits. While it would be interesting to conduct a complementary study over several months in an industry setting, we note that the heuristic walkthrough already mitigates this limitation by allowing evaluators to spend more time and investigate the tools in depth. To avoid further external threats to validity, we recruited participants with an industry background.

7 Conclusion

This paper responds to a call for usability evaluations of *developer* security tools. In this work, we conducted a heuristic walkthrough and a user study to evaluate the usability of four static analysis tools: Find Security Bugs, RIPS, Flawfinder, and CTool. To enable similar evaluations to be conducted even more broadly, we have made our heuristics, study protocols and study setting available in a replication package.

This work reveals usability issues that detract from all four of the tools we examined. We discuss potential mitigations for those types of issues. To help toolsmiths design more usable static analysis tools for security, we present a set of design recommendations. We hope that our work enables practitioners to improve the usability of their tools and inspires researchers to evaluate the usability of developers' security tools.

Acknowledgments

We would like to thank our study participants for their time. We would also like to thank Maria Riaz, Caitlin Sadowski, Sarah Heckman, William Enck, Amy Ko, Adam Meade, and Kathryn Stolee for their feedback.

References

[1] Apache poi source code. https://poi.apache.org/subversion.html, 2018.

[2] Checkmarx home page. https://www.checkmarx.com/, 2018.

[3] Codesonar home page. https://www.grammatech.com/products/codesonar, 2018.

[4] Eclipse home page. http://www.eclipse.org/, 2018.

[5] Evaluation guide. https://figshare.com/s/087f103905189f1a7ca0, 2018.

[6] Evaluation vm. https://figshare.com/s/bf91e24a3df3c4fff77c, 2018.

[7] Flawfinder home page. https://www.dwheeler.com/flawfinder/, 2018.

[8] List of issues. https://figshare.com/s/71d97832ae3b04e0ff1a, 2018.

[9] Nist test suites. https://samate.nist.gov/SRD/testsuite.php, 2018.

[10] Rips home page. http://rips-scanner.sourceforge.net/, 2018.

[11] Security tool interface dimensions. https://figshare.com/s/5255acbe659d3097d8a2, 2018.

[12] Spotbugs home page. https://spotbugs.github.io/, 2018.

[13] Find security bugs home page. http://find-sec-bugs.github.io/, 2019.

[14] Yasemin Acar, Sascha Fahl, and Michelle L. Mazurek. You are not your developer, either: A research agenda for usable security and privacy research beyond end users. In *Cybersecurity Development (SecDev), IEEE*, pages 3–8. IEEE, 2016.

[15] Yasemin Acar, Christian Stransky, Dominik Wermke, Michelle L. Mazurek, and Sascha Fahl. Security developer studies with github users: Exploring a convenience sample. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 81–95, Santa Clara, CA, 2017. USENIX Association.

[16] Hala Assal, Sonia Chiasson, and Robert Biddle. Cesar: Visual representation of source code vulnerabilities. In *Visualization for Cyber Security (VizSec), 2016 IEEE Symposium on*, pages 1–8. IEEE, 2016.

[17] Andrew Austin, Casper Holmgreen, and Laurie Williams. A comparison of the efficiency and effectiveness of vulnerability discovery techniques. *Information and Software Technology*, 55(7):1279–1288, 2013.

[18] T. Barik, Y. Song, B. Johnson, and E. Murphy-Hill. From quick fixes to slow fixes: Reimagining static analysis resolutions to enable design space exploration. In *2016 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 211–221, Oct 2016.

[19] T. Boland and P. E. Black. Juliet 1.1 c/c++ and java test suite. *Computer*, 45(10):88–90, Oct 2012.

[20] Manuel Burghardt and Sebastian Spanner. Allegro: User-centered design of a tool for the crowdsourced transcription of handwritten music scores. In *Proceedings of the 2Nd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2017, pages 15–20, New York, NY, USA, 2017. ACM.

[21] Brian Chess and Gary McGraw. Static analysis for security. *IEEE Security and Privacy*, 2(6):76–79, November 2004.

[22] Maria Christakis and Christian Bird. What developers want and need from program analysis: An empirical study. In *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering*, ASE 2016, pages 332–343, New York, NY, USA, 2016. ACM.

[23] Maria Christakis and Christian Bird. What developers want and need from program analysis: An empirical study. In *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering*, ASE 2016, pages 332–343, New York, NY, USA, 2016. ACM.

[24] Cristina Cifuentes, Nathan Keynes, Lian Li, Nathan Hawes, Manuel Valdiviezo, Andrew Browne, Jacob Zimmermann, Andrew Craik, Douglas Teoh, and Christian Hoermann. Static deep error checking in large system applications using parfait. In *Proceedings of the 19th ACM SIGSOFT Symposium and the 13th European Conference on Foundations of Software Engineering*, ESEC/FSE '11, pages 432–435, New York, NY, USA, 2011. ACM.

[25] Jeremy Clark, Paul C. Van Oorschot, and Carlisle Adams. Usability of anonymous web browsing: an

examination of tor interfaces and deployability. In *Proceedings of the 3rd symposium on Usable privacy and security*, pages 41–51. ACM, 2007.

[26] Yadolah Dodge. *Latin Square Designs*, pages 297–297. Springer New York, New York, NY, 2008.

[27] Dean Julian Dykstra. *A Comparison of Heuristic Evaluation and Usability Testing: The Efficacy of a Domain-Specific Heuristic Checklist*. A & M University, Texas, 1993.

[28] Paula J. Edwards, Kevin P. Moloney, Julie A. Jacko, and François Sainfort. Evaluating usability of a commercial electronic health record: A case study. *International Journal of Human-Computer Studies*, 66(10):718 – 728, 2008.

[29] Pär Emanuelsson and Ulf Nilsson. A comparative study of industrial static analysis tools. *Electron. Notes Theor. Comput. Sci.*, 217:5–21, July 2008.

[30] Kevin Gallagher, Sameer Patil, and Nasir Memon. New me: Understanding expert and non-expert perceptions and usage of the tor anonymity network. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 385–398, Santa Clara, CA, 2017. USENIX Association.

[31] Nathaniel S Good and Aaron Krekelberg. Usability and privacy: a study of kazaa p2p file-sharing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 137–144. ACM, 2003.

[32] Peter Leo Gorski, Yasemin Acar, Luigi Lo Iacono, and Sascha Fahl. Listen to developers! a participatory design study on security warnings for cryptographic apis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[33] Matthew Green and Matthew Smith. Developers are not the enemy!: The need for usable security apis. *IEEE Security & Privacy*, 14(5):40–46, 2016.

[34] Thomas RG Green. Cognitive dimensions of notations. *People and computers V*, pages 443–460, 1989.

[35] Xiaoqing Gu, Fengjia Gu, and James M Laffey. Designing a mobile system for lifelong learning on the move. *Journal of Computer Assisted Learning*, 27(3):204–215, 2011.

[36] William Hudson. The encyclopedia of human-computer interaction, 22. card sorting. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/card-sorting, 2014.

[37] N. Imtiaz, A. Rahman, E. Farhana, and L. Williams. Challenges with responding to static analysis tool alerts. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, pages 245–249, May 2019.

[38] Nasif Imtiaz, Brendan Murphy, and Laurie Williams. How do developers act on static analysis alerts? an empirical study of coverity usage. In *2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE)*, pages 323–333. IEEE, 2019.

[39] Pooya Jaferian, Kirstie Hawkey, Andreas Sotirakopoulos, Maria Velez-Rojas, and Konstantin Beznosov. Heuristics for evaluating it security management tools. *Human Computer Interaction*, 29(4):311–350, 2014.

[40] Brittany Johnson, Yoonki Song, Emerson Murphy-Hill, and Robert Bowdidge. Why don't software developers use static analysis tools to find bugs? In *Proceedings of the 2013 International Conference on Software Engineering*, ICSE '13, pages 672–681, Piscataway, NJ, USA, 2013. IEEE Press.

[41] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. page 159–174, 1977.

[42] Thomas D. Latoza and Brad A. Myers. Visualizing call graphs. In *in VL/HCC'2011: IEEE Symposium on Visual Languages and Human-Centric Computing*, pages 18–22.

[43] C. Lewis, Z. Lin, C. Sadowski, X. Zhu, R. Ou, and E. J. Whitehead. Does bug prediction support human developers? findings from a google case study. In *2013 35th International Conference on Software Engineering (ICSE)*, pages 372–381, May 2013.

[44] Duc Cuong Nguyen, Dominik Wermke, Yasemin Acar, Michael Backes, Charles Weir, and Sascha Fahl. A stitch in time: Supporting android developers in writing secure code. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1065–1077. ACM, 2017.

[45] Lisa Nguyen Quang Do. *User-Centered Tool Design for Data-Flow Analysis*. PhD thesis, Paderborn University, 2019.

[46] Jakob Nielsen. 10 usability heuristics for user interface design. *Nielsen Norman Group*, 1(1), 1995.

[47] Jakob Nielsen and Rolf Molich. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '90, pages 249–256, New York, NY, USA, 1990. ACM.

[48] Tulsidas Patil, Ganesh Bhutkar, and Noshir Tarapore. Usability evaluation using specialized heuristics with qualitative indicators for intrusion detection system. In *Advances in Computing and Information Technology*, pages 317–328. Springer, 2012.

[49] Nikhil Patnaik, Joseph Hallett, and Awais Rashid. Usability smells: An analysis of developers' struggle with crypto libraries. In *Proceedings of the Fifteenth USENIX Conference on Usable Privacy and Security*, SOUPS'19, page 245–257, USA, 2019. USENIX Association.

[50] Peter G. Polson, Clayton Lewis, John Rieman, and Cathleen Wharton. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies*, 36(5):741 – 773, 1992.

[51] Bradley Reaves, Jasmine Bowers, Sigmund Albert Gorski III, Olabode Anise, Rahul Bobhate, Raymond Cho, Hiranava Das, Sharique Hussain, Hamza Karachiwala, Nolen Scaife, et al. * droid: Assessment and evaluation of android application analysis tools. *ACM Computing Surveys (CSUR)*, 49(3):55, 2016.

[52] Joshua Reynolds, Trevor Smith, Ken Reese, Luke Dickinson, Scott Ruoti, and Kent Seamons. A tale of two studies: The best and worst of yubikey usability. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 872–888. IEEE, 2018.

[53] Caitlin Sadowski, Edward Aftandilian, Alex Eagle, Liam Miller-Cushon, and Ciera Jaspan. Lessons from building static analysis tools at google. *Communications of the ACM*, 61(4):58–66, 2018.

[54] Caitlin Sadowski, Jeffrey van Gogh, Ciera Jaspan, Emma Söderberg, and Collin Winter. Tricorder: Building a program analysis ecosystem. In *Proceedings of the 37th International Conference on Software Engineering - Volume 1*, ICSE '15, pages 598–608, Piscataway, NJ, USA, 2015. IEEE Press.

[55] Andrew Sears. Heuristic walkthroughs: Finding the problems without the noise. *International Journal of Human-Computer Interaction*, 9(3):213–234, 1997.

[56] Justin Smith, Brittany Johnson, Emerson Murphy-Hill, Bill Chu, and Heather Richter Lipford. Questions developers ask while diagnosing potential security vulnerabilities with static analysis. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*, ESEC/FSE 2015, pages 248–259, New York, NY, USA, 2015. ACM.

[57] Tyler Thomas, Heather Lipford, Bill Chu, Justin Smith, and Emerson R Murphy-Hill. What questions remain? an examination of how developers understand an interactive static analysis tool. In *WSIW@ SOUPS*, 2016.

[58] Daniela G. tom Markotten. User-centered security engineering. In *Proceedings of the 4th EurOpen/USENIX Conference–NordU2002*, 2002.

[59] Carmine Vassallo, Sebastiano Panichella, Fabio Palomba, Sebastian Proksch, Harald C Gall, and Andy Zaidman. How developers engage with static analysis tools in different contexts. *Empirical Software Engineering*, pages 1–39, 2019.

[60] Alma Whitten and J Doug Tygar. Why johnny can't encrypt: A usability evaluation of pgp 5.0. In *USENIX Security Symposium*, volume 348, 1999.

[61] Glenn Wurster and Paul C Van Oorschot. The developer is the enemy. In *Proceedings of the 2008 New Security Paradigms Workshop*, pages 89–97, 2008.

[62] Fiorella Zampetti, Simone Scalabrino, Rocco Oliveto, Gerardo Canfora, and Massimiliano Di Penta. How open source projects use static code analysis tools in continuous integration pipelines. In *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, pages 334–344. IEEE, 2017.

[63] Misha Zitser, Richard Lippmann, and Tim Leek. Testing static analysis tools using exploitable buffer overflows from open source code. In *ACM SIGSOFT Software Engineering Notes*, volume 29, pages 97–106. ACM, 2004.

[64] Nist source code security analyzers. http://samate.nist.gov/index.php/Source_Code_Security_Analyzers.html.

[65] Owasp source code analysis tools. http://owasp.org/index.php/Source_Code_Analysis_Tools.

[66] Web application security consortium static code analysis tools. http://projects.webappsec.org/w/page/61622133/StaticCodeAnalysisList.

[67] Static analysis tools for security. https://www.dwheeler.com/essays/static-analysis-tools.html.

A User Study Briefing

Thank you so much for agreeing to participate in this study. As I mentioned in the email, we are researchers interested in improving the usability of different security-oriented static analysis tools. This study will provide valuable information on how to design and develop better support for the use of static analysis tools.

Today I'll have you to use two static analysis tools and ask a few questions about your experience using them. During this session I'll be recording the computer screen and audio from our conversation. You can let me know at any point if you'd like to stop or pause the recording. If all that sounds ok to you, could you please sign this form for me...

Consent form: <obtain consent from participant>

B User Study Task Briefing

In this scenario, you have been tasked with evaluating the security of two applications. These applications have been analyzed by a static analysis tool that detects potential security vulnerabilities. I'll have you use the first tool for about 15-20 minutes, then I'll ask you a few questions. Then I'll have you use the second tool for about 15-20 minutes and I'll ask you a few more questions.

As you are working with the tools, try to think aloud. So, say any questions or thoughts that cross your mind regardless of how relevant you think they are. If you are silent for longer than 30 seconds or so, I'll gently remind you to KEEP TALKING.

(If participants are silent for more than 30 seconds raise a "KEEP TALKING" sign [Sugirin 1999])

Prioritized list of tasks

1. Choose a vulnerability that you'd be most likely to inspect first.
2. Determine whether the reported vulnerability is actually a vulnerability.
3. Propose a fix to the vulnerability.
4. Assess the quality of your fix.

Do you have any questions before we begin?

(TURN ON SCREEN RECORDER, AUDIO RECORDER, AND BACKUP RECORDER!)

C Post-Study Questions

1. Which issues did you encounter when using the tool?
2. Which functionalities of the tool did you like most?
3. Which functionalities of the tool did you dislike most?
4. Which functionalities of the tool did you find useful?
5. Were there moments when you were confused?
6. Would you use this tool in your development work?

D Heuristic Walkthrough Guide

Pass 1

You work at a software company and are responsible for the security of your product. Your team uses [toolname] to detect vulnerabilities early in the development process. The tool detects several potential vulnerabilities in the code, but you only have a limited amount of time to resolve them. Given these constraints, work through the following tasks:

Prioritized list of tasks

1. Choose a vulnerability that you'd be most likely to inspect first.
2. Determine whether the reported vulnerability is actually a vulnerability.
3. Propose a fix to the vulnerability.
4. Assess the quality of your fix.

Repeat these tasks until you feel satisfied with your assessments. Use the questions below to guide your evaluation. Record any usability problems you encounter during this phase.

Guiding questions:

1. Will users know what they need to do next? It is possible that they simply cannot figure out what to do next.
2. Will users notice that there is a control (e.g., button, menu) available that will allow them to accomplish the next part of their task? It is possible that the action is hidden or that the terminology does not match what users are looking for. In either case, the correct control exists but users cannot find it. The existence and quality of labels on controls and the number of controls on the screen influence the user's ability to find an appropriate control (Franzke, 1995).
3. Once users find the control, will they know how to use it (e.g., click on it, double click, pull-down menu)? For example, if the control is a pull-down menu but it looks like a normal button, users may not understand how to use it. Users may find the icon that corresponds to the desired action, but if it requires a triple-click they may never figure out how to use it.
4. If users perform the correct action, will they see that progress is being made toward completing the task? Does the system provide appropriate feedback? If not, users may not be sure that the action they just performed was correct.

Pass 2

Heuristics

Guided by the knowledge you gained in Pass 1, you are now free to explore any part of the system. Evaluate the system using each of the following heuristics, which are derived from Smith and colleagues' 17 information needs [56]. For your convenience, short summaries of each heuristic are included here:

- Preventing and Understanding Potential Attacks
  Information about how an attacker would exploit this vulnerability or what types of attacks are possible in this scenario.
- Understanding Approaches and Fixes
  Information about alternative ways to achieve the same functionality securely.
- Assessing the Application of the Fix
  Once a fix has been selected and/or applied, information about the application of that fix or assessing the quality of the fix.
- Relationship Between Vulnerabilities
  Information about how co-occurring vulnerabilities relate to each other.
- Locating Information
  Information that satisfies "where" questions. Searching for information in the code.
- Control Flow and Call Information
  Information about the callers and callees of potentially vulnerable methods.
- Data Storage and Flow
  Information about data collection, storage, its origins, and its destinations.
- Code Background and Functionality
  Information about the history and the functionality of the potentially vulnerable code.
- Application Context/Usage
  Information about how a piece of potentially vulnerable code fits into the larger application context (e.g., test code).
- End-User Interaction
  Information about sanitization/validation and input coming from users.
- Developer Planning and Self-Reflection
  Information about the tool user reflecting on or organizing their work.
- Understanding Concepts
  Information about unfamiliar concepts that appear in the code or in the tool.
- Confirming Expectations
  Does the tool behave as expected?
- Resources and Documentation
  Additional information about help resources and documentation.
- Understanding and Interacting with Tools
  Information about accessing and making sense of tools available. Including, but not limited to the defect detection tool.
- Vulnerability Severity and Rank
  Information about the potential impact of vulnerabilities, including which vulnerabilities are potentially most impactful.
- Notification Text
  Textual information that an analysis tool provides and how that text relates to the potentially vulnerable code.
- Other Usability Problems / Notes

E Summary of Participants' Experience

Table 3: Summary of Participant Experience

ID	Tool 1	Tool 2	Security Familiarity	Java Familiarity	C++ Familiarity	PHP Familiarity	Professional Exp. (Years)
P00	FSB	Rips	●●●●●	●●●●●	●●○○○	●○○○○	7
P01	FSB	Rips	●●●○○	●●●●○	●●●○○	●○○○○	3
P02	FSB	Flawfinder	●●●○○	●●●○○	●●○○○	●○○○○	2.5
P03	FSB	CTool	●●○○○	●●○○○	●○○○○	●○○○○	4
P04	RIPS	CTool	●●●○○	●●●○○	●●○○○	●○○○○	3
P05	Flawfinder	CTool	●●●●○	●●●●○	●●○○○	●●○○○	7
P06	RIPS	FSB	●●●●○	●●●●●	●●●●○	●●○○○	1
P07	Flawfinder	FSB	●○○○○	●●●○○	●●○○○	●○○○○	0.5
P08	CTool	FSB	●●●○○	●●●●○	●●●●○	●●○○○	3
P09	Flawfinder	Rips	●●○○○	●●●○○	●●○○○	●○○○○	6
P10	CTool	Rips	●○○○○	●●○○○	●●○○○	●●○○○	3
P11	CTool	Flawfinder	●●●○○	●●●●○	●●○○○	●○○○○	0
P12	RIPS	Flawfinder	●●●○○	●●●●○	●●●●○	●●●●○	8

F Heuristics Summarized

Table 4: Summary of heuristics from Smith and colleagues [56]

Heuristic	**Description**
Preventing & Understanding Potential Attacks	Information about how an attack would exploit this vulnerability or what types of attacks are possible in this scenario.
Understanding Alternative Fixes & Approaches	Information about alternative ways to achieve the same functionality securely.
Assessing the Application of the Fix	Once a fix has been selected and/or applied, information about the application of that fix or assessing the quality of the fix.
Relationship Between Vulnerabilities	Information about how co-occurring vulnerabilities relate to each other.
Locating Information	Information that satisfies "where" questions. Searching for information in the code.
Control Flow & Call Information	Information about the callers and callees of potentially vulnerable methods.
Data Storage & Flow	Information about data collection, storage, its origins, and its destinations.
Code Background & Functionality	Information about the history and the functionality of the potentially vulnerable code.
Application Context / Usage	Information about how a piece of potentially vulnerable code fits into the larger application context (e.g., test code).
End-User Interaction	Information about sanitization/validation and input coming from users. Does the tool help show where input to the application is coming from?
Developer Planning & Self-Reflection	Information about the tool user reflecting on or organizing their work.
Understanding Concepts	Information about unfamiliar concepts that appear in the code or in the tool.
Confirming Expectations	Does the tool behave as expected?
Resources & Documentation	Additional information about help resources and documentation.
Understanding & Interacting with Tools	Information about accessing and making sense of tools available. Including, but not limited to the defect detection tool.
Vulnerability Severity & Rank	Information about the potential impact of vulnerabilities, including which vulnerabilities are potentially most impactful.
Notification Text	Textual information that an analysis tool provides and how that text relates to the potentially vulnerable code.

Security, Availability, and Multiple Information Sources: Exploring Update Behavior of System Administrators

Christian Tiefenau
University of Bonn

Maximilian Häring
University of Bonn

Katharina Krombholz
CISPA Helmholtz Center for Information Security

Emanuel von Zezschwitz
University of Bonn, Fraunhofer FKIE

Abstract

Experts agree that keeping systems up to date is a powerful security measure. Previous work found that users sometimes explicitly refrain from performing timely updates, e.g., due to bad experiences which has a negative impact on end-user security. Another important user group has been investigated less extensively: system administrators, who are responsible for keeping complex and heterogeneous system landscapes available and secure.

In this paper, we sought to understand administrators' behavior, experiences, and attitudes regarding updates in a corporate environment. Based on the results of an interview study, we developed an online survey and quantified common practices and obstacles (e.g., downtime or lack of information about updates). The findings indicate that even experienced administrators struggle with update processes as the consequences of an update are sometimes hard to assess. Therefore, we argue that more usable monitoring and update processes are essential to guarantee IT security at scale.

1 Introduction

"Keep your systems up to date" is one of the most popular pieces of advice that security experts give to end-users [33,55]. Supporting this, Khan et al. found that there is a correlation between not deployed updates and infected machines [35]. Systems can easily be hardened against vulnerabilities like Heartbleed [1] by applying updates. Regardless of that, many systems in the wild remain vulnerable for two years or more [57]. A prominent example of a situation where an update could have prevented severe data leakage is the Equifax breach [2], which occurred in 2017. Similar incidents seem not unusual as is reported by an industry report [42].

Related work studied user perceptions and experiences with system updates and found that the results are often not in line with current recommendations of experts from a security perspective. In most cases, concerns about functional issues or unexpected UI changes hinder individuals from updating their systems [63]. In addition, users often do not understand the importance of non-visual changes [63], as they come with security updates. In contrast to users who are responsible only for managing their own personal devices, system administrators are in charge of large and complex IT infrastructures while also being users. We argue that their update behavior can have severe implications at a much larger scale. Marconato et al. [43] observed the vulnerability life-cycle on different platforms and found that the time to patch and disclose vulnerabilities is decreasing. This finding can be applied to the Equifax breach and suggests that administrators are required to react in a timely manner.

Although general user concerns about system updates have been investigated in user studies, little light has been shed on the perspective of specific user groups (e.g., administrators or operators). Investigating administrators, Dietrich et al. [12] found that insecure configurations are often caused by institutional and individual factors, as well as time constraints. We assume that similar factors can have a negative impact on update processes. Administrators are often overworked [12], and updates are time-consuming. Secure systems, however, rely on updates and therefore, require regular attention by administrators. As the body of literature is still in an early state regarding administrators' update behavior, we follow an inductive approach to explore the processes and obstacles that administrators face when updating in a corporate context.

[1] http://heartbleed.com/, accessed 02/25/2020.

[2] https://www.theverge.com/2017/10/3/16410806/equifax-ceo-blame-breach-patch-congress-testimony, accessed: 11/20/2019.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

Our contributions are as follows:

- We conducted **seven qualitative interviews to explore how administrators experience, perceive, and act during the update process**.
- We conducted an **online survey with 67 valid answer sets to test our observations on a larger scale.**
- We confirm that current **update processes and system factors tend to endanger IT security** and we discuss critical factors that need to be addressed to **support administrators**.

The results suggest that the update process differs among companies for various reasons. Administrators face a variety of obstacles in their update routines, e.g., downtime or hard-to-foresee situations, that often hinder them from performing updates in a timely manner. Overall, we argue that system administrators would benefit from more usable mechanisms, and that providing such mechanisms could effectively improve IT security at scale.

2 Related Work

Two areas of research are specifically important for our work: 1) studies about update behavior and 2) studies investigating the security behavior of expert users. In the following, we present lessons learned from both research areas.

2.1 Users' Update Behavior

According to security experts, keeping systems and software up to date is an important security recommendation [55]. However, users may not follow this advice for reasons that are not related to security [54], and only a minority of non-experts actually considers software updates an important security measure [33, 48]. It has been repeatedly shown that users often delay or even avoid updates [22, 47, 64].

Investigation of the root causes of such critical user behavior has become a very active field of research. Previous work revealed diverse reasons for avoiding updates. Many users think that updates are not important because the link to security aspects often is not obvious [15, 25, 45, 53, 63–65]. Furthermore, users are often afraid of functional changes (e.g., UI modifications) [7, 62–64] or fear making mistakes [22]. Inconvenience is an important factor as updates can cause interruptions and take time [45, 64, 65]. Finally, bad experiences with previous updates and negative online reviews hinder the installation of future patches [15, 45, 59, 62]. This problem seems self-perpetuating, because the frequency of security updates is influenced by the emergence of novel attacks and thus, cannot be controlled by the vendor alone [56]. However, high update frequencies can lead to further negative reviews [21, 52].

Several countermeasures for mitigating the problem of delayed updates have been proposed. As one straightforward solution, automatic updates [65] and silent updates [13, 56] have been deployed. Although such mechanisms are very effective in keeping software up to date, they often cause confusion and irritation as they hamper the user's understanding of what is happening on their machines [14, 65]. Furthermore, some users might have good reasons to refrain from performing certain updates [14]. Therefore, user-centered solutions, such as providing more information [44, 50, 58, 59] and designing better notifications [16, 17, 24], have been repeatedly suggested as complementary concepts to further increase compliance rates.

2.2 IT Professionals and IT Security

Recently, researchers have started focusing on security-related usability problems of specific user groups [2]. In contrast to security advocates [29] or security analysts [27], most of these people are not security professionals. They are often knowledgeable in a specific domain, related to IT. Several recent studies addressed the problems of software developers [1,3,4,40]. For example, Acar et al. [1,3] investigated available sources of information and how these sources influence code security. Gorski et al. [40] showed that software developers benefit from API-integrated security recommendations and that such usability-optimized concepts can significantly improve security [60].

Several human-centered studies with system administrators were published between 2001 and 2007. In 2001, Hrebec and Stiber [32] studied the mental models of system administrators and found that these experts often struggle to understand the complex systems that they need to manage. In addition, the study participants reported a lack of formal education and the desire to solve problems by themselves. Barrett et al. [6] found that system administrators often lack situational awareness. Haber and Kandogan [28, 34] and Botta et al. [10] observed the tools and work practices of security administrations and IT professionals. Their results show that security administrators perform a lot of different tasks and need various skills like pattern recognition or inferential analysis to perform these tasks. They proposed, that new classes of tools need to be developed to counter the ever increasing complexity of the systems and attack-vectors.

In contrast to this early work, a few recently published papers investigated more specific problems of system administrators. Fahl et al. [18] studied non-validating X.509 certificates and revealed that about 30% of the responsible webmasters misconfigured their web servers accidentally. Ukrop et al. [61] analyzed the corresponding warnings and found that rewording can help administrators to make better informed decisions. Krombholz et al. [37, 38] showed that the deployment process for HTTPS is far too complex and that administrators struggle with finding secure and compatible configura-

tions due to the lack of conceptual mental models. Dietrich et al. [12] investigated the administrators' general perception of misconfigurations and identified missing or delayed updates as one of the root causes of these problems.

There exists work that discussed update processes in companies [8, 9, 46, 64]. For example, Vitale et al. [64] performed three interviews with technical staff concerned with updates and found that these professionals prioritized security aspects and licensing issues over potential usability consequences. This finding confirmed previous findings [46] that in a corporate context, business needs rather than user requirements drive update decisions. In contrast, Blythe et al. [9] reported that employees often rely on "security experts" in the company to manage updates and often lack a feeling of responsibility. Finally, the update challenges of system administrators have been indirectly considered by various researchers who proposed automatic tools to improve the manageability of the update process (e.g., [5, 26, 39, 49]). However, none of these concepts have been evaluated in a user study.

Parallel to our work, Li et al. [41] published a closely related paper in which they studied US-based system administrators in a qualitative fashion. They as well researched the update process in companies and found several *pain points* within the process. In contrast, our interview sample was drawn from German companies, thus representing a different culture. Overall, our study confirms most of their findings. We will separately discuss our findings in comparison to Li et al.'s in section 5.5 in more detail.

3 Interview Study

Although recommendations for patch management have been published[3], we are aware of only one other study that systematically investigated the update behavior of system administrators [41]. Therefore, we started with an interview study to identify important factors of the problem space.

This interview study aimed to provide answers to the following research questions with an emphasis on administrators' perceptions, challenges, and tools they use in their update routines:

1. **How can the update processes be described, and what common patterns are there?**
   Administrators are usually paid professionals who are responsible for updating large and complex IT infrastructures. This raises the question, whether, and if so, where, system administrators' updates processes differ from end-users' processes [62].

2. **What issues and obstacles do professional administrators face in their update routines?**
   We specifically aim at understanding the problems of administrators and their perception of update processes. Identifying obstacles in relation to processes, tools, and environments is indispensable to define important directions for future work.

3. **How are administrators informed about updates, and which sources of information do they use?**
   Related work has indicated that the source of information can have a significant impact on software security [1, 20]. Thus, we aim at understanding how administrators gather information and what sources they use.

4. **What kind of tools do administrators use to manage system updates, and is there room for improvements?**
   As usable security researchers, we are specifically interested in the tools involved in the update process. We hypothesize that although some tools are used on purpose and other tools are unavoidable, such tools can either complicate or ease the process.

3.1 Study Design and Procedure

We conducted seven semi-structured interviews in June 2018 to explore the participants' opinions, thoughts, and experiences. Based on three pilot-study interviews, we refined the interview guidelines to balance between informing the research questions and supporting a flexible exploration of the problem space (i.e., leaving enough room to add further comments). The interview was structured into (1) general questions about the daily work routine of the participant, (2) general experiences with updates, (3) a more detailed assessment of specific aspects, and (4) additional comments. The guidelines are in Appendix B.

All but one interview were conducted by the same researcher. Both researchers are experts in computer science and spoke the same native language as the interviewees. After an introduction to the purpose of the study, the participants were asked to sign a consent form. All participants gave their consent to being audio-recorded. We conducted one interview in person and six via telephone. All interviews were held in German. During the interviews, the interviewee and the researcher were allowed to take notes. The interviews lasted between 34 and 67 minutes and ended with a short questionnaire that collected demographic information.

3.2 Recruitment and Participants

We did not restrict our invitations to administrators working with a specific operating system, infrastructure or type of update. The only criterion for inclusion was that participants had to be in charge of, or in contact with, any kind of updates. Personal contacts were used as entry points to larger organizations and asked to forward the announcement to their employers' IT department. Additionally, we directly

[3] https://www.infosec.gov.hk/english/technical/files/patch.pdf, accessed 02/25/2020.

approached representatives of medium-sized and large companies at CeBIT 2018, a large international computer expo[4].

In total, we recruited seven participants at companies that had an office based in Germany. All participants reported they were in charge of system administration, although they had various job descriptions and managed different types of systems. Table 7 in the appendix presents more details about the sample. All the participants were male. For ease of readability in the following sections, we assigned the participants random names.

3.3 Analysis

The interviews were transcribed, and coded by two researchers. We coded open answers inductively following the approach of Wertz, Charmaz et al. [66]. The two researchers categorized the data according to the research questions presented in Section 3. The first three interviews were coded in a batch to establish the first codebook. Each of the following four interviews was coded separately. Then, the conflicts were discussed, and new codes were added to the codebook. We calculated the combined Krippendorff's alpha [36] before (0.61) and after (0.98) the discussion phase for each interview. Our goal was to use the qualitative analysis solely as a first step and foundation for the following quantitative study. Therefore, we refrained from continuing with interviews until theoretical saturation [23] was reached.

3.4 Qualitative Results

In the following, we present the results from the interview study with respect to the research questions.

3.4.1 Update Processes

In Table 1, we present the sum of all extracted process phases, including all reported steps that were performed in these phases. Overall, the update process varied in time and structure among participants and tended to be variable even for individual administrators, depending on the software that needed an update. Cyril reported he worked in a client environment with Windows systems. He was concerned mainly with regular update cycles. Therefore, he was able to prepare for update events (e.g., briefing the team, allocating resources, allocating maintenance windows, and gathering information). Four out of seven participants reported they relied on fixed update cycles for client systems, although Zelko reported that this was not always possible in practice. In contrast, Lorenz, who worked at a smaller company, reported that employees at his company were responsible for their systems. When we discussed more specific software, the answers became more diverse. Milan usually builds packages to automate the distribution, but Markus tends to perform manual installations.

[4] https://www.cebit.de/, accessed 02/25/2020.

Although participants' responsibilities differed, we were able to identify common patterns in the update process. Most of these phases can be mapped to those of client users [62]. However, we identified three major differences:

Some administrators perform extensive testing before installing the update on a live system. For example, Julian utilized up to three stages. Zelko, who stated, that "[E]ven if there is a risk that the update breaks something, we install them timely", utilized two test stages. First, he tested the update with virtual machines that simulate the client landscape, and then he rolled out the updates for a small group of colleagues.

Updates are rolled out step by step. The participants reported that often not all systems are updated in one batch. This allows the administrators to minimize the number of misconfigurations once an update fails, but constraints on resources are also a reason for this. For example, Julian reported that the network would be used to capacity if all systems were patched at the same time.

The preparation step is structured and involves planning and research of resources and the allocation of time slots. Five participants explicitly reported they conduct online research before they install an update. In addition, Alexander told that important update decisions are often made in group discussions.

3.4.2 Obstacles

We identified various obstacles that hamper the administrators' task of performing updates. In Table 1, we connect and report obstacles to the phases of the update process. In the following, we discuss common obstacles in more detail:

Downtimes. The participants stated that downtimes are a serious obstacle in the update process which often cause delayed deployments. As soon as a reboot is necessary, and there is no redundant system, downtime is induced. Alexander gave anecdotal evidence of a mitigation strategy: Upgrading from Solaris 10 (which required significant downtime) to Solaris 11 (which supports near to hot-swap updates and an easy rollback) increased update frequencies from three times a year to once a week.

Dependencies. The participants reported patches that break dependencies usually delay the process. Although this may not be surprising, it highlights the problem of dealing with dependent systems that cannot be patched in time. Further dependencies are introduced as part of the infrastructure landscape. For example, some systems depend on other systems to be available at boot time (Markus). Assessing these dependencies and then following the right order makes the process highly complex. Another type of dependency is towards the vendor of the software or hardware. An example of this can be as trivial as no available patches, even if a vulnerability is public, as Lorenz reported for the Meltdown case.

High frequency and large files. Every update takes re-

Phase	Step	Obstacles
Information	Becoming aware	
	Further details	Unsatisfying communication with the publisher*
Deciding	Discussion	Stability (1); Risk of exploits (2); Performance (1); Priority (2); Missing expertise (1)
Preparation	Planning	Planning itself (3); Time of release (3); Communication (1); Missing documentation about the system and processes*
	Backup	
	Waiting for release	
	Obtaining the patch	Missing patches (1)
	Automating	
	Informing users	
Testing	Test system	Testing itself (1); Broken dependencies (4); Resources*; Frequency of updates*
	Pilot system	
	Problem solving with manufacturer	
Installation	Installation itself	Failure (2); Missing configuration options (1); Social pressure; System resources (2); Complexity (3); Missing tools (3); Heterogeneous system (6); Company structure (3); Impact on systems/users (2); Downtime (1); Installation method (manual/automatic) (1,1)
	User interaction	Waiting for users (1)
	Reboot	Reboot itself (3); Old/Slow hardware (1)
Post-Installation	Documentation	
	Testing/Monitoring	
	Troubleshooting	
	Reversing	Missing backup, failover, or redundancy*

Table 1: Overview of phases, steps, and obstacles. The number in brackets denotes the number of participants who mentioned this aspect in the interviews. *Additional obstacles were found through the questionnaire.

sources: for example, time, workforce, CPU, and data storage. Zelko reported that big update files, which are often a consequence of combining functional updates with security patches, can cause problems. To handle resource constraints, updates are rolled out in multiple but smaller batches (Julian).

Competing priorities. Similar to standard users, administrators' decisions to perform updates are influenced by various factors. Participants reported stability considerations, the risk of an exploit, and performance issues as influential aspects. The fact that some systems do not separate security and feature updates may intensify this situation. Finally, required resources are sometimes allocated to other processes that have higher priority. Alexander reported that "the decision [to update] is always based on the sum of available information". As mentioned in Section 3.4.1, group discussions are an important part of the process. However, the need for communication can also delay updates (Milan).

Human Factors. In addition to technological and structural constraints, the administrator faces other obstacles. Missing expertise or a lack of knowledge can lead to situations where administrators rely on third parties. In this regard, Lorenz acknowledged that he does not always know how to act correctly. Or as Markus put it, he has to trust the vendor that the classification of the patch is correct. System administrators have to trust the information they get from the software developer, vendor, or other source. Another factor we identified is social pressure, as Lorenz reported, "And you look like an idiot, when you kill a git server. [...] That chases me." Another aspect that makes updating harder for administrators was software which is managed by end-users. Such software is often installed without the knowledge of administrators and makes the update process more complicated because it is not integrated in standard processes.

3.4.3 Sources of Information

The participants reported they use various methods to inform themselves about security updates and vulnerabilities. Five out of seven participants reported they use third-party sources that were independent of the software publisher, such as popular news portals or blogs. This information is usually supplemented by publisher-related newsletters and specific mailing lists, such as the Ubuntu-security mailing list (Lorenz). Cyril mentioned specialized third-party services that push information about available patches. Others got more specific and

reported that they use tools like SCCM[5] or Nessus[6] which serve as sources of information.

3.4.4 Tools

The participants reported OS-integrated tools and special purpose tools that are used to update servers and clients and that serve as sources of information. The purpose of such tools ranged from monitoring systems (Julian) to complete automation of the update process, such as SCCM or WSUS[7] (Markus). Participants also named external services (e.g., Shavlik[8]) that test and pre-filter patches for companies. Although automation of update processes was an important goal for participants, it had not yet been fully implemented. Software that is not covered by such tools, meaning not integrated by default, has to be updated manually or integrated. This seems to be the case when the vendors or the operating systems differ (e.g., using Microsoft WSUS to update Adobe Flash Player). Although the integration is possible, it is connected to additional effort and is not always done (Markus), e.g., if it affects only a small group of clients (Milan). Concerning future developments, Lorenz was less optimistic and brought up that the time investment in tools that would ease the workflow was not a high priority.

3.5 Key Observations

We performed an interview study of administrators' update behavior. Based on the research questions, we were able to describe update processes, common obstacles, information retrieval, and the use of software tools. We extract a series of key observations to guide the construction of the quantitative study, following the interviews. Table 1 provides an overview of the process phases, tools, and obstacles that administrators face in their daily lives according to the participants. Table 2 presents nine key observations, which were formulated based on the qualitative findings and then categorized in three groups: "Update Process and Information," "Update Obstacles," and "Human Factors." In the next section, we report on a quantitative online survey which was performed to shed further light on the update behavior of system administrators.

[5] https://www.microsoft.com/en-us/cloud-platform/system-center-configuration-manager-features, accessed 02/25/2020.

[6] https://www.tenable.com/products/nessus/nessus-professional, accessed 02/25/2020.

[7] https://docs.microsoft.com/en-us/windows-server/administration/windows-server-update-services/get-started/windows-server-update-services-wsus, accessed 02/25/2020.

[8] https://www.ivanti.com/company/history/shavlik, accessed 02/25/2020.

ID	Observation
	Update Process and Information
U1	Online sources are an important source for administrators to get informed about updates.
U2	Small companies have no formal update process.
	Update Obstacles
O1	Performance considerations often hinder the installation of an update.
O2	Update-caused downtimes delay the installation of an update (e.g., reboots)
O3	Problems after the installation of an update on the live system are only a minor concern.
O4	Lack of information hinder the update process.
O5	User action (e.g., installing a software without the knowledge of the admin) can circumvent the update process and render it useless.
	Human Factors
P1	Administrators of big companies feel sufficiently trained.
P2	Administrators think that timely updates are important.

Table 2: Key observations based on qualitative results.

4 Quantitative Online Survey

Following the interviews, we performed a quantitative online survey. We created statements based on our observations in the interview study and developed an online questionnaire to quantify and enrich them.

4.1 Procedure and Structure

The recruitment process for the preliminary interview study indicated that system administrators are inherently short on time, and thus, minimizing the time to fill out the survey was indispensable to obtain a sufficient number of responses. Therefore, most of the questions were based on simple answer types, such as check boxes or rating scales. To further motivate participation, we offered an opt-in for a raffle of 3D prints. Every tenth participant had the chance to win a 3D-printed model of their choice. E-mail addresses were collected only for this raffle, stored separately, and deleted afterward. Twenty-three entered their contact email address of whom no one was interested in a print. After participants had given their consent to take part in the study, the survey started. Completion took about 10 minutes.

To support many different circumstances, we framed questions in a way that answers could be related to the current position or if not applicable, to the last position as system administrator. We started by collecting demographic data (e.g., age), personal information (e.g., years of experience), information about the work environment (e.g, their role, company size), and information about update processes (e.g., existence

ID	Statement	1	2	3	4	5	*	Plot	Median
O1	Performance considerations hinder the installation of an update.	24	27	7	9	0	0		2
O2	Downtimes caused by the update process hinder the installation of an update.	8	22	13	18	6	0		3
O4	A lack of information about the update hinder the installation of an update.	15	19	18	9	4	1		2
P1	I feel sufficiently trained as an administrator.	1	7	13	29	17	0		4

Table 3: Overview of the responses to statements regarding the frequency on a 5-point scale from "1 - Never" to "5 - Always" (* "Not sure") and their connection to the key observations.

ID	Statement	1	2	3	4	5	6	7	*	Plot	Median
O3	Post-installation problems in a live system are only a minor concern because they don't happen frequently.	8	9	8	5	12	16	9	0		5
O5	Users often install software without the knowledge of the administrator.	18	9	7	8	12	6	7	0		3
P2	Deploying security updates in a timely manner is important.	0	1	0	0	7	18	41	0		7

Table 4: Overview of the responses to statements regarding the attitude on a 7-point scale from "1 - Strongly disagree" to "7 - Strongly agree" (* "Not sure") and their connection to the key observations.

of formal processes). In the second phase, participants rated 1) the frequency of specific events using 5-point scales ranging from "1 - Never" to "5 - Always" and 2) indicated their agreement with different statements using 7-point scales ("1 - Strongly disagree" to "7 - Strongly agree"). The questions were presented in random order for each participant. The questions were chosen based on our observations and thus, examined the impact of obstacles (e.g., "Downtimes caused by the update process hinders the installation of an update"), human factors (e.g., "I feel that I am sufficiently trained as an administrator"), and information sources (e.g., selection of sources used). The questionnaire ended with an open-ended question about the biggest obstacles in the update process that we coded afterwards. The new categories are marked with an asterisk in Table 1.

To ensure the internal consistency of the collected data, we added an attention check based on the negation of one of these questions. Five participants, who answered both questions with a different polarity, were excluded from the evaluation.

4.2 Recruitment and Participants

To attract professional system administrators, we decided against using crowdsourcing platforms like Amazon Mechanical Turk. Instead, we reached out to community sites like Reddit and specialized forums. Additionally, we used Twitter and followed a similar approach as we did in the interview study. Posting in forums resulted in 66 answers, advertising on Twitter resulted in 67 responses, and using personal contacts in companies to spread the questionnaire contributed eight answers.

The English survey was active for 14 days in September 2018. During this time, the questionnaire was started 141 times and completed by 72 (51.1%) participants. As reported, five data sets were excluded from the analysis due to failed attention checks, resulting in 67 valid data sets. The participants' age ranged between 22 and 55 years. Fifty-eight of them were male, one female, three reported "Other" and five preferred did not specify their gender. More than 61% (41) work in European countries. The biggest group of the participants pool work in Germany (22), but we also received answers from other continents, like North America(19), Australia (2) or South America (1). Table 6 in the appendix provides an overview of the participants' demographics. The job-related education of our participants can be classified as "unspecified training," "vendor training," "self taught," and "experience at the job." Most of the participants worked in a team (39), 16 were a team leader, and 10 worked alone. In the following, we report on the data gathered by the questionnaire.

4.3 Results

In the following, the results of the online survey are presented structured by the main categories presented in Table 2. The observations from the interviews suggest that company size may have an influence on different factors. To assess this point, we divided the data sets in two groups: 34 companies with 250 employees or fewer were tagged as small and medium-sized enterprises (SMEs), and 33 companies with more than 250 employees were defined as large enterprises [11]. This was found to be a suitable comparison because post-hoc we had comparable group sizes. A controlled analysis of additional

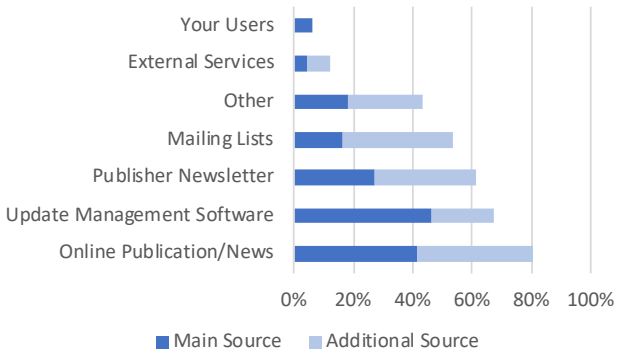


Figure 1: Distribution of information sources used by the administrators (n=67).

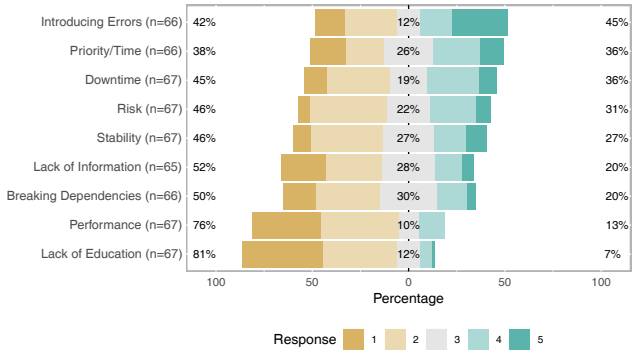


Figure 2: Frequency of considerations that hinder the installation of an update. The scale ranged from "1 – Never" to "5 – Always." Not included are "not sure" or missing answers.

factors was not feasible at this stage, and future work should consider other aspects (e.g., experience, type of systems, and team size). Tables 3 and 4 show the answers of the participants to the statement they were presented.

4.3.1 Update Process and Information

U1 Figure 1 presents the *sources of information* administrators use to learn about (new) updates. Most of the participants reported a median of three different sources. Third-party online publications are the most frequently used sources of information. They served as a source for 54 (81%) participants, and 28 of all 67 participants (42%) even declared them the main source of information. When focusing on the main source of information, we found that update management tools are essential for most administrators (46%). Fisher's exact test indicated no statistically significant differences between differently sized companies ($p = 0.2242$). Using an optional comment field, some administrators added other sources of information, such as vendors, the online community (e.g., Twitter), work experience, and active monitoring of systems. Due to the structure of the questionnaire, we cannot make statements about how the participants ranked the quality of those sources. We do not know whether they use one source to get informed about the occurrence of an update and then use another to capture details.

U2 To investigate the *existence of formal update processes*, we asked the participants if 1) "there is a written document," 2) "no document but an informal guideline," or 3) "no defined process" in their company. Twenty-eight (42%) participants indicated the existence of formal processes, 26 (39%) administrators had at least informal guidelines for performing updates, and 13 (19%) participants indicated that there are no predefined processes. A comparison of the use of formal, written update processes in differently sized companies revealed a statistically significant difference between large companies (57.6%) and smaller ones (26.5%), ($p = 0.0136$, $ratio = 3.769$, Fisher's exact test). This indicates that small companies make less use of formal update processes. The lack of such a process is not uncommon in our sample, as 10 out of 34 of the small companies did not report any kind of defined process.

4.3.2 Update Obstacles

Figure 2 shows the share of administrators who have faced specific obstacles during daily update routines. Quantifying the observations, we found that general risk assessments are known to most of the participants (94%) while deciding to deploy specific updates. Only four (6%) participants answered that they never considered assessing risks as an obstacle, while 63 agreed they did so at least sometimes.

O1 to O4 When asked about more specific obstacles, or risks, *stability considerations* represented the biggest issues that had been considered by 61 (91%) participants in the past. Similarly, 59 (88%) participants considered *downtime* as a specific obstacle. *Lack of information* (50, 77%), performance issues (43, 64%) and educational aspects (39, 58%) were the least prevalent obstacles in the sample. However, even those factors were considered by a majority of the participants. Finally, we performed Mann-Whitney U tests to investigate the impact of company size on the prevalence of obstacles: We could not find statistically significant differences [9].

Fifty-five percent seemed to agree that *problems after the installation of an update* are only a minor concern. However, eight participants strongly disagreed with the statement. Five

[9] stability considerations: $p = 0.814$, downtime: $p = 0.324$, lack of information: $p = 0.655$, performance issues: $p = 0.067$, educational aspects: $p = 0.752$, introducing errors: $p = 0.611$, risk considerations: $p = 0.415$, breaking dependencies: $p = 0.387$, priority: $p = 0.559$

Interval	
Hours to a day	11
Within a week	19
Within two weeks	8
Within one month	11
More than a month	9
No answer/no usable information (e.g., missing unit)	11

Table 5: Reported time intervals between the release of an update and deployment on all systems.

were undecided, and 25 (37%) disagreed in some way. To cover potential reasons for the answers, we assigned participants to two groups: those who do some kind of testing before installing updates on the live system ($n = 45, 67\%$) and those who do not ($n = 22, 33\%$). There was no statistically significant difference ($p = 0.2553$, Mann-Whitney U test) meaning that having a testing phase seems not to prevent all problems after the installation. Due to the sample size, we could not investigate if the company size is a significant factor in this regard.

O5 Another aspect in the interview study was the *user rights*. The agreement to the observation "Users often install software without the administrators' knowledge" was diverse. Although there was a tendency to disagree, as can be seen by the low median (3), there were also seven strong agreements. We found no statistical significance that would have supported our assumption that IT companies may have a different distribution on this than non-IT companies.

4.3.3 Human Factors

P1 Seventeen (25%) administrators reported that they always *feel sufficiently trained* for dealing with updates. However, 50 (74.6%) participants already faced situations for which they did not feel sufficiently trained. An evaluation of the impact of the administrator's company size indicates that administrators at large companies ($Median = 4$) more often feel sufficiently trained than their colleagues at smaller companies ($Median = 4$), Mann-Whitney U test: $U = 358.0, p < 0.01$, two-sided.

P2 Finally, all administrators (except one) somewhat agreed that *timely updates are important*. The self-reported time span between the release of an update and its installation can be seen in Table 5. While some participants reported deploying updates within a day, there were nine cases where updates needed more than a month. Optional comments given by the participants supported the findings that downtime, complexity, and dependencies are common reasons for such delays.

4.3.4 (Missing) Distinction between Security- and Feature-Updates

The interviews revealed that security- and feature-updates are often hard to distinguish. While we did not ask for the share of security-related patches in our interviews, the survey participants reported that 56% (ranging from 5-100%) of the overall updates involved security-related ones.

5 Discussion and Implications

Our work identified multidimensional problems that should be addressed by multiple stakeholders (e.g., software vendors or the companies themselves). In this section, we reflect on our results, provide actionable recommendations for these stakeholders and suggest directions for future research. We acknowledge that many aspects reported in this paper may seem like "common sense". With this work, we add to the scientific evidence in this very broad area with several factors that influence the update process and directions for further research and discussion.

5.1 Security Implications

Our results are in line with Li et al. and show that even professionals cannot always deploy updates in a timely fashion. This can be a security issue since outdated systems are often vulnerable to exploits. The administrators we asked were aware of this problem and agreed that deploying updates in a timely manner is important. However, we found that external factors such as compliance with company-specific rules, inflexible processes and communication overhead (e.g., leadership approval) still delay updating in practice. Future work needs to take a more holistic view and investigate technical and social factors in the update process. We need to understand which people are involved in these processes and how their communication can be supported. In addition, we need to develop approaches to better communicate the urgency of specific patches as today, the rating is often not clear [42].

5.2 Update Process

The results showed that the update processes of system administrators are diverse and complex. Although the update processes of administrators can be matched to the end-user phases [62], the identified phases differ in the details. In particular, gathering information and discussing update decisions were identified as important but time-consuming steps. As many administrators reported they make decisions in group meetings, we raise the question of how individual administrators can be supported in their decision-making process. The preparation process takes time and involves extensive testing. Although the testing processes were handled differently, they usually involved multiple iterative stages. This indicates that

administrators have to go through the whole update process multiple times. Two findings were primarily interesting: 1) Many companies lack formal processes, and 2) the update process is highly complex and lacks automation. The insights into this process provide important directions for future research and immediate action items for software vendors, such as the following:

- Formal processes seem to be more frequently used in large companies. Whether formal processes help to reduce the burden of decision-making and ease the overall process should be researched; that is, in what way they influence the update process (e.g., can well-defined responsibilities speed up the decision and do they lead to more and faster updates?) and where possible trade-offs can be expected (e.g., decreased complexity versus more time needed).
- The high number of iterative steps must be supported, e.g., with automation approaches. Thus, it is important to understand which phases of the process are critical and which parts can be effectively supported by tools.
- A possible approach for improving the process could be to connect more effectively virtual teams of administrators who share similar responsibilities and manage similar systems. Supporting such concepts with feasible tools can quickly lead to shared knowledge of best practices and experiences resulting in a better overview of the effects updates have on their systems. We hypothesize that especially smaller companies would profit from that.

5.3 Obstacles

The findings indicate that administrators face severe obstacles that often hinder them from performing timely updates. In line with Dietrich et al.'s work [12], the findings show that the problems administrators face are diverse and interconnected. Corresponding to Hrebec and Stiber's findings [32], individual-related factors, such as negative and positive experiences with updating, as well as education, come into play. The findings provide a baseline for future research questions and immediate action items for software vendors, such as the following:

- Due to the highly diverse landscape of large-scale systems, future research should further explore contextual factors and different populations of administrators. Differentiation of the various types of administrators could help to better categorize participants and understand their diverse problems and challenges. Related to this point, the check of the external validity of the research would benefit from better differentiation of types of administrators. However, a practicable taxonomy for this is still missing.
- Software development should focus on reducing downtime and providing rollback mechanisms that encourage administrators to take the risk of potential negative effects on availability.
- Researchers and software vendors should investigate on how to provide reliable information and accurate documentation of the effects of an update and occurring problems right in the moment and at the place the update is going to be installed.

Therefore, we hypothesize that supporting administrators' situational overview will have positive effects on timely updates. Finally, minimizing consequences by providing reversible updates, or just updates that have very small effects, could furthermore help administrators to update. As an example, dynamic software updates (DSU) [31] seems like a promising technique to contribute to this area and could be evaluated from this perspective.

5.4 Coping Strategies

As a consequence of facing obstacles, system administrators have developed a diverse set of coping strategies. Although the degree of usage varied among participants, an important countermeasure against the growing complexity is the use of tools that monitor update processes and support to (partly) automate installations. Because administrators expressed the desire for more automation, the findings emphasize the importance of the area of research that deals with the development of such concepts [5, 26, 39, 49].

To cope with the problem of limited resources combined with growing package sizes, the participants started to divide update processes into multiple batches. This can have the advantage of allowing more feedback loops and of reducing the load on the network. However, at the same time, this process increases the number of required iterations for single patches. Although we argue that the footprint (e.g., resources needed to roll out), especially of security updates, should be minimal, this may not always be possible.

Based on the findings, we provide the following recommendations to support existing coping strategies and for the development of novel solutions:

- Hot swap functionality and small-sized patches which enable administrators to estimate the impact of the installation on their systems, have the potential to further ease the update processes.
- Update management tools should better support the integration of third-party software.
- Administrators' coping strategies are still not sufficiently understood. Thus, researchers should focus on systematically investigating different coping strategies for various obstacles, identify desirable behavior and analyze in which way the human aspect contributes to this.

5.5 Comparison to Results by Li et al. [41]

As mentioned in Section 2, a thematically similar publication emerged independently while we were working on this research. Li et al. published a study on system update processes among US American system administrators, identifying an update process that was very similar to ours [41].

The Update Process

While Li et al.'s process emerged entirely from their interview response data, our update process was informed by theoretical work by Vaniea et al. [62]. This could explain minor differences such as the separate testing phase we introduced to highlight the difference to the end-user process. Despite this difference, overall, we consider the identified processes to be very similar. In alignment with their findings, we can confirm that in the information-phase, administrators use multiple sources to derive information about updates. We didn't find any statistical difference in the number of sources used between administrators working in different companies (big vs small) in our sample. Li et al. reports on the frequency of the used sources and that three quarter of their participants used security advisories or direct vendor notifications. In our data, 81% informed themselves using online publications and 63% relied on publisher newsletters. We can add that despite having multiple sources (median=3), our population uses update management tools as their main source followed by online resources.

Both works identified the deciding-phase. We can match most of our identified obstacles to the reported factors of Li et al. With a slightly different perspective, we can add an additional reported obstacle that focuses more on the administrator executing the process than the update: missing expertise.

We can support Li et al.'s finding that testing is an important phase in the process and we encountered the same approaches: "Staggered deployments" and "Dedicated testing environments". As 83 of 102 (81%) of their survey participants included some form of testing, a slightly smaller, but still the major, part of our participants 45/67 (67%) reported the same.

As for the remaining two phases, our works differed in focus. While Li et al. extensively discussed the method of deployment (automatic vs. manual) and the decision of when to deploy in the deployment phase, our work concentrates on the obstacles the administrators face in this phase. For the post-installation phase, their work presents the ways in which administrators deal with update issues, while we report on the frequency of the occurrence of such issues (O3) in Section 4.3.2.

Obstacles in the Update Process

Li et al. identified challenges faced by administrators within this update process that can be categorized as: (1) obtaining relevant information about relevant updates and deciding, (2) preparing, testing and deploying updates in a timely fashion, (3) recovering from update-induced errors, and (4) organizational and management influence [41]. Our identified obstacles (cf. Section 3.4.2) are in line with these obstacles. Li et al.'s work reports that identifying the relevant information in an update can be a challenging task. We can confirm this (O4) and show that this was mentioned by 77% of our participants.

Automation can help to deploy updates sooner and more frequently. Li et al. have found several obstacles such as dependency and compatibility considerations or host heterogeneity as factors that have an influence on update deployment. In addition to those, we have found additional ones such as missing tools or performance considerations in our data set. Table 1 provides a summary of our findings that assigns the problems to the phases in which they occur.

In general, while their work reveals the existence of those problems, we can complement these problems with the frequency of the problems that our survey participants stated. Li et al. report that the recovery of updated-induced errors is a problem that we can enrich with the fact that this seems to be of mixed importance (O3). This could indicate that this is a context-dependent factor, and a more detailed research must be undertaken in this regard.

Also, Li et al.'s work reports on the existence of organizational oversight that hinders or delays updates in some cases. We can also find this problem and show that this, among stability and risk considerations, is of more importance than factors such as performance considerations.

Demographics

While both Li et al.'s and our study are very similar in methodology, they differ in a key point: the recruited sample. Li et al. sampled only US-based administrators, while we recruited our interview-study population from Germany and our survey participants were mostly (41 of 67) European-based. Despite work culture in the US and Europe (e.g. in Germany [19,30,51]) being distinctively different (stemming from cultural differences in education, law, and professional socialization, among others), both studies report similar findings. We are thus in the fortunate situation to not only have our methodology and findings independently validated within a close distance in time, but also to confirm that the phenomena we identified are relevant across both US and European system administrators.

On interpreting the independently compiled findings, we have an indication that the system administration process is not as susceptible to cultural differences (at least in Western societies) as other fields of work. This might be connected to the rather globalized nature of IT infrastructure. Both participant pools used similar software, e.g., SCCM or WSUS (cf. Section 3.4.4). It is reasonable to assume that the technical challenges are similar. Comparing both papers, we could not

find any differences that originate in individual or organizational factors. If this can be confirmed in further studies within different countries such as China (the largest producer of IT hardware and systems[10]), Estonia (the often considered "most advanced" country within the EU in terms of digital transformation[11]), or Qatar (the largest economy in the Middle East according to GDP per capita [12]), this would significantly widen the recruitment possibilities for future studies within the field of system administration.

6 Limitations

The population we refer to as administrators is inherently diverse in terms of responsibilities, education, and previous experience. Depending on the size of a company, administrators have different responsibilities and work either in isolation or in larger teams. Furthermore, the security requirements depend on the types of products and services a company offers. Also, there is no unified career path for administrators, and one must not necessarily have a degree or certificate of any kind to become an administrator. Because of all these aspects, the results are not generalizable and thus applicable other populations of administrators with different demographics or training. The participants in the online survey were mainly from Europe and the United States. In these regions, technical staff like administrators are predominantly male which is why the sample was heavily biased in terms of gender. Due to our recruitment strategy for the quantitative study, the sample potentially suffered from self-selection bias, as was likely also due to the completion rate (51.1%) of the survey. Regarding our questionnaire, we did not ask the participants about their current employment status. This could result in answers from people that worked as an administrator previously and are now in a different position. However, due to the mentioned self-selection bias we think that the participants are still somehow active in this area. Also, we did not collect information about the systems and software, the administrators were in charge of. Because of this, we cannot report possible existing differences between, e.g., different operating systems or widespread versus niche software. The analysis is based on self-reported data, and thus, participant reports are highly subjective. We have no reason to believe that social desirability and recall bias are uncommonly strong in the sample because the interviews and related work showed that administrators tend to admit that they do not know about everything [32]. However, this must be taken it into account, especially when talking about risk, obstacle perception, and individual perception (e.g., P1). Finally, the qualitative interviews provided useful insights but did not reach saturation (cf. [23]). However, the potential lack of saturation is alleviated as the qualitative analysis was primarily used as an exploratory first step to build hypotheses. The answers to the free-text questions on the questionnaire did not bring up many new topics which make us confident that the most common real-world problems were covered. But, although several different issues were covered, we make no claim for completeness.

[10] https://www.mckinsey.com/~/media/mckinsey/featured%20insights/china/china%20and%20the%20world%20inside%20the%20dynamics%20of%20a%20changing%20relationship/mgi-china-and-the-world-full-report-june-2019-vf.ashx

[11] https://www.wired.co.uk/article/estonia-e-resident, accessed 11/21/2019.

[12] https://www.cia.gov/library/publications/the-world-factbook/rankorder/2004rank.html, accessed11/21/2019.

7 Ethical Considerations

Our institution located in central Europe does not have a formal IRB process for this type of study but has a series of guidelines to follow. According to these guidelines, we limited the collection of personal information as much as possible and collected data separately from contact information. Furthermore, all the processes complied with the European General Data Protection Regulation (GDPR). As the administration of services in a corporate environment is a sensitive topic, we did not collect detailed information about the companies' infrastructures. In addition, participants were explicitly given the chance to drop out at any time during the study. Finally, we emphasized the option to skip questions that participants preferred not to answer.

8 Conclusion and Future Work

This work contributes a mixed-methods study that revealed how administrators incorporate security updates in their daily work routines, what obstacles they experience, and what coping strategies they apply. We found that even experienced administrators find it hard to predict the direct consequences of applying an update and are heavily concerned about potential downtimes. Another interesting observation was that administrators often rely on information that is not provided by the (software) vendor but by online media or by their peers, who often face similar struggles. Among other things, the findings imply that there are aspects that vendors can influence such as, e.g., provide sufficient documentation or more granular updates, which can help to motivate administrators to update and support them in the update process.

Based on the insights presented in this paper, we recommend the following topics for future work: (1) Investigate current established formal processes and evaluate their effectiveness in supporting timely updates. (2) Create computer-supported solutions that enable better communication between administrators and in this way, enhance the transfer of knowledge. (3) Design and evaluate feasible tools that support situational awareness, e.g., by helping administrators to find out about relevant updates and provide them with the information they need.

9 Acknowledgements

We thank Karoline Busse for their support in the interview study and valuable input on the discussion. Also, we thank Jennifer Seifert for her help with the sociologial theory. Finally, we want to thank Matthew Smith, Frank Li and the anonymous reviewers that helped with their constructive feedback to improve this work.

References

[1] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. You Get Where You're Looking for: The Impact of Information Sources on Code Security. *Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016*, pages 289–305, 2016.

[2] Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. You are not your developer, either: A research agenda for usable security and privacy research beyond end users. In *Cybersecurity Development (SecDev), IEEE*, pages 3–8. IEEE, 2016.

[3] Yasemin Acar, Christian Stransky, Dominik Wermke, Charles Weir, Michelle L Mazurek, and Sascha Fahl. Developers need support, too: A survey of security advice for software developers. In *Cybersecurity Development (SecDev), 2017 IEEE*, pages 22–26. IEEE, 2017.

[4] Hala Assal and Sonia Chiasson. Security in the software development lifecycle. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 281–296. USENIX Association, 2018.

[5] Rekha Bachwani, Olivier Crameri, Ricardo Bianchini, and Willy Zwaenepoel. Mojave: A recommendation system for software upgrades. In *MAD*, 2012.

[6] Rob Barrett, Eser Kandogan, Paul P. Maglio, Eben M. Haber, Leila A. Takayama, and Madhu Prabaker. Field studies of computer system administrators. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work - CSCW '04*, page 388, New York, New York, USA, 2004. ACM Press.

[7] Ofer Bergman and Steve Whittaker. The cognitive costs of upgrades. *Interacting with Computers*, 30(1):46–52, 2017.

[8] John M. Blythe and Lynne Coventry. Costly but effective: Comparing the factors that influence employee anti-malware behaviours. *Computers in Human Behavior*, 87:87 – 97, 2018.

[9] John M Blythe, Lynne M Coventry, and Linda Little. Unpacking security policy compliance: The motivators and barriers of employees' security behaviors. In *SOUPS*, pages 103–122, 2015.

[10] David Botta, Rodrigo Werlinger, André Gagné, Konstantin Beznosov, Lee Iverson, Sidney Fels, and Brian Fisher. Towards understanding it security professionals and their tools. In *Proceedings of the 3rd symposium on Usable privacy and security*, pages 100–111. ACM, 2007.

[11] The commission of the european communities. Commission recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises, 2003.

[12] Constanze Dietrich, Katharina Krombholz, Kevin Borgolte, and Tobias Fiebig. Investigating system operators' perspective on security misconfigurations. In *Conference on Computer and Communications Security (CCS'18)*, 2018.

[13] Thomas Duebendorfer and Stefan Frei. Why silent updates boost security. *TIK, ETH Zurich, Tech. Rep*, 302, 2009.

[14] W. Keith Edwards, Erika Shehan Poole, and Jennifer Stoll. Security automation considered harmful? In *Proceedings of the 2007 Workshop on New Security Paradigms - NSPW '07*, page 33, New York, New York, USA, 2008. ACM Press.

[15] Michael Fagan and Mohammad Maifi Hasan Khan. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*, pages 59–75, 2016.

[16] Michael Fagan, Mohammad Maifi Hasan Khan, and Ross Buck. A study of users' experiences and beliefs about software update messages. *Computers in Human Behavior*, 51:504 – 519, 2015.

[17] Michael Fagan, Mohammad Maifi Hasan Khan, and Nhan Nguyen. How does this message make you feel? a study of user perspectives on software update/warning message design. *Human-centric Computing and Information Sciences*, 5(1):36, Dec 2015.

[18] Sascha Fahl, Yasemin Acar, Henning Perl, and Matthew Smith. Why Eve and Mallory (Also) Love Webmasters: A Study on the Root Causes of SSL Misconfigurations. In *Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security*, ASIA CCS '14, pages 507–512, New York, NY, USA, 2014. ACM.

[19] Thomas Faist. The volume and dynamics of international migration and transnational social spaces. 2000.

[20] F. Fischer, K. Böttinger, H. Xiao, C. Stransky, Y. Acar, M. Backes, and S. Fahl. Stack overflow considered harmful? the impact of copy amp;paste on android application security. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 121–136, May 2017.

[21] Marvin Fleischmann, Miglena Amirpur, Tillmann Grupp, Alexander Benlian, and Thomas Hess. The role of software updates in information systems continuance â an experimental study from a user perspective. *Decision Support Systems*, 83:83 – 96, 2016.

[22] Alain Forget, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, Marian Harbach, and Rahul Telang. Do or do not, there is no try: user engagement may not improve security outcomes. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 97–111, 2016.

[23] Jill J. Francis, Marie Johnston, Clare Robertson, Liz Glidewell, Vikki Entwistle, Martin P. Eccles, and Jeremy M. Grimshaw. What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology and Health*, 25(10):1229–1245, 2010.

[24] Alisa Frik, Serge Egelman, Marian Harbach, Nathan Malkin, and Eyal Peer. Better Late(r) than Never: Increasing Cyber-Security Compliance by Reducing Present Bias. In *Workshop on the Economics of Information Security (WEIS)*, page 20, Insbruck, Austria, 2018.

[25] Steve Furnell. Vulnerability management: not a patch on where we should be? *Network Security*, 2016(4):5 – 9, 2016.

[26] Jonathan Gallagher, Robin Gonzalez, and Michael E Locasto. Verifying security patches. In *Proceedings of the 2014 international workshop on privacy & security in programming*, pages 11–18. ACM, 2014.

[27] John R. Goodall, Wayne G. Lutters, and Anita Komlodi. I know my network: Collaboration and expertise in intrusion detection. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, CSCW '04, pages 342–345, New York, NY, USA, 2004. ACM.

[28] Eben M. Haber and Eser Kandogan. Security administrators: A breed apart. *Soups USM*, 2007.

[29] Julie M Haney and Wayne G Lutters. "it's scary... it's confusing... it's dull": How cybersecurity advocates overcome negative perceptions of security. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. USENIX Association, 2018.

[30] Norbert Hedderich. Three Approaches to Qualitative Content Analysis. *Global Business Languages: Vol. 2 , Article 14*, 2:162–172, 2010.

[31] Michael Hicks and Scott Nettles. Dynamic software updating. *ACM Trans. Program. Lang. Syst.*, 27(6):1049–1096, November 2005.

[32] Dennis G. Hrebec and Michael Stiber. A survey of system administrator mental models and situation awareness. *Proceedings of the 2001 ACM SIGCPR conference on Computer personnel research - SIGCPR '01*, pages 166–172, 2001.

[33] Iulia Ion, Rob Reeder, and Sunny Consolvo. "... no one can hack my mind": Comparing expert and non-expert security practices. In *SOUPS*, volume 15, pages 1–20, 2015.

[34] Eser Kandogan and Eben M Haber. Security administration tools and practices.

[35] Moazzam Khan, Zehui Bi, and John A. Copeland. Software updates as a security metric: Passive identification of update trends and effect on machine infection. In *MILCOM 2012 - 2012 IEEE Military Communications Conference*, pages 1–6. IEEE, oct 2012.

[36] Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433, jun 2004.

[37] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel von Zezschwitz. "If HTTPS Were Secure, I Wouldn't Need 2FA"-End User and Administrator Mental Models of HTTPS. *To appear in the IEEE Symposium on Security & Privacy, May 2019*, 2019.

[38] Katharina Krombholz, Wilfried Mayer, Martin Schmiedecker, and Edgar R. Weippl. "I Have No Idea What I'm Doing" - On the Usability of Deploying HTTPS. In *26th USENIX Security Symposium, USENIX Security 2017*, 2017.

[39] Mika Latimer. Trace-weighted binary comparison for software update management, 2017.

[40] Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke, Christian Stransky, Sebastian Moeller, Yasemin Acar, and Sascha Fahl. Developers Deserve Security Warnings, Too On the Effect of Integrated Security Advice on Cryptographic API Misuse. 2018.

[41] Frank Li, Lisa Rogers, Arunesh Mathur, Nathan Malkin, and Marshini Chetty. Keepers of the machines: Examining how system administrators manage software updates.

In *Proceedings of the Fifteenth USENIX Conference on Usable Privacy and Security*, SOUPS'19, pages 273–288, Berkeley, CA, USA, 2019. USENIX Association.

[42] PONEMON INSTITUTE LLC. COSTS AND CONSEQUENCES OF GAPS IN VULNERABILITY RESPONSE, 2019.

[43] Geraldine Vache Marconato, Vincent Nicomette, and Mohamed Kaâniche. Security-related vulnerability life cycle analysis. In *7th International Conference on Risk and Security of Internet and Systems (CRiSIS-2012)*, pages 1–8. IEEE Computer Society, 2012.

[44] Arunesh Mathur, Josefine Engel, Sonam Sobti, Victoria Chang, and Marshini Chetty. "they keep coming back like zombies": Improving software updating interfaces. In *SOUPS*, pages 43–58, 2016.

[45] Arunesh Mathur, Nathan Malkin, Marian Harbach, Eyal Péer, and Serge Egelman. Quantifying users' beliefs about software updates. *CoRR*, abs/1805.04594, 2018.

[46] Huoy Min Khoo and Daniel Robey. Deciding to upgrade packaged software: a comparative case study of motives, contingencies and dependencies. *European Journal of Information Systems*, 16(5):555–567, 2007.

[47] Andreas Möller, Florian Michahelles, Stefan Diewald, Luis Roalter, and Matthias Kranz. Update behavior in app markets and security implications: A case study in google play. In *Research in the Large, LARGE 3.0: 21/09/2012-21/09/2012*, pages 3–6, 2012.

[48] James Nicholson, Lynne Coventry, and Pamela Briggs. Introducing the cybersurvival task: Assessing and addressing staff beliefs about effective cyber protection. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 443–457. USENIX Association, 2018.

[49] Jon Oberheide, Evan Cooke, and Farnam Jahanian. If it ain't broke, don't fix it: Challenges and new directions for inferring the impact of software patches. In *HotOS*, 2009.

[50] Srivatsan Parthesarathy, Reiner Fink, Sean L Flynn, and Ray Sun. Software update notification, 2002. US Patent 6,353,926.

[51] Heike Pethe. *Internationale Migration hoch qualifizierter Arbeitskraefte*. DUV.

[52] Rahul Potharaju, Mizanur Rahman, and Bogdan Carbunar. A longitudinal study of google play. *IEEE Transactions on computational social systems*, 4(3):135–149, 2017.

[53] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How i learned to be secure: a census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 666–677. ACM, 2016.

[54] Elissa M Redmiles, Amelia R Malone, and Michelle L Mazurek. I think they're trying to tell me something: Advice sources and selection for digital security. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 272–288. IEEE, 2016.

[55] R. W. Reeder, I. Ion, and S. Consolvo. 152 simple steps to stay safe online: Security advice for non-tech-savvy users. *IEEE Security & Privacy*, 15(5):55–64, 2017.

[56] Armin Sarabi, Ziyun Zhu, Chaowei Xiao, Mingyan Liu, and Tudor Dumitraş. Patch me if you can: A study on the effects of individual user behavior on the end-host vulnerability state. In *International Conference on Passive and Active Network Measurement*, pages 113–125. Springer, 2017.

[57] Symantec. Internet Security Threat Report. *Network Security*, 21, 2016.

[58] Yuan Tian, Bin Liu, Weisi Dai, Lorrie Faith Cranor, and Blase Ur. Study on user's attitude and behavior towards android application update notification. *Usenix, Menlo Park, CA*, 2014.

[59] Yuan Tian, Bin Liu, Weisi Dai, Blase Ur, Patrick Tague, and Lorrie Faith Cranor. Supporting privacy-conscious app update decisions with user reviews. In *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*, pages 51–61. ACM, 2015.

[60] Christian Tiefenau, Emanuel von Zezschwitz, Maximilian Häring, Katharina Krombholz, and Matthew Smith. A usability evaluation of let's encrypt and certbot: Usable security done right. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pages 1971–1988, New York, NY, USA, 2019. ACM.

[61] Martin Ukrop, Lydia Kraus, Vashek Matyas, and Heider Ahmad Mutleq Wahsheh. Will you trust this tls certificate? perceptions of people working in it. In *35rd Annual Computer Security Applications Conference (ACSAC'2019)*. ACM, 2019.

[62] Kami Vaniea and Yasmeen Rashidi. Tales of software updates: The process of updating software. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, pages 3215–3226, 2016.

[63] Kami E Vaniea, Emilee Rader, and Rick Wash. Betrayed by updates: how negative experiences affect future security. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2671–2674. ACM, 2014.

[64] Francesco Vitale, Joanna McGrenere, Aurélien Tabard, Michel Beaudouin-Lafon, and Wendy E. Mackay. High costs and small benefits: A field study of how users experience operating system upgrades. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 4242–4253, New York, NY, USA, 2017. ACM.

[65] Rick Wash, Emilee Rader, Kami Vaniea, and Michelle Rizor. Out of the Loop: How Automated Software Updates Cause Unintended Security Consequences. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 89–104, Menlo Park, CA, 2014. USENIX Association.

[66] F.J. Wertz, Kathy Charmaz, L.M. McMullen, Ruthellen Josselson, R. Anderson, and Emalinda McSpadden. *Five Ways of Doing Qualitative Analysis: Phenomenological Psychology, Grounded Theory, Discourse Analysis, Narrative Research, and Intuitive Inquiry*. 01 2011.

Appendix

A Questionnaire

Information & Consent

Hello, we're Usable Security researchers from BLINDED and our mission is to make your challenges with system updates easier. As a first step, we need to understand your experiences and struggles with software updates in a corporate environment. We conducted interviews with seven colleagues of you and condensed interesting themes. This short questionnaire will take about 10 minutes to answer . We know that your time is precious, which is why every tenth participant gets a 3D-print of a model of her/his choice (max. 3x3x3cm and a reasonable model). If you are interested in this form of compensation just leave us your email address in the commentary field at the end. This email address will be stored separately from your answers and will only be used to communicate about your compensation. Please read all questions and instructions carefully. All of your answers will be checked, and your survey may be rejected in the case of inconsistent answers. Your data will be collected and processed in anonymized form, so that no connection to your person can be made. You can stop participating in this study at any time. If you have any questions please contact us via email.

*1. I have read and understood the information provided above and consent to take part in this study.

- I consent
- I do not consent

Demographics & General

*2. How old are you?
Text-input field
3. What ist your gender?
Text-input field
4. In what country do you work?
Text-input field
*5. For how many years have you worked as a professional system administrator?
Text-input field

Job information

All of the questions on this page refer to a specific company. If you currently work as an administrator, please answer these questions about your current company. Instead, if you do not currently work as an administrator, please answer these questions about the last company at which you worked as an administrator.

6. Is this company an IT company (software/hardware development, hosting, ISP, ...)?

- Yes
- No
- Other (please specify): *Text-input field*

7. Which of the following statements best describes your role in this company?

- My primary responsibility was system administration
- My primary responsibility was not system administration, but I spent at least 20% of my time on system administration
- My primary responsibility was not system administration, but I spent between 1% and 19% of my time on system administration
- I did not perform system administration at that company

8. In a few words, what would you consider as your main task in the company you are working at?
Text-input field
9. What is your main task as a system administrator? If it is the same as in the previous answer, please answer: same.
Text-input field
10. What kind of systems do you administer?

- Clients (e.g. workstations)
- Servers
- Mobile Clients (eg. tablet, smartphone)
- Other (please specify): *Text-input field*

* 11. How big is the company you work at as a system administrator?

- less than 10 employees
- up to 50 employees
- up to 250 employees
- more than 250 employees

12. Do you work in a team?

- Yes, as a team leader
- Yes, as a team member
- No
- Other (please specify): *Text-input field*

*13. What kind of job related education did you receive? (e.g. training, certificate, university)
Text-input field
14. Which of the following statements best describes the security-related training you have received concerning system administration?

- I received security-related training for system administration at that company
- I did not receive security-related training for system administration at that company, but I have received such training at a previous company or school
- I have never received security-related training for system administration

Update Process

Please be reminded that we do not collect or store identifying information. In the following we are interested in your honest opinion.
15. Among all software updates you install for operating systems or any other software running on systems, approximately what percentage do you estimate are **security** updates?
Slider [0-100]
16. Within your job as a system administrator, how much effort does it take you to keep the software on your systems up-to-date?
7-point Likert scale from "1 - Nearly none" to "7 - Nearly all my capacity"
17. What pre-deployment steps do you take before installing an update on a live system?

- We install it on a test system.
- We install it on a small number of production systems before deploying it to all systems or to everyone.
- We install it directly on all production systems.
- Other (please specify): *Text-input field*

18. What is the share of security related updates in relation to all updates (in %)?
Slider [0-100]
19. Which of the following statements best describe the update process in the company?

- There is a written document, that formally describes the steps in the update process.
- There is no written document but an informal guideline that is followed in the update process.
- There is no defined update process.

20. What is the typical time-span between the release of an update to the installation in a normal update process?
Text-input field
*21. Please indicate how often the following situations occur:
Table of the following questions, with a 6-point Likert scale from "1 - Never" to "5 - Always" and the option "Not sure", per question.

- I feel that I am not sufficiently trained as an administrator.
- I think of work- related consequences when doing tasks that have, in case of a failure, an impact on my company (e.g. downtime of a service that everyone uses).
- I feel personally responsible for keeping the software on my systems up-to-date.

22. Please indicate how often the following situations occur:
Table of the following questions, with a 6-point Likert scale from "1 - Never" to "5 - Always" and the option "Not sure", per question.

- Stability considerations hinder the installation of an update.
- Risk considerations hinder the installation of an update.
- Performance considerations hinder the installation of an update.
- Priority/time considerations hinder the installation of an update.

- Software updates are prevented because of other software (e.g. dependencies).

23. Please indicate how often the following situations occur:
Table of the following questions, with a 6-point Likert scale from "1 - Never" to "5 - Always" and the option "Not sure", per question.

- System stability considerations are irrelevant to the installation of an update.
- The risk of breaking dependencies hinder the installation of an update.
- A patch that is known to introduce errors hinder the installation of an update.
- Downtimes caused by the update process hinder the installation of an update.
- Lack of information about the changes an update introduced hinder the installation of an update
- Lack of education and knowledge hinder the installation of an update.

24. Please indicate how much you would agree/disagree with the statements.
Table of the following questions, with a 7-point Likert scale from "1 - Strongly disagree" to "4 - Undecided" to "7 - Strongly agree", per question.

- Deploying security updates in a timely manner is important.
- Post-installation problems in a live system are only a minor concern because they don't happen frequently.
- Users often install software without the knowledge of the administrator.

25. Who makes the decision whether to update or not?

- My team.
- Myself.
- My colleague(s).
- My supervisor.
- None of the above, please specify: *Text-input field*

26. Please indicate how often the following situations occur:
Table of the following questions, with a 6-point Likert scale from "1 - Never" to "5 - Always" and the option "Not sure", per question.

- I feel sufficiently trained as an administrator.
- I can oversee the impact an update would have on our systems.
- I can oversee the impact of a failed update on our system.
- I can oversee the security impact of updates on our systems.

Source and Tools

*27. What sources do you use to get information about current system updates?

- Online publications/news (e.g. cnet.com, Hacker News, heise,...)
- Update management software
- (Software) Publisher newsletters
- External services (e.g. a company that is contracted to inform you)
- Mailing lists
- My users
- Other (please specify): *Text-input field*

*28. What ist your main source to get information about current system updates?

- Online publications/news (e.g. cnet.com, Hacker News, heise, ...)
- Update management software
- (Software) Publisher newsletters
- External services (e.g. a company that is contracted to inform you)
- Mailing lists
- My users
- Other (please specify): *Text-input field*

29. Please explain your previous answer:
Text-input field

Thank you!

30. What do you think are the biggest obstacles in the update process?

Text-input field

31. Thank you for your participation! If you have any further comments for us: Don't hesitate to use the textbox!

Text-input field

32 . If you are interested in the 3D model print just leave your email in this field. We will only use this mail for the communication and will not link it to your answers.

Text-input field

B Interview Guidelines

Questions to explore

1. What does the update process look like?
2. What obstacles are there?
3. Who is involved?
4. What is his/her personal experience and assessment?

Introduction

1. How long has he/she done the job? What is the training? What is he/she doing on a daily basis?
2. What are the systems?
3. Does he/she work in a team?
4. What is the scope of his/her actions?
5. What tools are used?

General update process (or a specific update story)

1. How does he/she come in contact with updates?
2. What is the time frame and the process?
3. What tools are used?
4. Who is involved?
5. Where does the information come from?

(Optional) A second story

1. How does he/she come in contact with updates?
2. What is the time frame and the process?
3. What are the tools?
4. Who is involved?
5. Where does the information come from?

End

1. Do they have a fixed update policy?
2. Are there any feelings connected to new updates or the installation?
3. Is he/she aware of potential impacts of not installed update/failures of the installation? (Are there stories?)
4. Are there wishes concerning the process/tools?
5. Questionnaire

C Demographics

Survey demographic data

n	**67**	
Gender	1	Female
	58	Male
	3	Other
	5	Not specified
Location	19	North America
	41	Europe
	7	Rest of the world
Age	**22 – 55**	
Statistics	$md = 34, mn = 34.5, sd = 7.8$	
Experience	**0.1– 30.0** years	
Statistics	$md = 10.0, mn = 11.1, sd = 7.0$	
Company	34	IT-related
	29	Non IT-related
	4	Other
Company Size	4	≤ 10
	15	$10 < x \leq 50$
	15	$50 < x \leq 250$
	33	> 250
Role	50	Full-time admin
	11	Not primary, but $> 20\%$ of time
	6	Not primary, but $< 20\%$ of time
Administered Systems	28	Clients
	63	Servers
	14	Mobile
	13	Other

Table 6: Demographic data from the online survey.

Interview demographic data

Pseudonym	Gender	Position/Task	Age	Exp. (Years)	Team size	Supervised Machines
Markus	M	Administrator	25–35	6	7	300–350 clients, 150 virtual servers
Lorenz	M	Update management	25–35	2	n/a	5 servers
Cyril	M	Administrator	25–35	6	15	10,000 virtual, ca. 100 physical
Milan	M	Help desk	25–35	2.5	12	600 clients, number of servers
Zelko	M	Administrator	25–35	10	2	16 physical, 35 virtual, 80 clients
Alexander	M	Update management	> 35	23	5	26 physical, 170 instances
Julian	M	Management	> 35	29	20	n/a

Table 7: Interview participants.

An investigation of phishing awareness and education over time: When and how to best remind users

Benjamin Reinheimer*, Lukas Aldag*, Peter Mayer*, Mattia Mossano*, Reyhan Duezguen*,
Bettina Lofthouse†, Tatiana von Landesberger°, Melanie Volkamer*
* *SECUSO - Security, Usability, Society, Karlsruhe Institute of Technology*
† *Landesamt für Geoinformation und Landesvermessung Niedersachsen*
° *Interactive Graphics Systems Group, Technische Universität Darmstadt*
* *firstname.lastname@kit.edu,* ° *tatiana.von.landesberger@gris.tu-darmstadt.de*
† *Bettina.Lofthouse@lgln.niedersachsen.de*

Abstract

Security awareness and education programmes are rolled out in more and more organisations. However, their effectiveness over time and, correspondingly, appropriate intervals to remind users' awareness and knowledge are an open question. In an attempt to address this open question, we present a field investigation in a German organisation from the public administration sector. With overall 409 employees, we evaluated (a) the effectiveness of their newly deployed security awareness and education programme in the phishing context over time and (b) the effectiveness of four different reminder measures – administered after the initial effect had worn off to a degree that no significant improvement to before its deployment was detected anymore. We find a significantly improved performance of correctly identifying phishing and legitimate emails directly after and four months after the programme's deployment. This was not the case anymore after six months, indicating that reminding users after half a year is recommended. The investigation of the reminder measures indicates that measures based on videos and interactive examples perform best, lasting for at least another six months.

1 Introduction

Maintaining information security is an important challenge for organisations, and also for governmental and public administration sectors. The so-called German national IT Planning Council [71] requires German organisations in the public administration sector to implement information security management systems (ISMS). One of the goals of such ISMSs is to enhance employees' information security awareness and knowledge. A common approach to satisfying this requirement is to roll out *security awareness and education programmes*. They typically raise general security awareness (e.g., everyone can potentially become a victim, the technological protection mechanisms need users' support, potential consequences of successful attacks) and convey knowledge about information security (including how to identify various attacks, how to reduce one's risks of becoming a victim of cyber attackers, and who shall be contacted in case of questions and incidents). These programmes may include *security awareness and education measures* that cover different aspects and/or topics using different media types, such as self-learning measures [63], e-learning platforms [9], on-site tutorials [102], or games [94]. Although such measures are widely deployed, an evaluation of their effectiveness related to their ability to enhance employees' information security skills over an extended time period is often missing. This, however, is of the essence: if employees are never, or only rarely, confronted with attacks that are included in a security awareness and education programme, the acquired awareness and knowledge might dissipate over time, as is the case with any other awareness and knowledge programmes. While waning of awareness and dwindling of knowledge is to be expected, it poses a problem to the maintenance of organisational information security. Therefore, it is crucial to know: (a) *when* awareness and knowledge levels should be renewed, i.e., how long the effect of a security awareness and education programme can be expected to last, and (b) *which* type of measures are best suited to restore users' awareness and knowledge.

Researchers could show the effectiveness of security awareness and education measures directly after roll-out [4, 64, 94, 98, 102, 106, 111, 119] and that the significant improvements endured over different spans of time. [22, 60, 116]. However, what is missing, is an insight into how long the impact of a security awareness and education measure lasts and how awareness and knowledge should be renewed. In order to gain these insights, we adopted and evaluated the phishing aware-

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

ness and education measure from [84]. We customised it for a German organisation from the public administration sector: a German State Office for Geoinformation and State Survey (SOGSS). We replaced example messages with those more suited to SOGSS and by removing irrelevant content. The content was presented face-to-face in on-site tutorials. A '*train the instructor*' approach was used, which involved eleven instructors being trained by the Chief Information Security Manager. The participation in the tutorial was mandatory for all employees.

We evaluated employees' skills in distinguishing phishing emails from legitimate emails at several points in time. First, data was collected just before and directly after the on-site tutorials. To study how long the effect lasted, we collected retention data four months after employees had participated in the tutorial. We were prepared to continue doing so every second month as long as we continued to see a significant enhancement of participants' skills.

Our first contribution: we systematically measured the retention for the tutorial. Compared to previous studies we measured until the significant improvement wore off plus another measure after this point in time. The impact wore off after six months.

Our second contribution: we developed suitable reminder measures to replenish the employees' phishing awareness and knowledge after receiving not significant results for the first time after the tutorial. We developed four different ones (three presenting the content using text, one using video and one using interactive email examples). The success of the reminder measures was evaluated right after their deployment, and again after six months.

Our third contribution: we accompanied an organization for a total of twelve months to both check for the effectiveness of the tutorial and the reminder measures. The awareness and knowledge levels of participants having either seen the video measure or the interactive examples after six months were still significantly higher twelve months after the initial tutorial.

As a consequence, SOGSSs decided to use the video and the interactive examples measures and to distribute these to all employees on a regular basis at six month intervals.

2 Related Work

This section commences by providing phishing definitions from the research literature. Related work is discussed next with regard to different types of security interventions, different study designs used to evaluate these interventions, and different types of tested users groups. Finally research into the impact of phishing security awareness and education measures over time are discussed.

Phishing Definitions: there are many different definitions of phishing in the literature. Correspondingly, researchers' focus is different: (1) those who focus on phishers who want their victims to provide sensitive information (e.g. passwords, bank details) using an authentic-looking phishing web page [1–4,7,16–19,21,22,30,33,35,38,51,54,57,58,64,68–70,75–77,80,82,88,91,93,94,96,98,104,105,108,113–115,117], or (2) those who focus on phishers who distribute malware when recipients click on links in messages or open attached files [1,6,8,9,12–15,20,23,24,31,32,34,41,43,45,48–50,55,56,59–62,74,79,81,85–87,90,92,99,102,103,106,107,109–112,116,119]. Because it is safer to check the URL before clicking on it, instead of only checking the URL after opening the web page people are redirected to, *we focus on the second one*.

Types of Interventions[1]*:* various studies evaluating different types of interventions to test their effectiveness exist. Researchers evaluated a range of tools that are supposed to provide further support (e.g. additional security indicators or displaying existing security indicators in different ways) [2,5,18,29,36,40,66,68,69,72,88,96,114,115,117,118]. Different evaluated security awareness and education measures are a range of videos [46,107], games [10,11,21,22,64,94], various on-site instructor based tutorials [98,102,116] – *as studied in the research presented in this paper* – and a multitude of text-based measures [4,47,60,65,84,92,98,102,104,110,116,119]. Additionally, there is research evaluating users' skills in phishing detection without any interventions [6,7,12,14,15,26,30,31,34,37,39,45,48–51,53,54,59,73,74,80,81,85–87,89,90,95,100,103,109] (e.g. to understand decision making, to identify a baseline, or to motivate further research).

Study Designs: various types of lab studies have been employed, some with a cover story [4,14,15,35,36,61,69,86,91] and others without one [7,11,37,40,98,103,116], i.e., having security as participants' primary goal by telling participants the goal of the user study. A number of remote studies have been carried out, including various types of online surveys, with phishing messages sent to the study participants own email accounts (not study specific) [30,37,40,59,81,87,102,104], as well as to remotely accessible study-specific accounts [88,90,109,110,119]. Surveys include those that (1) show screenshots to be judged either as phishes or legitimate [54,70,100] *as we did in our study*. In some cases real phishes were used; others used examples created by the researchers; and (2) online surveys asking general questions such as the definition of phishing and the existing attack types [20,51,52,79].

Types of User Groups: studies have targeted different user groups, i.e. mixed groups on a variety of panels without deliberately isolating specific kinds of participants [16,35,40,54,70], employees [26,43,49], or university faculty or students [6,11,14,15,31,53,86,87]. *Our target users were employees of a governmental organisation.*

Forgetting Rate of Different Age Groups: [44] evaluated the ability to recall visual cues after 20-30 minutes and 75

[1] Interventions can be tools or security awareness and education measures.

days. They did not find any age differences in the recall ability of these visual cues. [101] examined the recall ability of verbal cues after 1 and 62 days with different age groups. They conclude that the encoding of the information in the beginning is slower, but the rate of forgetting is comparable afterwards. *Retention Periods of IT-Related Training:* while most of the previously mentioned phishing studies evaluated the impact of the their interventions straight after roll-out, a few also evaluated the effect after some time had elapsed. These mainly showed that the effect still held and did not systematically determine for how long the effect was still evident. These retention studies were mainly conducted in the context of security awareness and education measures. In [28, 60], retention was evaluated after approximately a month. In [116], retention was evaluated after 45 days. In [107], the retention was evaluated after 8 weeks. In [22], retention was evaluated after 5 months. All showed that the effect was still perceptible but was often no longer significant. [78] examined the ability to judge insecure password-related behaviour. The participants received awareness-raising materials and were tested again after 6 months. The participants were able to retain significant knowledge. *In our case, we study exactly how long the effect lasts.*

3 Use Case: Organisation Description

A State Office for Geoinformation and State Survey (SOGSS) is a public administration sector organisation. Its core activities relate to land register and real estate cadastre. Overall, SOGSS has about 2200 employees, 83% of whom have a technical background in either surveying, geodesy, geoinformatics or other related fields, such as photogrammetry. 60% of the employees are over 50 years of age. Only 14% are between 25 and 35 years of age. 40% of all employees are female, 60% are male. All employees use passwords and their SOGSS email account on a daily basis. Email communication with colleagues, citizens, and partners from business, science and other authorities is indispensable to employees.

SOGSS has nine regional head offices, a central operational office and a central head office, each with several departments and each in a different city. Like all organisations in the German public administration sector, SOGSS is required by the national IT Planning Council [71] to implement an information security management system (ISMS). Thus, SOGSS established the position of a chief information security manager (CISM) and the role of *'person of contact for information security concerns'* (PoC-InfoSec) was introduced as organisational interface between the CISM and the local offices. The managers of all ten human resources and administration departments (from nine regional head offices and the central operational office), as well as the manager of the central head office, perform this role. Furthermore, SOGSS decided to develop a security awareness & education programme containing one mandatory on-site tutorial, which was delivered to all employees. Most tutorial sessions were held in October 2018.

4 Security Awareness and Education Measure

We describe the design decisions made for the mandatory measure rolled out in 2018. Afterwards, we describe their structure and content. Finally, we introduce the reminder measures.

4.1 Design Decisions

The organisation decided to use on-site tutorials instead of other delivery measures, such as web-based training, for two reasons. Firstly, on-site tutorials are common practice at SOGSS and therefore employees' acceptance of such tutorials was expected to be higher than for other formats. Secondly, the search for a suitable third-party web-based training, and the obligatory call for tenders, would have taken too long. Due to room size constraints, it was decided to deliver training to forty participants in each tutorial. Furthermore, based on the experiences from other on-site tutorials, it was decided that the tutorial should last three to four hours. Thus, the content had to fit into this allotted time.

A decision was made to adopt a 'train the instructor' approach, instead of having the CISM delivering all the tutorials. The 'train the instructor' approach was chosen since it represents a resource-efficient way to deliver tutorials to a large number of employees over a reasonably short period of time. The eleven PoC-InfoSec were trained by the CISM. To support them, a Power Point presentation was developed in two versions: 1) an instructor version, supplemented with explanations and instructions on how to facilitate audience interaction, and 2) the actual presentation to be used during the tutorial.

4.2 Content Overview

It was decided that the tutorial would address the following three topics, as threat reports and the organisation's experiences identified these as the most relevant ones: *Topic-1*: General security awareness, *Topic-2*: Phishing, and *Topic-3*: Password best practice. While the first and third parts were developed by the CISM from scratch, the second part was an adaptation of the awareness and education measure reported by [84]. Correspondingly, the focus of our investigation was on Topic-2[2]. Its content will be described in more detail in the following two subsections.

[2]It could be argued that we should only have addressed this one topic in the tutorial if only one was going to be evaluated. However, this would not have been sufficient to be compliant with the ISMS. We could indeed have conducted two different tutorials, but this would have been much less efficient and there might have still been a bias because the study ran over several months and therefore everyone would have had to participate in the second tutorial in the same period of time, too. Thus, because we wanted to able to conduct our evaluation in the field, we had to accept this trade-off.

With respect to the content of the tutorials, *Topic 1* provided a general introduction to the information security topic and information about how security incidents have to be dealt with in SOGSS. The organisation's threat statistics and typical threat vectors such as email (being one of the most common in public administration) were introduced. Where suitable, examples of anonymised in-house IT-security incidents were provided to demonstrate vulnerabilities when using the Internet. The content of *Topic 3* made employees aware of the risks of using weak passwords and introduced methods of building strong passwords. This part ended with a short interactive quiz, where examples were used to assess the respective password strength.

4.3 Topic 2 Content: Phishing

We customised a security awareness and education measure, which was developed and evaluated[3] by Neumann *et al.* [84]. This material is very well suited: it has been evaluated in an organisational context attesting its effectiveness and it is freely available in German. The original content was prepared for self study use, i.e., reading a pdf or integrated it into an e-learning platform. Thus, it needed to be customised. The content of the measure had two parts:

Part-A provided general information about phishing, including: (a) why everyone can potentially become a victim, (b) that phishers don't just use email messages but any type of message, (c) that there are various types of phishing messages (including those asking for sensitive information, those including dangerous links, and those with dangerous files attached), (d) what the potential consequences of falling for a phishing attack are, and (e) the recommendation to delete phishing messages and to search for further information when the person is uncertain.

Part-B: commenced by explaining that a number of plausibility checks should be carried out, including checking the language, the style, and the sender information. Afterwards, the focus was on phishing messages which look plausible, at first glance, but which actually contain potentially dangerous links and/or attached files. Legitimate messages might be used as template, with the sender address being spoofed, and the URL behind a link and/or the attachment being replaced. *First*, employees were shown how to check whether an embedded link was dangerous to click on. This covers several attack types that phishers use to trick people (for more information see Appendix B). Furthermore, it was explained that these tricks are combined by phishers and that the presence of `https` is an unreliable indicator of trustworthiness. *Second,* they were shown how to identify dangerous file(s) and told which tricks phishers apply to trick people, incl. using two file extensions or unknown extensions.

[3]Note that effectiveness was evaluated straight after rolling-out the measure.

Both parts contained several example messages. All examples were synthetic as we came up with our own brands. There were example messages to illustrate various aspects of each attack type, as well as example messages to practice what was learned. Various misconceptions (such as that people tend to classify long URLs as Phishing links) identified from the literature were addressed throughout the measure, too.

Customization. Before customization the measure was usable only for self-studies: it is full text PDF documents that contains full sentences. This was not suitable for on-site tutorials, thus it had to be adopted to be used in on-site tutorials with Power Point presentations. Furthermore, *Part-A* (b), i.e. the information that phishers may use various message media, was not addressed in SOGSS's version. The use of email on mobile devices and/or social media is severely restricted to a small number of SOGSS's employees. Hence the tutorial focused on desktop application emails, as they are the only relevant target for phishers at SOGSS.

All examples from the original measure were replaced by ones more related to the employees' daily work. For example, where possible, anonymised examples from reported phishing emails were used. The PoC-InfoSecs were asked to show one example after the other. For each example, they were required to ask the audience whether it was a phish or not and one person from the audience was to justify the answer. Afterwards, the PoC-InfoSecs were supposed to explain the correct answer and comment on the answers given by the audience. This approach was used to attract the audience's attention. Finally, a summary of the most important findings and recommendations to check for was included at the end of the phishing part of the Power Point presentation.

4.4 Reminder Measures

Once the initial effect of the security awareness and education programme on-site tutorial has worn off, a reminder measure should be distributed to remind users of the information in the programme. Due to the lack of research into these measures, the goals of this research project was first to identify appropriate ones (i.e. by evaluating several). Correspondingly the reminder measures described in this section are currently – unlike the on-site tutorials– not part of the security awareness and education programme. So far, only participants in the corresponding study groups saw the reminder measures.

To the best of our knowledge, this aspect has not been studied in terms of which kind of presentation is most appropriate in the information security field. Four different types of reminder measures were developed and then evaluated: a text measure, a video measure, a interactive examples measure, and a short text measure. To inform the development of the reminder measures, we wanted to satisfy two requirements. *Firstly*, the measure must stand for itself. Apart from the shown measure, no further references or information should

be necessary In particular, no instructions or introductions from another person should be required. *Secondly*, the measure content should match the on-site tutorial, i.e. it should not contain any new content that does not represent previously-learned knowledge from the on-site tutorial. The text and and interactive examples measures contain exactly the same content (part A+B of the on-site tutorial). The video also covers this same content, but presents it as a story in order to make it more appealing. The short text cuts down the content to only include part-B and minimised descriptions of the attacks and defence strategies. In detail, the four reminder measures are:

Text: this measure is depicted in Figure 6 in the Appendix. It is a text, in German, with six figures. These visualise explanations such as the structure of URLs and that the actual linked URL is displayed in a status bar or a tooltip.

Video measure: this measure presents the same content as the previous one but relies mostly on visual explanations and narration, instead of text. Figure 7 in the Appendix gives an example of the video measure.

Interactive examples: this measure uses an interactive presentation. The content is presented as two interactive examples of phishing emails (see Figure 8 in the Appendix). Each of the emails has multiple interactivity-points marked with red dots, which reveals information about the respective part of the email when hovered over. In order to finish this measure, the trainee has to click at least once on each area.

Short Text: this measure represents a text-based measure with curtailed content compared to the previous measures. It contains only Part-B, i.e. it focuses only on the recommendations for detecting phishing messages (see Figure 9 in the Appendix).

5 Methodology

We first introduce the research questions and the hypotheses. Then, we discuss our study design decisions. Afterwards, we provide details about the study, i.e. recruitment, group assignment, used email examples, and actual study procedure as well as ethical considerations.

5.1 Research Questions

We want to answer three research questions. The first one is: **How long does the effect of the on-site tutorial last, i.e. when should the gained awareness and knowledge be reminded?**

The following pre-condition needs to hold: the measure significantly strengthens participants' skills in distinguishing between phishing emails and legitimate emails, straight after the on-site tutorial. Therefore we phrase the following hypothesis for this pre-condition:

$\boldsymbol{H}_{M_{pre}-M_{0M}}$:*Participants have an enhanced skill in terms of distinguishing between phishing and legitimate emails directly after the on-site tutorial, i.e. 0 months after it, as compared to before participating in the on-site tutorial.*

In order to investigate the effectiveness[4] of the on-site tutorial over time, we formulate the following hypothesis for the continued testing:

$\boldsymbol{H}_{M_{pre}-M_{\Delta\tilde{\imath}M}}$:*Participants have an enhanced skill in terms of distinguishing between phishing and legitimate emails after* $\Delta t = 4 + 2i$ *months, where* $i \in \{0,1,2,3,4\}$*, as compared to before participating in the on-site tutorial.*

We decided to start the follow-up evaluations after four months due to results from related work in the phishing context [22, 60, 116] reporting significant effects from security awareness and education measures lasting from 45 days to up to five months. Therefore either 45 days or five months after the on-site tutorial should be chosen for the first follow-up evaluation. However, conducting the first follow-up evaluation after 5 months would have increased the likelihood that the effect of the on-site tutorial had decreased below a significant improvement. Therefore, we decided on a more conservative approach, i.e. to start the follow-up evaluations earlier. Starting to early would have required too many participant groups due to the between-subjects approach. Therefore, we commenced the follow-up evaluation after four months, since it represented the best trade-off and allowed a meaningful study design despite the limited overall number of participants available. Note that the scheduled maximum duration of the evaluation was set to twelve months due to legislative constraints[5] of SOGSS.

The second research question is: **Which of the four reminder measures performs best – straight after its roll-out?**

The reminder measures were distributed as soon as $H_{M_{pre}-M_{\Delta\tilde{\imath}M}}$ no longer held, i.e. after six months[6].

To study this second research question, the following pre-condition needs to hold: the potentially best reminder measure needs to significantly strengthen participants' skills in distinguishing between phishing emails and legitimate emails - right after the reminder measure was rolled out. Correspondingly, we use the following hypothesis:

[4] When we talk about effectiveness from now on, we only talk about the effectiveness of the phishing part of the on-site tutorial.

[5] The next security awareness and education measure is required after one year due to compliance reasons

[6] We analysed the data as soon as they arrived to distribute the reminder measures as soon as the performance is no longer significantly better as compared to before they participated in the on-site tutorial. As described in Section D.1, our evaluation yields the result $\Delta\tilde{\imath}$=6, i.e. the measurement after six months did not detect a significantly enhanced skill in terms of distinguishing between phishing and legitimate emails. We provide this information here, to facilitate description of the remaining two research questions.

$\boldsymbol{H}_{M_{pre}-M_{Reminder_x}-6M}$: *Participants have an enhanced skill to distinguish between phishing and legitimate emails directly after the distribution of reminder measure* $Reminder_x \in \{Text, Video, Interactive\ examples, Short\ Text\}$*, as compared to before participating in the on-site tutorial.*

For those reminder measures for which this pre-condition holds, we compare the measured effects, to see one is more superior than any others.

The third research question is: **How long does the effect of reminder measures last?** To measure this, we evaluated the performances of those reminder measures for which the precondition from the second research question holds in a six month retention (i.e. 12 months after the roll-out of the on-site tutorial). The corresponding hypothesis is:

$\boldsymbol{H}_{M_{pre}-M_{Reminder_y}-12M}$: *Participants still have an enhanced skill to distinguish between phishing and legitimate emails six months after distribution of the measure for which* $\mathbf{H}_{M_{pre}-M_{Reminder_x}-6M}$ *holds, compared to before participating in the on-site tutorial.*

This research question has two pre-conditions: (1) participating twice (once after six months and again after twelve months) should not have a significant impact on the measured effect (2) other events in the organisation should not lead to a significant improvement again.

5.2 Design Decisions for Study Design

The selection of the study type was driven by the need to gain a high participation rate and outcome quality of the study within the SOGSS environment. *We decided on a study design that would enable remote participation.* This allowed us to reach more participants than a lab study. Moreover, remote participation was less time consuming for the SOGSS' employees (being distributed over several locations and cities) and it was less likely to interrupt their work as they could participate during the allotted time frame at their convenience.

There are two main ways of conducting the evaluation of the on-site tutorial with remote participation: a multiple choice test, e.g. asking to define phishing and name attack types; or evaluating participants' actual skill to identify phishing and legitimate emails. Multiple choice tests would have provided very little information about the enhancement of employees' skills in terms of distinguishing phishing and legitimate emails from each other as we would not be able to determine whether emails' other properties (e.g. the deceptively trustworthy design or the sender name) may have led them to judge phishing emails as legitimate (or the other way round). *We decided to employ the second option which asked participants to classify emails as either phishing or legitimate.*

There are two main ways to evaluate participants' actual skills in terms of distinguishing between phishing and legitimate emails: (i) sending them phishing emails (with or without announcing the fact that phishing emails will be sent) and then e.g. asking them to report phishing emails; or (ii) displaying a set of emails in a survey style and asking them to decide which were phish and which were legitimate. The first option might be considered to be closer to assessing real behaviour but such behaviour might well be influenced by the fact that participants know they will receive phishing emails, but they don't know when they will receive them. Their daily business remains their main task, not security, which is ecologically valid. In contrast, the second approach measures skills in a 'best case' scenario as security is the participants' primary goal in this case, and this is unrealistic.

However, the first approach – actually sending phishing emails – was not feasible in our study setting at SOGSS. Some reasons were: the research goal of assessing gained awareness and knowledge over time required us to evaluate all data in a very short time span for each group (i.e., for each time Δt_i). To ensure that the received phishing emails amount was realistic, we could not have sent more than one test phishing email per day – but also not send one every day. As the goal was to evaluate skills for all five attack types, this evaluation have taken too long. Moreover, sending phishing emails to employees of German organisations would have required extra permissions e.g. from work councils. There are also some general issues with this approach regarding data collection quality (as e.g. discussed by [83]). For these reasons, *we employed the first approach – displaying email screenshots in an online survey*, where participants assessed all emails in one session and decided, for each, whether it was a phish or legitimate. This allowed us to evaluate all attack types in the shortest possible amount of time.

5.3 Recruitment and Group Assignment

Due to SOGSS organisational requirements, participation in the on-site tutorial was mandatory, but participating in the evaluation was voluntary. The information about the evaluation and a corresponding link to the survey was emailed to employees by SOGSS's CISM. Every group only got one survey sent to them (we did not reuse groups/participants for other groups). Once they received the notification to take part in the survey, they had a week to do so. A reminder email was sent that emphasised the importance of personal participation due to the cyber security situation. This was sent to everyone in the group as it was not known who had actually participated as yet. We collected data for two weeks.

We planned for eleven groups (see Figure 1): seven retention groups and four reminder groups. For the assignment, we considered the fact that besides the October on-site tutorials, a few were scheduled for end of 2018 / beginning of

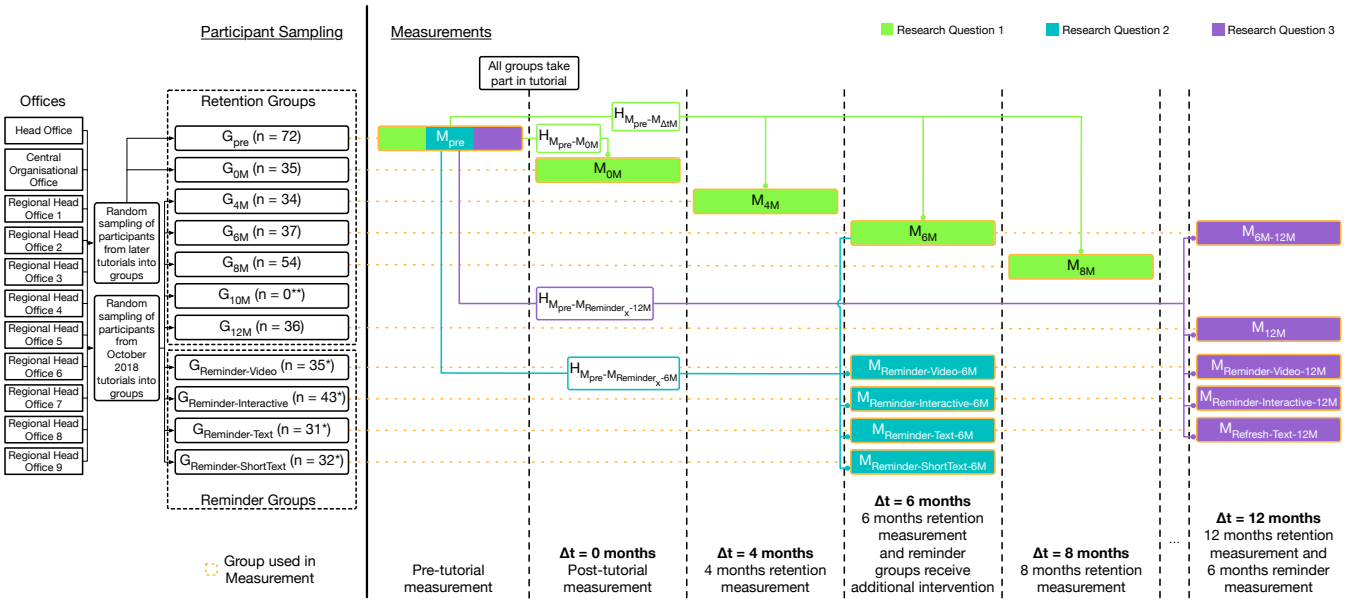


Figure 1: Overview of the participant sampling into the ten study groups, the measurements taken for each of the groups, and the hypotheses in relation to the measurements used in the corresponding statistical tests. Group sizes marked with an '*' are the overall group size, not those participants used for the linked analyses (see Section 6.2). Group G_{10M} (marked with '**') was originally scheduled and participants assigned to it, but due to the lack of significant effects in G_{6M} and G_{8M} the group was never tested. The • indicates for each hypothesis in which measurement the participants are expected to perform better.

2019 to enable those employees who could not attend any of the October options to participate. To prevent the introduction of variance into the later measurements, all participants from these later tutorials were randomly assigned to either G_{pre} or G_{0M}. All other participants were randomly assigned to one of the retention or reminder groups. Thereby, we ensured that participants from the same office were equally distributed among all groups. Thus, participants from each office were represented in each group. This was important as employees were taught by different instructors at different locations. Since it was a requirement of SOGSS, we did not collect any demographic data in the study, and no other parameters were used for the sampling. Every measurement consisted of unique individuals except for the reminder measurements at month 12 and the measurement at 12 months for measuring participants twice. The numbers in Figure 1 reflect these unique participants. The linked participants are a part of the full reminder measurements that we could link based on a code they entered for the 6-month measurement and the 12-month measurement. We will add this description to section 5 to make the distinction clearer and earlier. For $G_{Reminder-Interactive}$, $G_{Reminder-Text}$, $G_{Reminder-Video}$, and G_{6M}, the measurement after 12 months is longitudinal.Therefore, a subset of the overall participants from e.g. G_Reminder-Interactive after 6 months build the corresponding group after 12 months. We denote those participants in each of the four subsets as linked. The overall participants in each group after 6 months are called unmatched.

Note that all assignments were implemented by the CISM for data protection reasons.

5.4 Email Screenshots

As outlined in Section 5.2, the participants' performance was measured using screenshots of emails, each of which has to be classified as phish or legitimate. In an ideal setting, one would have evaluated all combinations, relevant to SOGSS, of five attack types (see Section 4.3), operating systems (including at least Windows and macOS), email clients (including at least Outlook, Thunderbird, Apple Mail), and web browsers (including at least Firefox, Chrome, Safari). However, this would have resulted in an infeasibly large number[7] of screenshots that each participant would have needed to rate, leading to fatigue effects or abandonment. Therefore, we selected a representative subset of all possible screenshots. The goal was to cover all attack types with a variability within the remaining characteristics. (This resulted in ten different phishing email screenshots. This set of phishing email screenshots was complemented by an equal set of ten legitimate email screen-

[7] At least $180 = 5 \times 2 \times 3 \times 3 \times 2$, assuming there are as many phishes as legitimate emails.

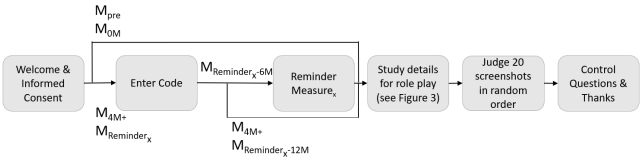


Figure 2: The survey process and measurement times.

shots[8]. We decided to use 'https' for all URLs (both phishing and legitimate)[9]. The screenshots with a dangerous link were generated in a way that the mouse was already situated next to this link, i.e. the actual URL was displayed depending on the environment in the tooltip or the status bar. The screenshots of the phishing emails used in our evaluation are provided in the supplementary material (see Figure 11 to Figure 15 in the appendix). The following phishing emails were used: one easily to detected phishing email with implausible email content: a dubious job offer, or an offer of unrealistic amounts of money. Eight phishing emails with plausible content, but including dangerous links: four emails presenting the URL in the toolbar and four in the status bar; two with a mismatch and one with a faked tooltip; URLs were either arbitrary, had the domain name as subdomain, a typo in the domain, or the domain name was extended. One phishing email with plausible email content, including a dangerous attachment.

5.5 Survey Design

The survey was designed to match the corporate design of the organisation and it was implemented on the SoSci Survey platform[10]. The overall structure is depicted in Figure 2.

For all measurements, after opening the corresponding SoSci Survey link, participants received *explanations* about the evaluation and the cooperation between their employer (SOGSS) and our university for this evaluation. We tried to mitigate external factors by also explaining that participants ought not to use external sources (e.g. web search) and to concentrate on the survey during the session. We tried to increase participation by highlighting the possibility of using this evaluation as a self-assessment of their own skills. Moreover, we assured the participants that they did not need to fear consequences if they performed poorly. Thereafter, the survey's structure was explained to them. Participants were informed that they could terminate their participation at any time without providing any reason and that, in this case, their data would not be used (*informed consent*). Next, participants from all groups except G_{pre} and G_{0M} saw a page where they were supposed to enter an individual self-selected *code*. This code was necessary to permit us to link initial and subsequent measurements (without violating anonymity).

Afterwards for the $M_{Reminder-x-6M}$ the corresponding *reminder measure* was displayed with a short introduction of what to expect on the page. The three measures were designed to be similar: text, video and interactive examples required the same amount of time (8 minutes). Short Text, which was intentionally shorter, required 3 minutes to complete. Due to a technical error interactive examples was also set to 3 minutes.

Next, all groups saw the same page again. We used a *role play* approach. We told participants, before displaying the screenshots, that they should assume that they are someone called 'Martin Müller'. Relevant details about Martin were provided (see Appendix C). Then, the *email screenshots were shown in a random order*, one per page. For each of the 20 screenshots, they had to decide whether it was a phish or legitimate. At the end, participants had to answer a few *control questions* such as their usage of the Internet or revising the tutorial material. For the measurements after twelve months, participants were also asked whether they had already participated after six months.

5.6 Ethics and Data Protection

Participation in the evaluation was optional and the survey could have been completed at a time of the participant's choice. Participation was not remunerated in any way but they could have participated during their working hours. Due to strong privacy regulations in Germany, the anonymity of participants was a mandatory requirement. Therefore, we used SoSciSurvey to collect the data (they are compliant with the new European Data Protection Regulation). The previously described process to assign participants to groups, to invite and to remind them, as well as the fact that no demographic data was collected, was discussed with and approved by the works council, as this prevented any kind of individual performance monitoring. All information about the process, the anonymity, the agreement of the works council, as well as the fact that they did not need to fear consequences for poor performance was provided to the participants in the invitation email. It was also advised to get in touch with the CISM in case of any ambiguity or questions about the received email.

6 Results

We first provide information about our participants and then present the results for each of our three research questions.

[8] While this does not represent the usual ratio of phishing to legitimate emails employees might find in their inboxes, it is appropriate for knowledge assessments in particular to compare the performance against guessing

[9] The alternatives were either to use only http (which could have led to a misleading message that dangerous links always use http and a conclusion that https is always secure) or to combine each attack type applied with both which would have lead to twice as many email screenshots.

[10] https://www.soscisurvey.de/

6.1 Participants and Data Cleaning

A total of 439 participants completed the online survey (several due to the two measure points for research question 2 and 3). We performed the following data cleansing steps: (1) We excluded four participants whose answers evidenced specific patterns. They had 100% phishing email identification and 0% legitimate email identification respectively; i.e., they judged all emails as phishing emails. (2) We excluded 26 participants who admitted using the Internet or other sources to answer or right before answering the questions. Thus, the data from 409 participants was analysed.

6.2 Analysis Methods

We used the Signal Detection Theory (SDT) [97] to measure the participants' performance, i.e. whether participants' skills in distinguishing between phishing and legitimate emails improved, as compared to before the tutorial. This theory has been used in other studies evaluating phishing identification [14–19, 38, 75, 76, 79, 82, 93, 94]. SDT enables us to discern between signal (phishing emails) and noise (legitimate emails). In line with above-mentioned literature, we used the following two output values: sensitivity (d') and criterion (C). In the context of our research, sensitivity defined the skill to distinguish phishing emails (signal) from legitimate ones (noise). The larger d', the better the participants' performance in distinguishing signal from noise. Criterion (C) was defined as the response tendency, e.g. in our case whether participants were more cautious after the tutorial, i.e. more legitimate emails were classified as phishing (more false negatives), or did they take more risks, i.e. more phishing emails are classified as legitimate (more false positives). The closer this criterion was to 0, the more accurately they decided whether a signal was phish or legitimate.

We evaluated the assumptions relevant for calculating SDT parameters, i.e. equal variance and Gaussian distribution. Afterwards, we calculated the SDT parameters for sensitivity and criterion per participant. We then calculated the mean values per measurement using SPSS. To evaluate our hypotheses, we analysed the differences for participants' sensitivity and criterion values using one-way ANOVAs (using SPSS). ANOVA is a common tool to analyse forgetting curves as it overcomes the problem of initial learning levels [44]. For every ANOVA, we started off the analysis by checking the assumptions for both the sensitivity and the criterion. Since both sensitivity and criterion only violated the normal distribution assumption, and the ANOVA is relatively robust against the violation of this assumptions [42], we continued the analysis. For the descriptive results for the sensitivity see Figure 3, 4, and 5. The hit-rates for phishes and legitimate results are provided there as well. For simplicity and readability reasons we will only state the significant results in the following subsections (for full version see Appendix D).

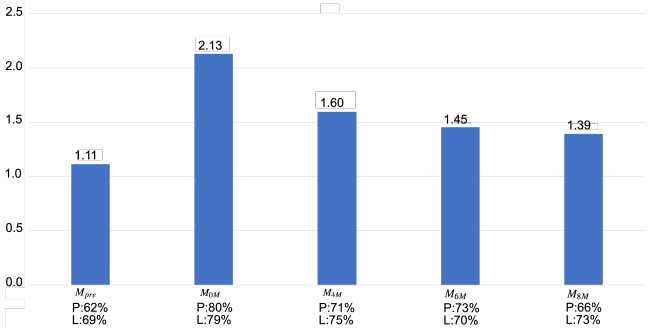


Figure 3: Sensitivity score and hit rates (Phish/Legitimate) for the measure of RQ1.

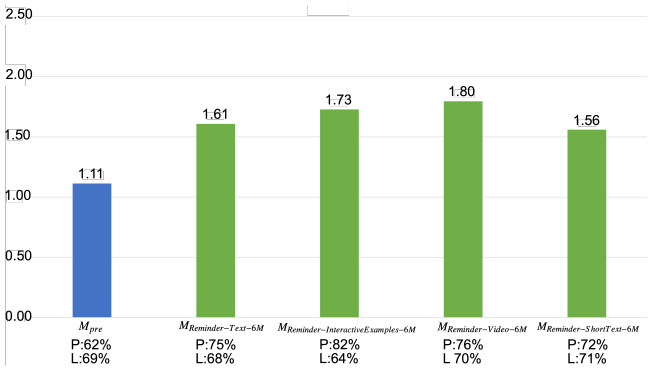


Figure 4: Sensitivity score and hit rates (Phish/Legitime) for the measure of RQ2 (blue = only tutorial, green = reminder measure groups).

6.3 Results for Research Question 1

As stated before, we analysed the data as soon as possible so that we could distribute the reminder measure as soon as the performance was no longer significantly better as compared to their performance before they participated in the on-site tutorial. We discovered that after a period of six months, performance was no longer significantly different from before the on-site tutorial. However, we decided to continue collecting data after eight months to strengthen our findings. We wanted to make sure that the difference between those two groups was not due to variance of the participants in our between-subjects design.

For the reporting of the results, we combine the analyses of $\boldsymbol{H}_{M_{pre}-M_{0M}}$ and $\boldsymbol{H}_{M_{pre}-M_{\Delta i M}}$ (for $\Delta t = 4+2i$ months, where $i \in \{0,1,2\}$) We checked for a significant difference between the corresponding five groups (see Figure 1) using a one-way ANOVA. There was statistical significance between the groups ($F(4,227) = 5.457$, $p < 0.001$) for the sensitivity (d'). For the effect size we calculated $\omega^2 = .093$, which is a medium effect size according to [42]. A LSD post-hoc showed that

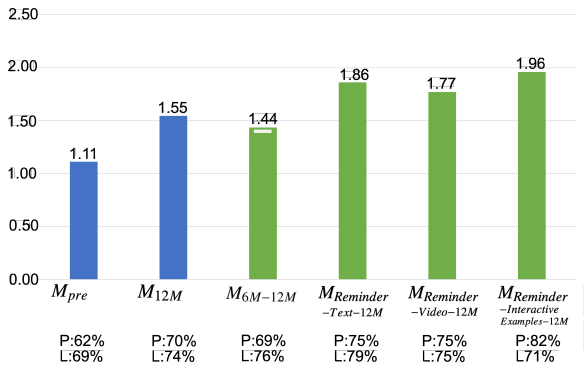


Figure 5: Sensitivity score and hit rates (Phish/Legitime) for the measure of RQ3 (blue = only tutorial, green = reminder measure groups).

the sensitivity for the M_{0M} ($d' = 2.13$, $SD = 1.15$) was significantly higher than for the M_{pre} ($d' = 1.11$, $SD = 1.12$) with ($p < .001$). The LSD post-hoc test showed that the sensitivity for the M_{4M} ($d' = 1.60$, $SD = 1.01$) was significantly higher than for the M_{pre} ($d' = 1.11$, $SD = 1.12$) with ($p = .034$). Note, there was no statistical significance between the groups ($p = 0.623$) for the criterion (C).

In summary: We accept $\boldsymbol{H}_{M_{pre}-0Month}$ and $\boldsymbol{H}_{M_{pre}-4Months}$.

6.4 Results for Research Question 2

First, we checked for which reminder measures the hypothesis $\boldsymbol{H}_{M_{pre}-M_{Reminder_x}-6M}$ holds. We checked for a significant difference between M_{pre} and the four months retention groups (see Figure 1) using a one-way ANOVA. There was statistical significance between the groups ($F(5,244) = 2.410$, $p = 0.037$) for the sensitivity (d'). For the effect size we calculated $\omega^2 = .027$, which is a small effect size [42]. A LSD post-hoc showed that the sensitivity for the $M_{Reminder-Text-6M}$ ($d' = 1.61$, $SD = 1.18$) with ($p = .005$), $M_{Reminder-Video-6M}$ ($d' = 1.80$, $SD = 1.42$) with ($p = .005$) and $M_{Reminder-InteractiveExamples-6M}$ ($d' = 1.73$, $SD = 1.19$) with ($p = .007$) were significantly higher than for the M_{pre} ($d' = 1.11$, $SD = 1.12$). Note, there was statistical significance between the groups ($p = 0.013$) for the criterion (C). A LSD post-hoc showed that the criterion for the $M_{Reminder-Text-6M}$ ($C = -.23$, $SD = .59$) with ($p = .043$) and $M_{Reminder-InteractiveExamples-6M}$ ($C = -.43$, $SD = .65$) with ($p < .001$) were significantly different from the M_{pre} ($C = .12$, $SD = .84$).

In summary: We accept $\boldsymbol{H}_{M_{pre}-M_{Reminder_x}-6M}$ for text measure, video measure, and interactive examples measure.

In order to test whether one of the three remaining reminder measures performs best, we also checked the ANOVA values for between the reminder measures. There is no significant difference between these measures. From the descriptive data, the interactive examples measure performs slightly better than the video measure (see Figure 5).

6.5 Results for Research Question 3

Based on the results from RQ2 we decided to not collect data from the short text group after 12 months. In order to address the pre-conditions from Section 5.1 we kind of extended $\boldsymbol{H}_{M_{pre}-M_{Reminder_y}-12M}$ accordingly, i.e. six measurements were considered (see Figure 2).

We linked participants using the provided codes. This resulted in 20 participants in M_{6M-12M}, 17 participants in $M_{Reminder-Text-12M}$, 17 participants in $M_{Reminder-Video-12M}$, and 12 participants in $M_{Reminder-InteractiveExamples-12M}$. We analysed the data from participants that we could link via code. We checked for a significant difference between the corresponding six measurements. There was statistical significance between the groups ($F(5,172) = 2.721$, $p = 0.022$) for the sensitivity (d'). The LSD post-hoc showed that the sensitivity for the $M_{Reminder-Text-12M}$ ($d' = 1.93$, $SD = 1.17$) with ($p = .009$), the $M_{Reminder-Video-12M}$ ($d' = 1.77$, $SD = 1.32$) with ($p = .031$) and $M_{Reminder-InteractiveExamples-12M}$ ($d' = 1.96$, $SD = 1.34$) with ($p = .016$) were significantly higher than for the M_{pre} ($d' = 1.11$, $SD = 1.12$). For the effect size we calculated $\omega^2 = .047$, which is a small effect size according to [42]. Note, there was no statistical significance between the measurements ($p = 0.274$) for the criterion (C).

In summary: The pre-conditions hold and we accept $\boldsymbol{H}_{M_{pre}-M_{Reminder_y}-12M}$ for video measure and interactive examples measure.

7 Discussion

We first discuss some general implications of our study, then our results for the three research questions and then the limitations of our work.

We excluded four participants because they marked all screenshots in the survey as phishing emails and therefore had 100% phishing detection but also 0% legitimate detection. In addition, we excluded 26 participants for seeking help for answering the survey. Seeking help is very useful in the real world. But as we could not control what kind of help they got and we explicitly mentioned that they should fill out the survey without help, we decided to exclude those that violated our rule.

We discussed advantages and disadvantages of publishing the results with the organisation. In particular the potential risk to the organization caused by publishing the results was evaluated. Together with the organisation it was decided to name a few key facts about the organization. We wanted to give the opportunity to other researchers to know the study setting in order to allow transferring the information to other contexts and making sure that our results can be correctly interpreted.

7.1 Discussion of RQ-1

The participants' skills in identifying phishing messages improved significantly straight after attending the tutorialsOur results are in line with those of the study evaluating the effectiveness of the original measure [84]. Thus, the customisation, as well as the switch to an instructor-based tutorial approach, seems not to have affected the efficacy of the content.

While the pre-condition for the first research question holds and the phishing detection rate increased from 62% to 80%, a closer inspection of the results begs the question whether an 80% phishing identification rate in the M_{0M} measurement leaves the participants sufficiently protected, considering the fact that security was their primary goal in our evaluation. After some internal deliberation about these numbers, the improvements were seen as a success at SOGGS, since this was the first organisation-wide security awareness and education measure and employees skills in distinguishing legitimate and phishing emails were significantly enhanced by the tutorial. We will again deliver security awareness and education measures at SOGGS with the goal of further increasing these numbers.

With respect to the performance over time, we found that after six and eight months participants' skills were no longer significantly better than before participating in the tutorial. This also aligns with prior research reporting results of retention tests, albeit all of these studies had shorter retention periods. While they all used different interventions and also different evaluation techniques, they all found that the effect lasted until they conducted the retention study (which was max. after 5 months in [22]).

Our results show that current reminder periods required by standards such as PCI-DSS [27] – which usually require an interval of twelve months – should be re-considered.

We are aware that the SOGGS has a higher age average, as 60% of the employees are older than 50 years. One might assume that we would have achieved different results with younger participants. According to [44, 101] the age does not increase the forgetting rate significantly. Therefore, we argue that our results also hold true for younger people.

7.2 Discussion of RQ-2

Concerning the results for the reminder measures, both the video measure and the interactive examples measure stand out from the others in terms of sensitivity (1.797 and 1.728 versus 1.559 for the text measures after 6 months; and the text measure not being significantly better after 12 months compared to the measurement before the tutorial). Thus, our results indicate that – in line with related work such as [25, 67] – static measures lead to a poorer performance than dynamic measures. Furthermore, our results show that even short reminder measures can be very effective and it is neither necessary nor recommended (because of the time needed) to use the main security awareness and education measure as reminder measure. Yet, it must also be noted that there is a lower bound to the information which must be included in the reminder material, as evidenced by the insufficient performance of the short text measure.

Overall, for SOGGS, the combination of costly on-site tutorials and an efficient reminder measure after six months looks very promising.

Whether the video measure or the interactive examples measure perform better is not that easily answered as there is no significant difference between the two. Considering the criterion values, we could argue the video measure achieved the best results. With respect to sensitivity, there is no clear 'winner' after 12 months. Note that although the performance after 6 months for the video measure and interactive examples measure is not significantly lower than directly after the on-site tutorial, the ultimate goal must be to get as close as possible to the performance achieved directly after the tutorial (see Figure 4).

7.3 Discussion of RQ-3

Following the discussion of RQ-1, the performance of participants who received the video measure and the interactive examples measure is so good that a refreshment might not even be necessary after another six months has elapsed. Such results should also be taken into account when reconsidering time intervals of international standards, as discussed in Section 7.1. In order to know when subsequent reminders should be scheduled, future research into their long-term effects is required.

7.4 Recommendations for Future Studies

We faced several challenges during our research. We discuss here how they were addressed to assist researchers planning similar studies - which we would welcome.

One challenge with our study was to avoid reporting an effect for our reminder measures while the effect was actual caused by external factors: e. g. media reports, or internal discussions. This was addressed by including the 12 month group and comparing their performance to those of the reminder groups.

The next challenge is that the exact point in time when we measured the performance could not be controlled (also in comparison to the participation in the tutorials which were offered for an entire month). To address this, we limited the time span for filling out the survey to two weeks after having received the invitation email. Note that while this might be an unusual design choice, this is reasonable in a field study as not all employees would receive a security awareness and education measure on the same day.

7.5 Limitations

Even though all instructors used the same measures and the same instructions for conducting the tutorial, it is still possible that there were small differences in the course of the held tutorials.Some of these groups were trained by the CISM herself. *This limitation was mitigated by the random assignment of employees from different locations to study groups.*

Our study design selected measures for best case scenarios, with security being the primary goal. We argue that it is worth testing in such scenarios as it is a pre-condition for identifying a phishing email during any working day. The results show that this pre-condition is far from being a given (the M_{pre} only detects 63% of the phishing emails). Furthermore, most of the example emails could only be identified as such when the URL behind the link and/or the file type were checked. Thus the phishing emails used in the evaluation were more difficult to identify as compared to the average phishing emails received by SOGSS employees. *For the purpose of our research, best case and poor performance before participating in the tutorial, this approach is appropriate.* We acknowledge that for statements in the actual working environment with actual received phishing emails, the study design would need to be different.

In addition, due to the restrictions of SoSci Survey, we could only provide screenshots, i.e. it was not necessary for participants to move the mouse to the link as it was already in the correct position, with the URL displayed. Furthermore, it was not possible to check several of the integrated links or get additional information such as the html source of the email. Thus, on the one hand we made it a bit easier and on the other hand a bit more difficult as compared to actual phishing email detection. Thus, in reality, the detection rate of the evaluated phishing emails is likely to be different if employees would have received them in their inboxes and were asked to judge them there. However, this would have made a lab setting necessary, which was not possible.

In order to keep the duration of the study feasible, we were restricted in the number of evaluated phishing emails. We selected a representative sample of emails from the large variety of possible options. It might be that different combinations would have given us different results. However, we believe that due to the selection of representative examples, the findings would not have been significantly different.

Our study was customised for, and conducted at, a German public sector organisation (SOGSS). Therefore, our participant sample is biased by the type of work and the technical background of the employees. We would need to replicate our study in other types of organisations and organisations with different employee characteristics. This is part of future work.

Due to a technical error, the group interactive examples had to spend only 3 minutes engaging with the material and not 8 minutes as planned (similar to text and video). Even though this meant that the participants spent less time with the interactive examples measure than initially planned (about 7.3 minutes on average see Table 5 in the appendix), it still produced excellent results. There is currently no evidence to suggest that a more extended time spent with the material would have had a negative effect. Since, despite the aforementioned technical error, the measure achieved significant results with both $M_{Reminder-InteractiveExamples-6M}$ and $M_{Reminder-InteractiveExamples-12M}$, one would expect either the same or a better effect over a more extended period. *Therefore, we believe that this did not impact our results.*

Finally, it is worth mentioning that our results for phishing hold although in the tutorial three topics were addressed (and not just phishing). Thus, it might be that the effect would last slightly longer if only one topic were addressed. It is open to discussion which scenario is more realistic (a single topic tutorial or one with some similar topics.

8 Conclusion

We presented a study on how effective security awareness and education measures are over time, and what the best way is to remind users' awareness and knowledge. To this end, we carried out a field investigation within a German "State Office for Geoinformation and State Survey". We considered three research questions: i) *How long does the effect of the on-site tutorial last, i.e., when should the gained awareness and knowledge be reminded?*, ii) *Which of the four developed reminder measures performed best?*, and iii) *How long does the effect of reminder measures last?*.
From the almost 2000 employees, 409 voluntarily participated. From the fourth month after the on-site tutorial, we evaluated groups every two months to measure awareness and knowledge retention. After six months, we saw no improved performance in distinguishing phishing and legitimate emails. Four reminder measures were distributed to four groups (one per group): a) text, b) video measure, c) interactive examples, and d) a short text. Twelve months after the tutorial, we compared the knowledge retention of the four reminder groups with the the pre-group. Among the four reminder measures, the video measure and the interactive examples measure performed best, with their impact lasting at least six months after being rolled-out.

Acknowledgments

The research reported in this paper has furthermore been supported by the German Federal Ministry of Education and Research within the framework of the project KASTEL_SKI in the Competence Center for Applied Security Technology (KASTEL).

References

[1] Pieter Agten, Wouter Joosen, Frank Piessens, and Nick Nikiforakis. Seven months' worth of mistakes: A longitudinal study of typosquatting abuse. In *Proceedings of the 22nd Network and Distributed System Security Symposium (NDSS 2015)*. Internet Society, 2015.

[2] Devdatta Akhawe and Adrienne Porter Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *22th USENIX Security CHI*, 2013.

[3] Abdullah Alnajim and Malcolm Munro. An Anti-Phishing Approach that Uses Training Intervention for Phishing Websites Detection . In *6th International Conference on Information Technology: New Generations*, pages 405–410. IEEE, 2009.

[4] Abdullah Alnajim and Malcolm Munro. An Anti-Phishing Approach that Uses Training Intervention for Phishing Websites Detection. *2009 Sixth International Conference on Information Technology: New Generations*, pages 405–410, 2009.

[5] Abdullah Alnajim and Malcolm Munro. An approach to the implementation of the anti-phishing tool for phishing websites detection. In *2009 International Conference on Intelligent Networking and Collaborative Systems*, pages 105–112. IEEE, 2009.

[6] Ibrahim Alseadoon. *The impact of users' characteristics on their ability to detect phishing emails.* PhD thesis, (Doctoral dissertation, Queensland University of Technology), 2014.

[7] Mohamed Alsharnouby, Furkan Alaca, and Sonia Chiasson. Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies*, 82, 2015.

[8] Kholoud Althobaiti, Ghaidaa Rummani, and Kami Vaniea. A review of human-and computer-facing url phishing features. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 182–191. IEEE, 2019.

[9] Kholoud Althobaiti, Kami Vaniea, and Serena Zheng. Faheem: Explaining urls to people using a slack bot. In *Symposium on Digital Behaviour Intervention for Cyber Security (AISB)*, pages 1–8, 2018.

[10] Nalin AG Arachchilage, Ivan Flechais, and Konstantin Beznosov. A game storyboard design for avoiding phishing attacks. In *Proceedings of the 11th Symposium On Usable Privacy and Security (SOUPS)*, 2014.

[11] Nalin Asanka Gamagedara Arachchilage, Steve Love, and Konstantin Beznosov. Phishing threat avoidance behaviour: An empirical investigation. *Computers in Human Behavior*, 60:185–197, 2016.

[12] Zinaida Benenson, Freya Gassmann, and Robert Landwirth. Unpacking spear phishing susceptibility. In *Financial Cryptography and Data Security*, 2017.

[13] Clemens Bergmann and Gamze Canova. Design, implementation and evaluation of an anti-phishing education app. Master's thesis, Technische Universität Darmstadt, 2014.

[14] M Butavicius, K Parsons, M Pattinson, A McCormac, D Calic, and M Lillie. Understanding Susceptibility to Phishing Emails: Assessing the Impact of Individual Differences and Culture. *International Symposium on Human Aspects of Information Security & Assurance (HAISA)*, 2017.

[15] Marcus Butavicius, Kathryn Parsons, Malcolm Pattinson, and Agata McCormac. Breaching the human firewall: Social engineering in phishing and spear-phishing emails. *Australasian Conference on Information Systems*, 2016.

[16] Casey Canfield, Alex Davis, Baruch Fischhoff, and Forget -A on Usable .... Replication: Challenges in using data logs to validate phishing detection ability metrics. *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[17] Casey Canfield, Baruch Fischhoff, and Alex Davis. Using Signal Detection Theory to Measure Phishing Detection Ability and Behavior. In *SOUPS*, 2015.

[18] Casey Canfield, Baruch Fischhoff, and Alex Davis. Quantifying Phishing Susceptibility for Detection and Behavior Decisions. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 58:1158–1172, 2016.

[19] Casey Inez Canfield and Baruch Fischhoff. Setting Priorities in Behavioral Interventions: An Application to Reducing Phishing Risk. *Risk Analysis*, 38:826–838, 2018.

[20] Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Roland Borza. NoPhish: An Anti-Phishing Education App. In *Security and Trust Management*, pages 188–192. LNCS, 2014.

[21] Gamze Canova, Melanie Volkamer, Clemens Bergmann, Roland Borza, Benjamin Reinheimer, Simon Stockhardt, and Ralf Tenberg. Learn to Spot Phishing URLs with the Android NoPhish App. In *WISE 9*, pages 87–100. Springer, 2015.

[22] Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Benjamin Reinheimer. NoPhish App Evaluation: Lab and Retention Study. In *USEC*. Internet Society, 2015.

[23] Madhusudhanan Chandrasekaran, Krishnan Narayanan, and Shambhu Upadhyaya. Phishing E-mail Detection Based on Structural Properties. *NYS Cyber Security conference*, 2006.

[24] Sidharth Chhabra, Anupama Aggarwal, Fabricio Benevenuto, and Ponnurangam Kumaraguru. Phi. sh/$ ocial: the phishing landscape through short urls. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, pages 92–101, 2011.

[25] Hee Jun Choi and Scott D. Johnson. The Effect of Context-Based Video Instruction on Learning and Motivation in Online Courses. *American Journal of Distance Education*, 19(4):215–227, 2005.

[26] Dan Conway, Ronnie Taib, Mitch Harris, Shlomo K Berkovsky, Kun Yu, and Fang Chen. A qualitative investigation of bank employee experiences of information security and phishing. *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[27] PCI Security Standards Council. PCI DSSv3.2.1 - May 2018, 2018. https://www.pcisecuritystandards.org/document_library accessed: 27.02.2020.

[28] Eugène JFM Custers. Long-term retention of basic science knowledge: a review study. *Advances in Health Sciences Education*, 15(1):109–128, 2010.

[29] Rachna Dhamija and J. D. Tygar. The battle against phishing: Dynamic Security Skins. In *Symposium on Usable Privacy and Security*, pages 77–88, New York, NY, USA, 2005. ACM.

[30] Rachna Dhamija, J. D. Tygar, and Marti Hearst. Why phishing works. In *Proceedings of CHI 2006 Human Factors in Computing Systems*, pages 581–590. ACM, 2006.

[31] Alejandra Diaz, Alan T Sherman, and Anupam Joshi. Phishing in an academic community: A study of user susceptibility and behavior. *Cryptologia*, pages 1–15, 2019.

[32] Matt Dixon, Nalin AG Arachchilage, and James Nicholson. Engaging Users with Educational Games: The Case of Phishing. *Woodstock*, 2019.

[33] Ronald C Dodge, Curtis Carver, and Aaron J Ferguson. Phishing for user security awareness. *Computers and Security*, 26(1):73–80, Elsevier, 2007.

[34] Julie S Downs, Mandy Holbrook, and Lorrie Faith Cranor. Behavioral response to phishing risk. *Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit*, pages 37–44, 2007.

[35] Julie S Downs, Mandy B Holbrook, and Lorrie Faith Cranor. Decision strategies and susceptibility to phishing. In *Proceedings of the Second Symposium on Usable Privacy and Security*, pages 79–90, 2006.

[36] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. In *CHI*, pages 1065–1074. ACM, 2008.

[37] Serge Egelman and Eyal Peer. Scaling the security wall: Developing a security behavior intentions scale (sebis). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2873–2882, 2015.

[38] Iain Embrey and Kim Kaivanto. Many Phish in the C: A Coexisting-Choice-Criteria Model of Security Behavior. *arxiv*, 2018.

[39] Jussi-Pekka Erkkilä. Why we fall for phishing. In *Conference on Human Factors in Computer Systems*. ACM, 2011.

[40] Adrienne Porter Felt, Robert W Reeder, Alex Ainslie, Helen Harris, Max Walker, Christopher Thompson, Mustafa Embre Acer, Elisabeth Morant, and Sunny Consolvo. Rethinking connection security indicators. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*, pages 1–14, 2016.

[41] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In *WWW*, pages 649–656. ACM, 2007.

[42] Andy Field. *Discovering statistics using IBM SPSS statistics*. sage, 2013.

[43] Dorota Filipczuk, Charles Mason, and Stephen Snow. Using a Game to Explore Notions of Responsibility for Cyber Security in Organisations. In *Proceedings of CHI 2019 Human Factors in Computing Systems*, pages 1–6, 2019.

[44] Anders M Fjell, Kristine B Walhovd, Ivar Reinvang, Arvid Lundervold, Anders M Dale, Brian T Quinn, Nikos Makris, and Bruce Fischl. Age does not increase rate of forgetting over weeks—neuroanatomical volumes and visual memory across the adult life-span. *Journal of the International Neuropsychological Society*, 11(1):2–15, 2005.

[45] Waldo Rocha Flores, Hannes Holm, Marcus Nohlberg, and Mathias Ekstedt. Investigating personal determinants of phishing and the effect of national culture. *Information and Computer Security*, Volume 23:178–199, 2015.

[46] Vaibhav Garg, Jean Camp, Lesa Mae, and Katherine Connelly. Designing risk communication for older adults. In *Symposium on Usable Privacy and Security (SOUPS)*. Citeseer, 2011.

[47] Robin Gonzalez and Michael E Locasto. An interdisciplinary study of phishing and spear-phishing attacks. *SOUPS*, 2015.

[48] Tzipora Halevi, Jim Lewis, and Nasir Memon. Phishing, Personality Traits and Facebook. *arXiv*, 2013.

[49] Tzipora Halevi, Nasir Memon, and Oded Nov. Spear-Phishing in the Wild: A Real-World Study of Personality, Phishing Self-Efficacy and Vulnerability to Spear-Phishing Attacks. *SSRN Electronic Journal*, 2015.

[50] Kyung Wha Hong, Christopher M. Kelley, Rucha Tembe, Emerson Murphy-Hill, and Christopher B. Mayhorn. Keeping Up With The Joneses: Assessing phishing susceptibility in an email task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57:1012–1016, 2013.

[51] Cristian Iuga, Jason RC Nurse, and Arnau Erola. Baiting the hook: factors impacting susceptibility to phishing attacks. *Human-Centric Computing and Information Sciences*, 6:8, 2016.

[52] Ankit Jain and BB Gupta. Phishing Detection: Analysis of Visual Similarity Based Approaches. *Security and Communication Networks*, 2017:1–20, 2017.

[53] Markus Jakobsson, Alex Tsow, Ankur Shah, Eli Blevis, and Youn-Kyung Lim. What instills trust? A qualitative study of phishing. In *Financial Cryptography*, pages 356–361. LNCS, 2007.

[54] Timothy Kelley and Bennett I Bertenthal. Real-World Decision Making: Logging Into Secure vs. Insecure Websites. *USEC*, 2016.

[55] Sungjin Kim, Jinkook Kim, and Brent ByungHoon Kang. Malicious url protection based on attackers' habitual behavioral analysis. *Computers & Security*, 77:790–806, 2018.

[56] Panagiotis Kintis, Najmeh Miramirkhani, Charles Lever, Yizheng Chen, Rosa Romero-Gómez, Nikolaos Pitropakis, Nick Nikiforakis, and Manos Antonakakis. Hiding in plain sight: A longitudinal study of combosquatting abuse. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 569–586, 2017.

[57] Iacovos Kirlappos and Martina A Sasse. Security education against phishing:A modest proposal for a major rethink. *IEEE Security & Privacy*, 10(2):24–32, 2012.

[58] Iacovos Kirlappos and Martina Angela Sasse. Security Education against Phishing: A Modest Proposal for a Major Rethink. *Security and Privacy*, 10(2):24–32, 2012.

[59] Sabina Kleitman, Marvin KH Law, and Judy Kay. It's the deceiver and the receiver: Individual differences in phishing susceptibility and false positives with item profiling. *PLOS ONE*, 13:e0205089, 2018.

[60] Ponnurangam Kumaraguru, Justin Cranshaw, and Alessandro Acquisti. School of phish: a real-world evaluation of anti-phishing training. *SOUPS*, 2009.

[61] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Protecting People from Phishing: The Design and Evaluation of an Embedded Training Email System. In *CHI*, pages 905–914. ACM, 2007.

[62] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)*, 10:7, 2010.

[63] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching Johnny Not to Fall for Phish. *Transactions on Internet Technology*, 10(2):1–7, ACM, 2010.

[64] Alexandra Kunz, Melanie Volkamer, Simon Stockhardt, Sven Palberg, Tessa Lottermann, and Eric Piegert. Nophish: Evaluation of a web application that teaches people being aware of phishing attacks. In *Informatik*, pages 15–24. GI, GI, LNI, 2016.

[65] Elmer Lastdrager, Ines C Gallardo, Pieter Hartel, and Marianne Junger. How effective is anti-phishing training for children? *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[66] Elaine Lau and Zachary NJ Peterson. A research framework and initial study of browser security for the visually impaired. In *Symposium on Usable Privacy and Security (SOUPS)*, 2015.

[67] D. Lewalter. Cognitive strategies for learning from static and dynamic visuals. *Learning and Instruction*, 13(2):177 – 189, 2003.

[68] Peter Likarish, Donald E Dunbar, Juan Pablo Hourcade, and Eunjin Jung. Bayeshield: conversational anti-phishing user interface. In *SOUPS*, volume 9, pages 1–1, 2009.

[69] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. Does domain highlighting help people identify phishing sites? In *Proceedings of CHI 2011 Human Factors in Computing Systems*, pages 2075–2084. ACM, 2011.

[70] Gang Liu, Guang Xiang, Bryan A Pendleton, Jason I Hong, and Wenyin Liu. Smartening the crowds: computational techniques for improving human verification to fight phishing scams. *Symposium on Usable Privacy and Security (SOUPS)*, 2011.

[71] Hans-Henning Lühr. Warm welcome to the website of the it planning council, 2019. https://www.it-planungsrat.de/EN/home/home_node.html;jsessionid=C4D654065D8D53558E2CDDD6287D957B.1_cid332 accessed: 11.09.2019.

[72] Samuel Marchal, Giovanni Armano, Tommi Grondahl, Kalle Saari, Nidhi Singh, and N. Asokan. Off-the-Hook: An Efficient and Usable Client-Side Phishing Prevention Application. *IEEE Transactions on Computers*, 66, 2017.

[73] Claudio Marforio, Ramya Jayaram Masti, Claudio Soriente, Kari Kostiainen, and Srdjan Čapkun. Evaluation of personalized security indicators as an anti-phishing mechanism for smartphone applications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 540–551. ACM, 2016.

[74] Charlie Marriott. *Through the Net Investigating How User Characteristics Influence Susceptibility to Phishing.* PhD thesis, Dublin Institute of Technology, 2018.

[75] Jaclyn Martin. *Something Looks Phishy Here: Applications of Signal Detection Theory to Cyber-Security Behaviors in the Workplace*. PhD thesis, University of South Florida, 2017.

[76] Jaclyn Martin, Chad Dubé, and Michael D Coovert. Signal Detection Theory (SDT) Is Effective for Modeling User Behavior Toward Phishing and Spear-Phishing Attacks. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 60:1179–1191, 2018.

[77] Max-Emanuel Maurer and Dennis Herzner. Using visual website similarity for phishing detection and reporting. *Conference on Human Factors in Computing Systems (CHI)*, 2012.

[78] Peter Mayer, Christian Schwartz, and Melanie Volkamer. On the systematic development and evaluation of password security awareness-raising materials. In *Proceedings of the 34th Annual Computer Security Applications Conference*, pages 733–748. ACM, 2018.

[79] Christopher B Mayhorn and Patrick G Nyeste. Training users to counteract phishing. *Work (Reading, Mass.)*, 41 Suppl 1:3549–52, 2012.

[80] Jamshaid G Mohebzada, Ahmed El Zarka, Arsalan H Bhojani, and Ali Darwish. Phishing in a university community: Two large scale phishing experiments. In *2012 International Conference on Innovations in Information Technology (IIT)*, pages 249–254. IEEE, 2012.

[81] Gregory D Moody, Dennis F Galletta, and Brian Dunn. Which phish get caught? An exploratory study of individuals' susceptibility to phishing. *European Journal of Information Systems*, 26:564–584, 2017.

[82] María M Moreno-Fernández, Fernando Blanco, Pablo Garaizar, and Helena Matute. Fishing for phishers. Improving Internet users' sensitivity to visual deception cues to prevent electronic fraud. *Computers in Human Behavior*, 69:421–436, 2017.

[83] Steven J Murdoch and Martina A Sasse. Should you phish your own employees? https://www.benthamsgaze.org/2017/08/22/should-you-phish-your-own-employees/, 2017. accessed: 18.09.2019.

[84] Stephan Neumann, Benjamin Reinheimer, and Melanie Volkamer. Don't be deceived: the message might be fake. In *International Conference on Trust and Privacy in Digital Business*, pages 199–214. Springer, 2017.

[85] Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam Soliman, Tian Lin, and Natalie Ebner. Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, page 6412–6424, 2017.

[86] Kathryn Parsons, Agata McCormac, Malcolm Pattinson, Marcus Butavicius, and Cate Jerram. Phishing for the Truth: A Scenario-Based Experiment of Users' Behavioural Response to Emails. pages 366–378, 2013.

[87] M Pattinson, C Jerram, K Parsons, A McCormac, and M Butavicius. Managing Phishing Emails: A Scenario-Based Experiment. pages 74–85, 2011.

[88] Justin Petelka, Yixin Zou, and Florian Schaub. Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings . In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[89] Richard Roberts, Yaelle Goldschlag, Rachel Walter, Taejoong Chung, Alan Mislove, and Dave Levin. You are who you appear to be: A longitudinal study of domain impersonation in tls certificates. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2489–2504, 2019.

[90] Dawn M Sarno, Joanna E Lewis, Corey J Bohil, Mindy K Shoss, and Mark B Neider. Who are Phishers luring?: A Demographic Analysis of Those Susceptible to Fake Emails. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61:1735–1739, 2017.

[91] Stuart E Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The emperor's new security indicators. In *2007 IEEE Symposium on Security and Privacy (SP'07)*, pages 51–65. IEEE, 2007.

[92] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of CHI 2010 Human Factors in Computing Systems*, pages 373–382. ACM, 2010.

[93] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. *Symposium on Usable Privacy and Security (SOUPS)*, 2007.

[94] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 88–99. ACM, 2007.

[95] Akbar Siami Namin, Rattikorn Hewett, Keith S Jones, and Rona Pogrund. Sonifying internet security threats. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2306–2313. ACM, 2016.

[96] Gunikhan Sonowal, KS Kuppusamy, and Ajit Kumar. Usability evaluation of active anti-phishing browser extensions for persons with visual impairments. *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 1–6, 2017.

[97] Harold Stanislaw and Natasha Todorov. Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1):137–149, 1999.

[98] Simon Stockhardt, Benjamin Reinheimer, Melanie Volkamer, Peter Mayer, Alexandra Kunz, Philipp Rack, and Daniel Lehmann. Teaching Phishing-Security: Which Way is Best? *IFIP SEC*, pages 135–149, 2016.

[99] Janos Szurdi, Balazs Kocso, Gabor Cseh, Jonathan Spring, Mark Felegyhazi, and Chris Kanich. The long "taile" of typosquatting domain names. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pages 191–206, 2014.

[100] Christopher Thompson, Martin Shelton, Emily Stark, Maximilian Walker, Emily Schechter, and Adrienne Porter Felt. The web's identity crisis: understanding the effectiveness of website identity indicators. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1715–1732, 2019.

[101] Tom N Tombaugh and Anita M Hubley. Rates of forgetting on three measures of verbal learning using retention intervals ranging from 20 min to 62 days. *Journal of the International Neuropsychological Society*, 7(1):79–91, 2001.

[102] Kai Florian Tschakert and Sudsanguan Ngamsuriyaroj. Effectiveness of and user preferences for security awareness training methodologies. *Heliyon*, 5:e02010, 2019.

[103] Anthony Vance, Brock Kirwan, Daniel Bjorn, Jeffrey Jenkins, and Bonnie Anderson. What Do We Really Know about How Habituation to Warnings Occurs Over Time?: A Longitudinal fMRI Study of Habituation and Polymorphic Warnings. In *Proceedings of CHI 2017 Human Factors in Computing Systems*, pages 2215–2227, 2017.

[104] Melanie Volkamer, Karen Renaud, and Paul Gerber. Spot the phish by checking the pruned URL. *Information and Computer Security*, Volume 24:372–385, 2016.

[105] Melanie Volkamer, Karen Renaud, and Benjamin Reinheimer. Torpedo: Tooltip-powered phishing email detection. In *IFIP SEC*, pages 161–175. Springer, 2016.

[106] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. User experiences of torpedo: Tooltip-powered phishing email detection. *Computers & Security*, 2017.

[107] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, Philipp Rack, Marco Ghiglieri, Peter Mayer, Alexandra Kunz, and Nina Gerber. Developing and Evaluating a Five Minute Phishing Awareness Video. *Trust, Privacy and Security in Digital Business (TrustBus)*, pages 119–134, 2018.

[108] Melanie Volkamer, Simon Stockhardt, Steffen Bartsch, and Michaela Kauer. Adopting the cmu/apwg anti-phishing landing page idea for germany. In *STAST*, pages 46–52. IEEE, 2013.

[109] Jingguo Wang, Yuan Li, Columbia College, H Raghav Rao, and The University of Texas at San Antonio. Overconfidence in Phishing Email Detection. *Journal of the Association for Information Systems*, 17:759–783, 2016.

[110] Rick Wash and Molly Cooper. Who Provides Phishing Training? In *Proceedings of CHI 2018 Human Factors in Computing Systems*, 2018.

[111] Zikai Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. What.Hack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game. page 108, 2019.

[112] Zikai Alex Wen, Yiming Li, Reid Wade, Jeffrey Huang, and Amy Wang. What.Hack: Learn Phishing Email Defence the Fun Way. pages 234–237, 2017.

[113] Min Wu, Robert C Miller, and Simson L Garfinkel. Do security toolbars actually prevent phishing attacks? *CHI*, pages 601–610, 2006.

[114] Weining Yang, Jing Chen, Aiping Xiong, Robert W Proctor, and Ninghui Li. Effectiveness of a phishing warning in field settings. *the 2015 Symposium and Bootcamp*, page 14, 2015.

[115] Ka-Ping Yee. Designing and evaluating a petname anti-phishing tool. In *Poster presented at Symposium on usable Privacy and Security (SOUPS)*, pages 6–8, 2005.

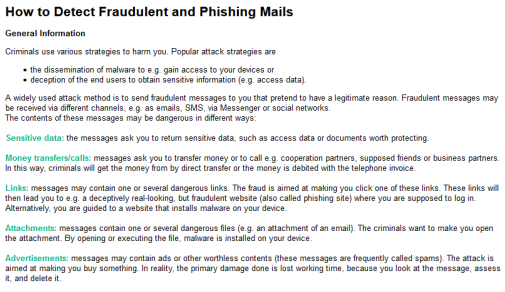
[116] Tianjian Zhang. Knowledge Expiration in Security Awareness Training. *Conference on Digital Forensics, Security and Law (ADFSL)*, 2018.

[117] Yue Zhang, Serge Egelman, Lorrie Faith Cranor, and Jason Hong. Phinding Phish: Evaluating Anti-Phishing Tools. In *NDSS*. School of Computer Science, Internet Society, 2007.

[118] Yue Zhang, Jason I Hong, and Lorrie F Cranor. CANTINA: A Content-Based Approach to Detecting Phishing Web Sites. In *16th International World Wide Web Conference*, pages 639–648, 2007.

[119] Olga A Zielinska, Rucha Tembe, Kyung Hong, Xi Ge, Emerson Murphy-Hill, and Christopher B Mayhorn. One Phish, Two Phish, How to Avoid the Internet Phish. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58:1466–1470, 2014.

A Reminder Measures



How to Detect Fraudulent and Phishing Mails

General Information

Criminals use various strategies to harm you. Popular attack strategies are

- the dissemination of malware to e.g. gain access to your devices or
- deception of the end users to obtain sensitive information (e.g. access data).

A widely used attack method is to send fraudulent messages to you that pretend to have a legitimate reason. Fraudulent messages may be received via different channels, e.g. as emails, SMS, via Messenger or social networks.
The contents of these messages may be dangerous in different ways:

Sensitive data: the messages ask you to return sensitive data, such as access data or documents worth protecting.

Money transfers/calls: messages ask you to transfer money or to call e.g. cooperation partners, supposed friends or business partners. In this way, criminals will get the money from by direct transfer or the money is debited with the telephone invoice.

Links: messages may contain one or several dangerous links. The fraud is aimed at making you click one of these links. These links will then lead you to e.g. a deceptively real-looking, but fraudulent website (also called phishing site) where you are supposed to log in. Alternatively, you are guided to a website that installs malware on your device.

Attachments: messages contain one or several dangerous files (e.g. an attachment of an email). The criminals want to make you open the attachment. By opening or executing the file, malware is installed on your device.

Advertisements: messages may contain ads or other worthless contents (these messages are frequently called spams). The attack is aimed at making you buy something. In reality, the primary damage done is lost working time, because you look at the message, assess it, and delete it.

Figure 6: Excerpt of the *text* reminder measure.



Figure 7: An impression of the video reminder measure.

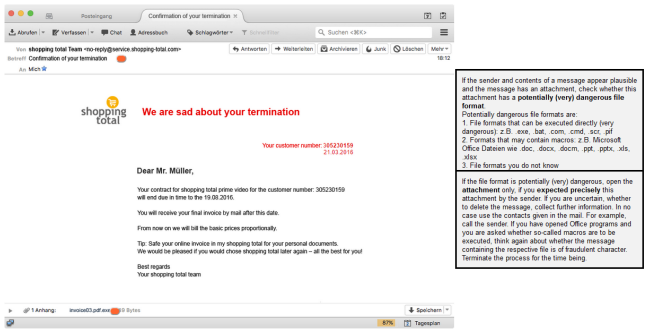
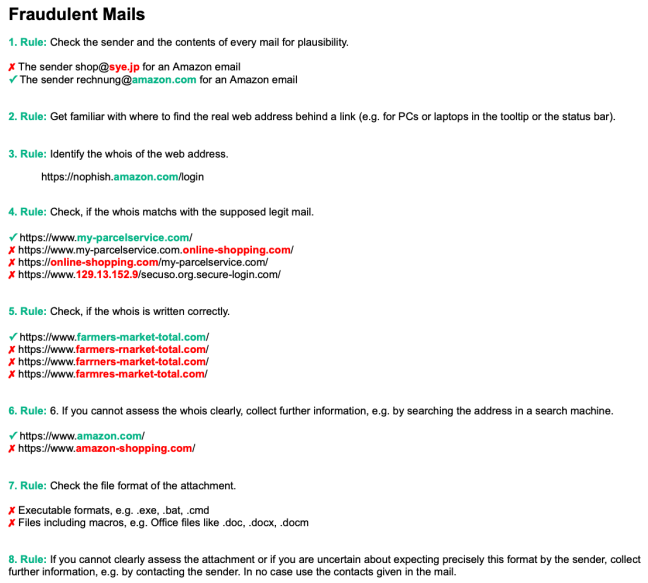


Figure 8: One of the interactive example measure used. The red dots represent the interactivity-points where participants can reveal more information about the respective area of the email.



Fraudulent Mails

1. Rule: Check the sender and the contents of every mail for plausibility.

✗ The sender shop@**sye.jp** for an Amazon email
✓ The sender rechnung@**amazon.com** for an Amazon email

2. Rule: Get familiar with where to find the real web address behind a link (e.g. for PCs or laptops in the tooltip or the status bar).

3. Rule: Identify the whois of the web address.

https://nophish.**amazon.com**/login

4. Rule: Check, if the whois matchs with the supposed legit mail.

✓ https://www.**my-parcelservice.com**/
✗ https://www.my-parcelservice.com.**online-shopping.com**/
✗ https://**online-shopping.com**/my-parcelservice.com/
✗ https://www.**129.13.152.9**/secuso.org.secure-login.com/

5. Rule: Check, if the whois is written correctly.

✓ https://www.**farmers-market-total.com**/
✗ https://www.**farmers-rmarket-total.com**/
✗ https://www.**farmers-market-total.com**/
✗ https://www.**farmres-market-total.com**/

6. Rule: 6. If you cannot assess the whois clearly, collect further information, e.g. by searching the address in a search machine.

✓ https://www.**amazon.com**/
✗ https://www.**amazon-shopping.com**/

7. Rule: Check the file format of the attachment.

✗ Executable formats, e.g. .exe, .bat, .cmd
✗ Files including macros, e.g. Office files like .doc, .docx, .docm

8. Rule: If you cannot clearly assess the attachment or if you are uncertain about expecting precisely this format by the sender, collect further information, e.g. by contacting the sender. In no case use the contacts given in the mail.

Figure 9: The shortened text measure.

B Attack Types

(a) Phishers may try to trick recipients by display the legitimate URL in the email's message text (and hope that recipients do not check the actual link destination). (b) Phishers may try to trick recipients by 'scipting' a tooltip with the legitimate URL next to the link. (c) Phishers replace the legitimate URL with a domain that they own (which has no connection to the expected domain). They might adopt either the subdomain (e.g., `https://www.amazon.com.phisher.com`) and/or the path so that the expected domain appears to allay suspicions. (d) Phishers replace the legitimate URL by a link to a domain that they own, and which looks very similar to the expected one (e.g., `arnazon.de`). (e) Phishers replace the legitimate URL by one with a domain they own, and which extends the expected domain name -– most likely a word before or after the original name is added (e.g. amazon-secure.com).

C Study Scenario

D Result Related Information

Sensitivity (d')				
M_{pre}	M_{0M}	M_{4M}	M_{6M}	M_{8M}
1.11	2.13	1.60	1.45	1.39
Criterion (C)				
0.12	−0.05	.0.09	−0.02	0.15

Table 1: SDT mean results for RQ1

- You are Martin Müller and have two e-mail addresses: martin.mueller.77@web.de and martin.mueller.77@posteo.de
- You speak German and English
- Your boss is Wolfgang Lange. His e-mail address is wolfgang.lange.54@web.de
- Your colleague is Jonas Schmidt. He has two e-mail addresses: jonas.schmidt.77@web.de and jonas.schmidt.77@posteo.de
- You use all the services used in this evaluation
- You use the various operating systems and programs used in this evaluation

Figure 10: Study details for role play (translated from German for the purpose of this paper).

– Continued on next page –

D.1 Results for Research Question 1

For the reporting of the results, we combine the analyses of $\boldsymbol{H}_{M_{pre}-M_{0M}}$ and $\boldsymbol{H}_{M_{pre}-M_{\Delta\bar{t}M}}$ (for $\Delta t = 4+2i$ months, where $i \in \{0,1,2\}$)

We checked for a significant difference between the corresponding five groups (see Figure 1) using a one-way ANOVA. There was statistical significance between the groups ($F(4,227) = 5.457$, $p < 0.001$) for the sensitivity (d'). For the effect size we calculated $\omega^2 = .093$, which is a medium effect size according to [42]. A LSD post-hoc showed that the sensitivity for the M_{0M} ($d' = 2.13$, $SD = 1.15$) was significantly higher than for the M_{pre} ($d' = 1.11$, $SD = 1.12$) with ($p < .001$).

The LSD post-hoc test showed that the sensitivity for the M_{4M} ($d' = 1.60$, $SD = 1.01$) was significantly higher than for the M_{pre} ($d' = 1.11$, $SD = 1.12$) with ($p = .034$). But the sensitivity for the M_{6M} ($d' = 1.46$, $SD = 1.01$) with ($p = .123$) and for the M_{8M} with ($d' = 1.39$, $SD = 1.42$) with ($p = .155$) was not significantly higher.

Note, there was no statistical significance between the groups ($p = 0.623$) for the criterion (C).

In summary: We accept $\boldsymbol{H}_{M_{pre}-0Month}$ and $\boldsymbol{H}_{M_{pre}-4Months}$.

D.2 Results for Research Question 2

First, we checked for which reminder measures the hypothesis $\boldsymbol{H}_{M_{pre}-M_{Reminder_x-6M}}$ holds.

We checked for a significant difference between M_{pre} and the four months retention groups (see Figure 1) using a one-way ANOVA. There was statistical significance between the groups ($F(5,244) = 2.410$, $p = 0.037$) for the sensitivity (d'). For the effect size we calculated $\omega^2 = .027$, which is a small effect size [42]. A LSD post-hoc showed that the sensitivity for the $M_{Reminder-Text-6M}$ ($d' = 1.61$, $SD = 1.18$) with ($p = .005$), $M_{Reminder-Video-6M}$ ($d' = 1.80$, $SD = 1.42$) with ($p = .005$) and $M_{Reminder-InteractiveExamples-6M}$ ($d' = 1.73$, $SD = 1.19$) with ($p = .007$) were significantly higher than for the M_{pre} ($d' = 1.11$, $SD = 1.12$). The sensitivity for the $M_{Reminder-ShortText-6M}$ was not significantly higher ($d' = 1.56$, $SD = 1.11$) with ($p = .075$). Note, there was statistical significance between the groups ($p = 0.013$) for the criterion (C). A LSD post-hoc showed that the criterion for the $M_{Reminder-Text-6M}$ ($C = -.23$, $SD = .59$) with ($p = .043$) and $M_{Reminder-InteractiveExamples-6M}$ ($C = -.43$, $SD = .65$) with ($p < .001$) were significantly different from the M_{pre} ($C = .12$, $SD = .84$). The criterion for the $M_{Reminder-ShortText-6M}$ ($C = .03$, $SD = .77$) with ($p = .603$) and $M_{Reminder-Video-6M}$ ($C = -.06$, $SD = .70$) with ($p = .273$) was not significantly different.

In summary: We accept $\boldsymbol{H}_{M_{pre}-M_{Reminder_x-6M}}$ for text measure, video measure, and interactive examples measure.

In order to test whether one of the three remaining reminder measures performs best, we also checked the ANOVA values for between the reminder measures. There is no significant difference between these measures. From the descriptive data, the interactive examples measure performs slightly better than the video measure (see Figure 5).

D.3 Results for Research Question 3

Based on the results from RQ2 we decided to not collect data from the short text group after 12 months. In order to address the pre-conditions from Section 5.1 we kind of extended $\boldsymbol{H}_{M_{pre}-M_{Reminder_y-12M}}$ accordingly, i.e. six measurements were considered (see Figure 2).

We linked participants using the provided codes. This resulted in 20 participants in M_{6M-12M}, 17 participants in $M_{Reminder-Text-12M}$, 17 participants in $M_{Reminder-Video-12M}$, and 12 participants in $M_{Reminder-InteractiveExamples-12M}$.

We analysed the data from participants that we could link via code. We checked for a significant difference between the corresponding six measurements. There was statistical significance between the groups ($F(5,172) = 2.721$, $p = 0.022$) for the sensitivity (d'). The LSD post-hoc showed that the sensitivity for the $M_{Reminder-Text-12M}$ ($d' = 1.93$, $SD = 1.17$) with ($p = .009$), the $M_{Reminder-Video-12M}$ ($d' = 1.77$, $SD = 1.32$) with ($p = .031$) and $M_{Reminder-InteractiveExamples-12M}$ ($d' = 1.96$, $SD = 1.34$) with ($p = .016$) were significantly higher than for the M_{pre} ($d' = 1.11$, $SD = 1.12$). The sensitivity for the M_{12M} ($d' = 1.55$, $SD = 1.06$) with ($p = .060$) was not significantly higher. The sensitivity for the M_{6M-12M} ($d' = 1.43$, $SD = 0.77$) with ($p = .256$) was not significantly higher. For the effect size we calculated $\omega^2 = .047$, which is a small effect size according to [42]. Note, there was no statistical significance between the measurements ($p = 0.274$) for the criterion (C).

Sensitivity (d')				
M_{pre}	$M_{Reminder-Text-6M}$	$M_{Reminder-Video-6M}$	$M_{Reminder-InteractiveExamples-6M}$	$M_{Reminder-ShortText-6M}$
1.11	1.61	1.73	1.80	1.56
Criterion (C)				
0.12	−0.02	−0.23	−0.43	−0.06

Table 2: SDT mean results for RQ2

Sensitivity (d')						
Measure	M_{pre}	M_{12M}	M_{6M-12M}	$M_{Reminder-Text-12M}$	$M_{Reminder-Video-12M}$	$M_{Reminder-InteractiveExamples-12M}$
Linked	1.11	1.55	1.44	1.86	1.77	1.96
Unmatched	1.11	1.55	1.60	1.54	1.73	1.69
Criterion (C)						
Linked	0.12	0.12	0.12	−0.07	−0.05	−0.38
Unmatched	0.12	0.12	0.01	−0.08	0.00	−0.02

Table 3: SDT mean results for RQ3

Measure	M_{12M}	M_{6M-12M}	$M_{Reminder-Text-12M}$	$M_{Reminder-Video-12M}$	$M_{Reminder-InteractiveExamples-12M}$
Linked	36	20	17	17	12

Table 4: Number of participants per measure for linked and Unmatched groups

Minutes	$M_{Reminder-Text-6M}$	$M_{Reminder-Video-6M}$	$M_{Reminder-InteractiveExamples-6M}$	$M_{Reminder-ShortText-6M}$
Median	9.93	9.50	7.32	3.35
Std. Deviation	7.52	3.68	12.81	15.75

Table 5: The time needed by participants

E Email Screenshots of Phishes

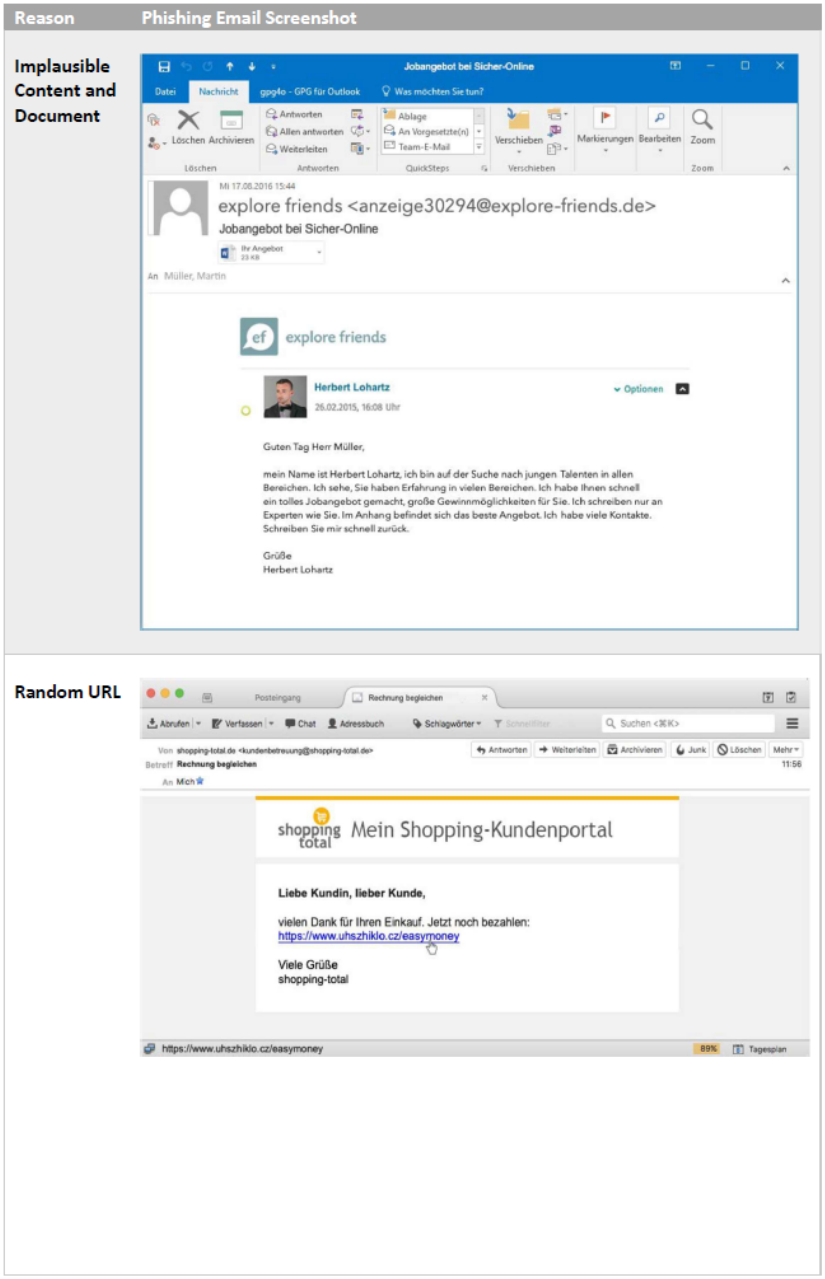


Figure 11: Phishing email screenshots (part 1)

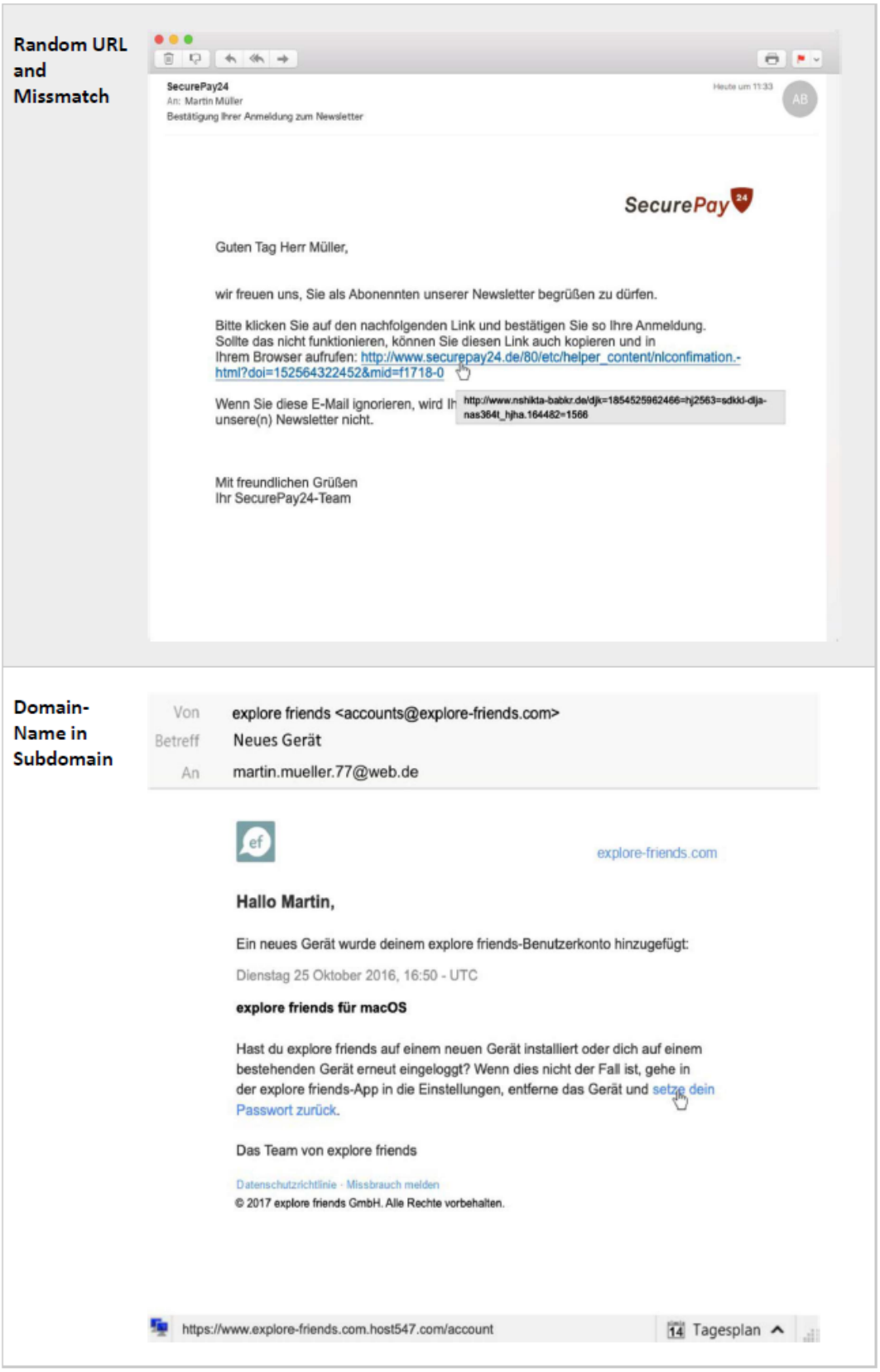


Figure 12: Phishing email screenshots (part 2)

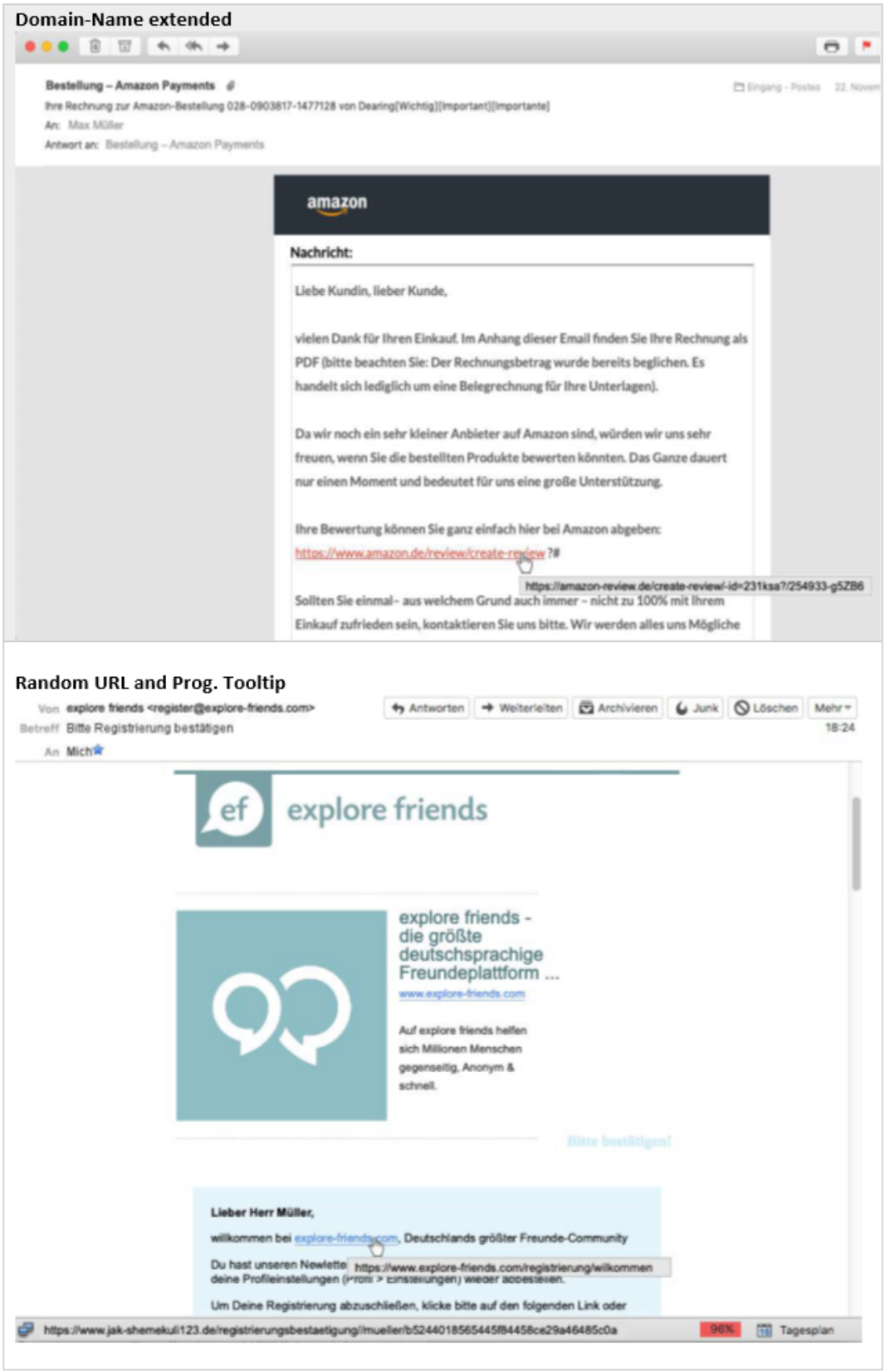


Figure 13: Phishing email screenshots (part 3)

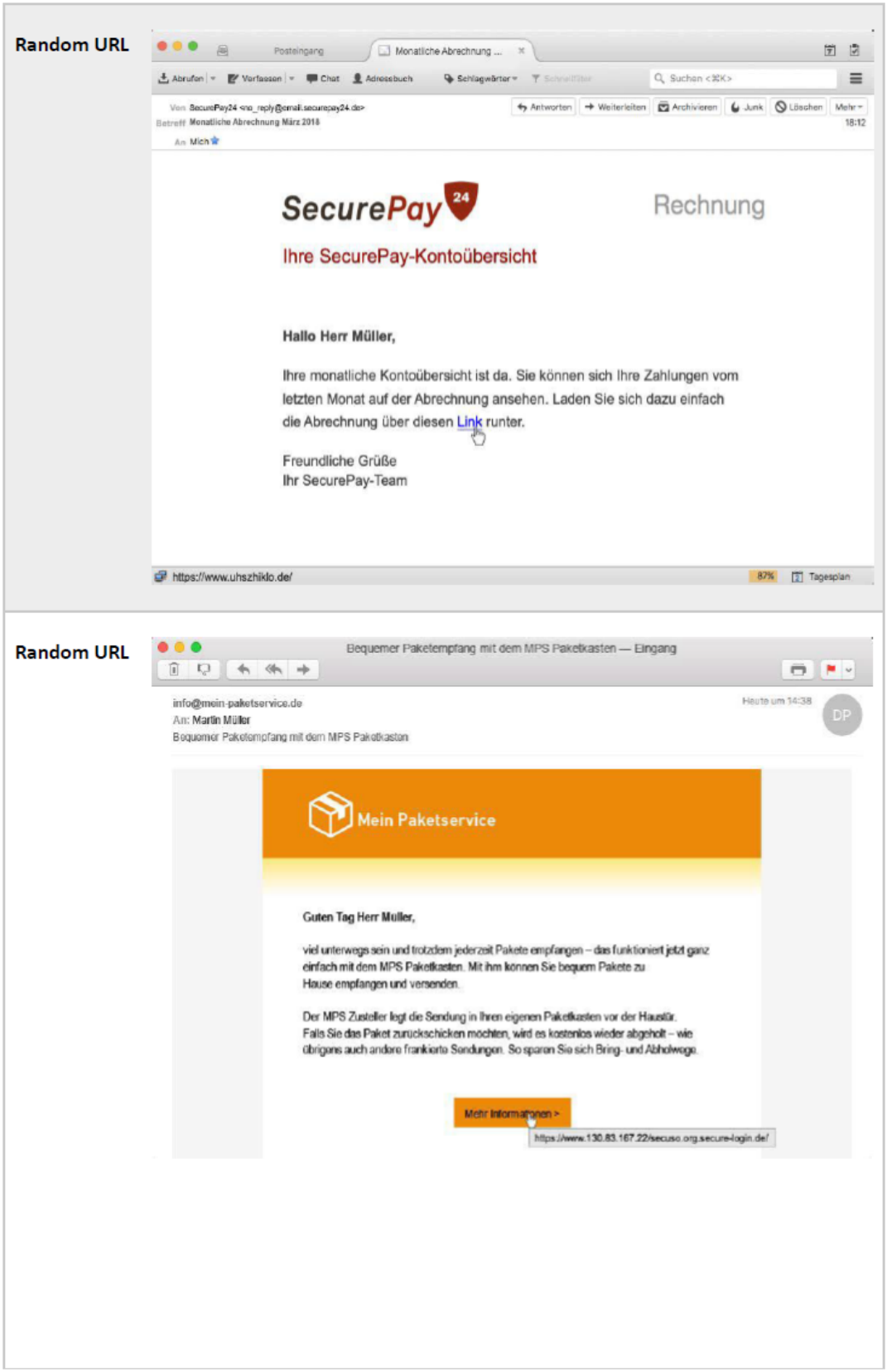


Figure 14: Phishing email screenshots (part 4)

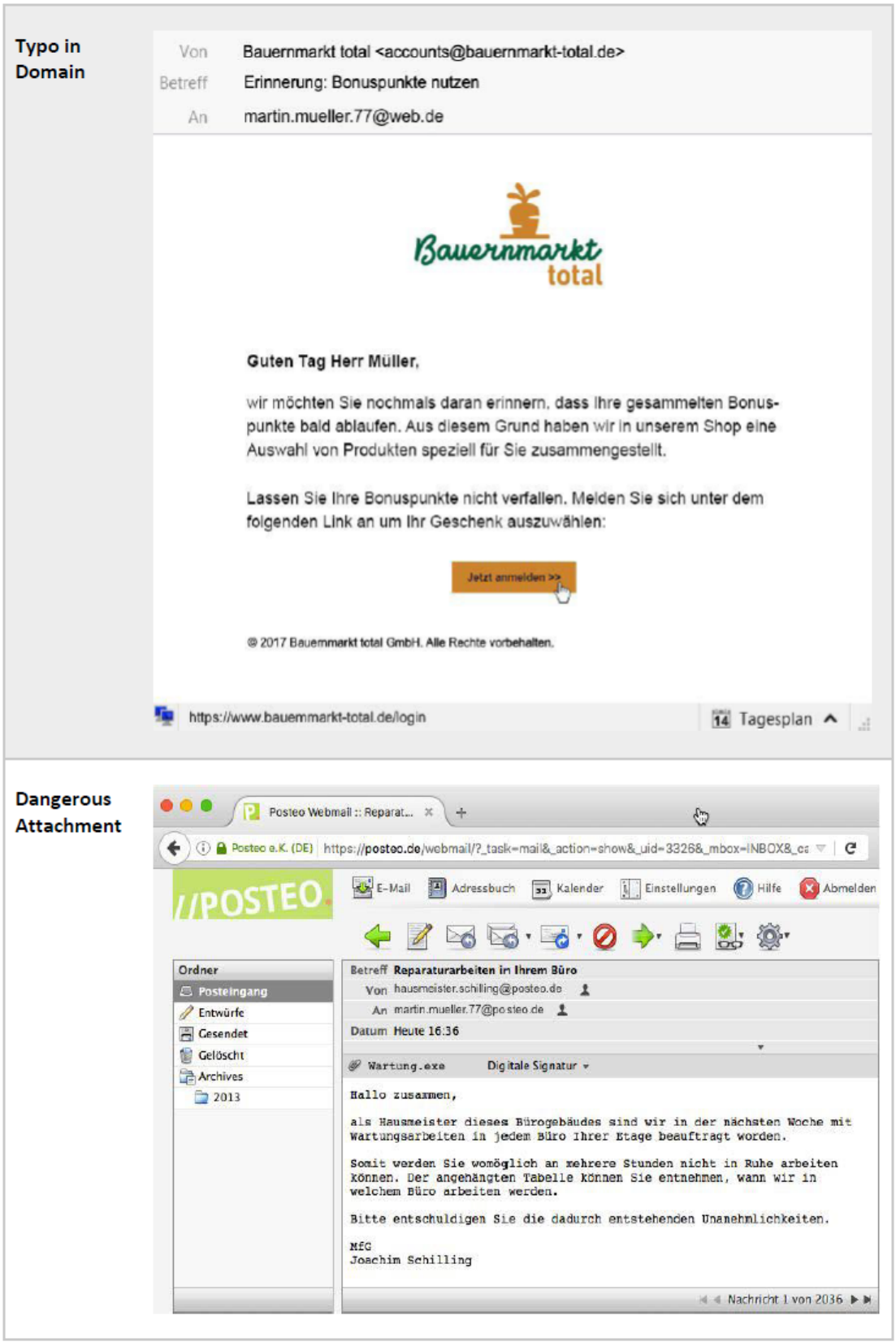


Figure 15: Phishing email screenshots (part 5)

An Exploratory Study of Hardware Reverse Engineering Technical and Cognitive Processes

Steffen Becker*†‡, Carina Wiesen*†‡, Nils Albartus†‡, Nikol Rummel†, and Christof Paar‡
†*Ruhr University Bochum;* ‡*Max Planck Institute for Cybersecurity and Privacy*
{steffen.becker, carina.wiesen, nils.albartus, nikol.rummel}@rub.de; christof.paar@csp.mpg.de
**Both authors contributed equally to this work.*

Abstract

Understanding the internals of Integrated Circuits (ICs), referred to as Hardware Reverse Engineering (HRE), is of interest to both legitimate and malicious parties. HRE is a complex process in which semi-automated steps are interwoven with human sense-making processes. Currently, little is known about the technical and cognitive processes which determine the success of HRE.

This paper performs an initial investigation on how reverse engineers solve problems, how manual and automated analysis methods interact, and which cognitive factors play a role. We present the results of an exploratory behavioral study with eight participants that was conducted after they had completed a 14-week training. We explored the validity of our findings by comparing them with the behavior (strategies applied and solution time) of an HRE expert. The participants were observed while solving a realistic HRE task. We tested cognitive abilities of our participants and collected large sets of behavioral data from log files. By comparing the least and most efficient reverse engineers, we were able to observe successful strategies. Moreover, our analyses suggest a phase model for reverse engineering, consisting of three phases. Our descriptive results further indicate that the cognitive factor Working Memory (WM) might play a role in efficiently solving HRE problems. Our exploratory study builds the foundation for future research in this topic and outlines ideas for designing cognitively difficult countermeasures ("cognitive obfuscation") against HRE.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

1 Introduction

By definition, every computing system is based on hardware components, in particular on Integrated Circuits (ICs). Their internals are typically completely opaque to the user and largely even to the developers of the system. Understanding the internals of (digital) hardware components, which requires Hardware Reverse Engineering (HRE), is of interest for both malicious and legitimate reasons [27]. For instance, the sensitive topic of low-level backdoors (i.e., hardware Trojans), which underlies the current discussion about foreign-built communication and computer equipment [29, 30], requires HRE for detection of such manipulations. On the other hand, adversaries might also need to reverse engineer the hardware they plan to subvert. HRE is also widely-used in practice for detection of Intellectual Property (IP) infringements [13]. On the adversary side, a malicious party needs to reverse areas of interest or even entire ICs. Moreover, there is a host of Trojan detection techniques [33] that require a flawless model of the target IC.

Despite its relevance, HRE is relatively poorly understood compared to many other areas of both hardware design and computer security [13]. We argue that it is desirable to obtain a better understanding of HRE. First, it will aid with assessing the threat posed by adversaries performing HRE. This is particularly prudent because there is undoubtedly expertise about HRE within government agencies with malicious intent [20]. Second, HRE performed for defensive purposes such as IC verification will benefit from having a better understanding of the involved time and costs. Third, having a better grasp on HRE will allow us to derive sound obfuscation techniques that exploit cognitive limitations of humans.

HRE is a multilayered process, where high-level information is extracted from a low-level circuit, consisting of two major stages [1]: In the first stage, a gate-level netlist is obtained from an IC either directly from the device or possibly through (online) interception of design information. A gate-level netlist is a logical circuit description and is usually composed of Boolean gates and their respective interconnections.

The technical steps required for netlist extraction, including decapsulation and imaging, are relatively well-covered in the literature [27, 35]. The objective of the second stage is to *make sense* of the recovered netlist, that is, to *understand* the netlist [1]. This second stage has only been rudimentarily addressed in the literature and is the topic of our investigation.

In this contribution, we present an exploratory study with the goal of obtaining initial insights into the complex processes of the sense-making part of HRE. This undertaking is particularly challenging as it depends on non-trivial technical steps as well as on cognitive factors of the human reverse engineer. To the best of our knowledge, it is the first time such a study based on observing a group of reverse engineers has been conducted.

Overview on our Exploratory Study

In order to gain such insights, we conducted an empirical study that explores both the technical and the underlying cognitive processes of hardware reverse engineers. In an ideal scenario, we would examine expert reverse engineers working on real-world tasks with tools they are familiar with. However, we face the methodological problem that those experts are few to begin with. They are primarily active in government agencies and a few highly specialized companies, and are generally unavailable to the scientific community. We approach this problem by observing students enrolled in a 5-years BSc-MSc program in cyber security while solving a realistic HRE task, which was developed in collaboration with HRE experts. Prior to the study, the students had been exposed to an extensive 14-week HRE training. For the study, we selected eight students based on their performance during the HRE training.

Despite the difficulty of engaging HRE experts as study participants, we were able to recruit one expert via the professional network of one of the authors. This expert solved the same task as our participants. The collected data served as a sanity check for our student sample with respect to solution time, as well as phases of and applied strategies during HRE.

In our exploratory study, we collected detailed behavioral data during an HRE-based attack and measured cognitive factors of eight participants. The HRE task is based on a realistic setting, where the analyst has to circumvent an IP protection mechanism in an unknown circuit. Analysis of the data revealed initial insights into problem solving strategies and relevant cognitive factors in HRE. We were also able to derive first hypotheses for a novel class of obfuscation measures that take the boundaries of cognitive abilities into account. Although our work was faced with methodological challenges that we discuss in the limitations section, we render the following main contributions advancing the current scientific knowledge of technical and cognitive processes in HRE.

1) We propose and examine an HRE phase model based on an exploratory behavioral study of human reverse engineers, who solved a realistic HRE task.
2) Based on behavioral analyses we explore more and less efficient HRE strategies of reverse engineers.
3) We explore the role of different cognitive factors in efficiently solving HRE problems.
4) Based on our findings, we derive hypotheses for a novel class of HRE countermeasures called cognitive obfuscation and outline future research directions.

2 Background

In this section, we present the relevant technical background of HRE and propose a three-phase model encompassing human processes during HRE. Furthermore, we introduce related work on cognitive processes in reverse engineering. Against this background, we identify the research gap and derive research questions we seek to answer in this work.

2.1 Hardware Reverse Engineering

Reverse engineering is the process of extracting knowledge or design information from anything man-made in order to comprehend its inner structure [28]. As mentioned above, in the case of HRE, there are two distinct stages [1]. In the first stage, a gate-level netlist is obtained directly from an IC or a Field Programmable Gate Array (FPGA) or through (online) interception of design information. Although netlist extraction requires sophisticated technical methods, research has shown that netlists can be extracted reliably by trained specialists from both, ICs and FPGAs [9, 23, 35].

In the second, sense-making stage of HRE, the netlist is transformed into higher levels of abstraction that enable a detailed analysis. This often involves module recognition, identification of blocks of interest, and detailed understanding of Boolean sub-circuits [1, 14, 32]. The analysis typically serves a specific objective, for example, finding and understanding IP blocks or extraction of cryptographic keys [39]. Due to the nature of this stage — which requires human ingenuity, sense-making, and in many cases customized solutions — fully automated tools do not exist [13]. Instead, the analyst typically employs HRE tools which enable interaction with the target netlist. Tools may provide semi-automated support for the human analyst, for example, for running specific algorithms on the netlist [8], as well as, features for manual analysis of netlist components [38].

Even with tool support, the cognitive processes and human problem-solving strategies are crucial for HRE, yet remain poorly understood. Against this background, it is hardly surprising that hardware obfuscation, which is a widely used countermeasure against HRE, is largely based on ad-hoc methods [41]. We argue that a comprehension of human processes in HRE will open a pathway for the development of novel obfuscation techniques. This paper will conclude with first guidelines for how such cognitive obfuscation measures might look like.

Reverse engineering of large and complex netlists is commonly driven by an objective more narrow than full understanding of the entire netlist. We model a situation under which HRE is typically performed in practice through the following conditions:

- An (error-free) netlist of the entire design or the area of interest is at hand.
- An HRE tool is available that allows interaction with the target netlist.
- There is a clear objective, for example, removal of an IP protection mechanism.
- (First) hypotheses related to the objective exist, for example, which IP protection mechanism is implemented.

If these preconditions are met, a human analyst attempts to understand the target netlist through two principal means:

- **Manual analyses**: Detailed manual inspection of netlist components and explorative navigation through textual and graphical representations of the netlist.
- **Semi-automated analyses**: Customized scripts and programs that facilitate structural and functional analyses of the netlist.

In Section 4, we will explore which role these two mechanisms play during HRE.

2.2 Phase Model for Gate-level Netlist Reverse Engineering

Due to the sheer complexity of the sense-making stage, it is plausible that human strategies during this process can be divided into sub phases of human sense-making. Thus, we propose a three-phase model derived from several HRE-based attacks from literature [1, 14], and two technical HRE workflow descriptions [8, 32]. Even though not explicitly formulated in a model before, the three phases are an implicit hypothesis about the inner workings of HRE. In a very recent work, Votipka et al. introduce a similar model of human processes during software reverse engineering based on expert interviews [37]. In the following passages, we briefly characterize each phase under the assumption that the preconditions described above are fulfilled. Also, since netlist reverse engineering is a continuous process, the consecutive phases can blend into each other rather than being strictly disjoint.

Phase 1: Candidate Identification

The goal of Phase 1 is to identify single candidates and sub-circuits which are potentially relevant in the context of the reverse engineering process. Therefore, the analyst starts exploration via structural analyses of the netlist topology with the goal to identify blocks of interests. The result are sub-circuit candidates, which are further inspected in Phase 2. Suited methods for this phase are semi-automated structural analyses (e.g., graph clustering algorithms or sub-circuit matching) as well as custom-tailored structural analyses incorporating hypotheses about searched structures. Additionally, manual netlist exploration can provide a starting point for the reverse engineer, for example, by inspecting global in- and outputs.

Phase 2: Candidate Verification

The goal of Phase 2 is the verification of extracted candidates from Phase 1 in order to narrow them down and select target components for Phase 3. If no target components are remaining, new candidates have to be identified in Phase 1 iteratively by refining the methods used. Phase 2 incorporates static analyses methods such as Boolean functionality analysis or sub-circuit matching. Manual inspection of candidates can support the reverse engineer by testing the existing hypotheses before solving the problem algorithmically.

Phase 3: Realization

While Phases 1 and 2 can be generalized for most HRE tasks, the goals and procedures of Phase 3 differ significantly depending on the objective. Methods employed include sub-circuit interpretation and annotation in order to obtain a more abstract netlist model; netlist simulation to analyze sequential behavior; or preparation of malicious netlist manipulation, for example, add, remove, or change functionality of netlist components. In many cases, Phase 3 incorporates and combines several of the aforementioned methods.

2.3 Cognitive Processes in Reverse Engineering

As outlined above, HRE always involves sense-making processes. Thus, the success of HRE heavily depends on skills, knowledge, and expertise of the performing reverse engineer. Surprisingly, underlying cognitive processes in HRE are understudied and remain poorly understood [13]. Despite this observation, prior research on HRE almost solely focuses on technical factors. Nonetheless, one prior work explored cognitive processes by defining reverse engineering of Boolean systems as a specific type of human problem solving [19]. In general, a problem exists when a person lacks in knowledge which enables the problem solver to achieve a desired goal [12]. Problem solving is defined as a sequence of cognitive operations (e.g., problem solving strategies) in order to solve a task for which the individual does not possess a suitable routine operation [25]. The success of problem solving is influenced by several factors, for example, prior domain specific knowledge [5], or cognitive abilities (e.g., intelligence and sub factors like Working Memory) [17]. Baddeley demonstrated that brain systems like the Working Memory are essential in solving complex cognitive tasks like problem solving [2]. Besides cognitive abilities, the level of expertise [7] is another important factor in determining problem solving performances. Larkin et al. showed that experts were quicker in solving physics problems than novices [18].

In 2013, Lee and Johnson-Laird analyzed problem solving behavior in reverse engineering of Boolean systems by

conducting five experiments in a laboratory setting [19]. The tasks students were asked to solve merely involved drawing Boolean circuits controlling an electric light. Consequently, the ecological and external validity of these experiments appears low in the context of HRE. Thus, it remains unclear to what extent the results of Lee and Johnson-Laird can be generalized to reverse engineering of entire ICs, which commonly consist of hundred of thousands or millions of logic components. Nevertheless, we transfer human problem solving processes to the cognitive processes in HRE by observing problem solving strategies of more and less successful reverse engineers and by measuring cognitive factors which might play a role in HRE problem solving.

2.4 Research Gap and Research Questions

In this work, we aim to close the existing research gap by providing first insights into the technical and cognitive processes during a realistic HRE task. We qualitatively examine the occurrence of the 3-phase model for netlist reverse engineering in the participants' behavior, and investigate HRE problem solving strategies and their influencing cognitive factors on the basis of behavioral and cognitive data. This allows us to derive hypotheses for cognitive obfuscation techniques, i.e., novel countermeasures impeding HRE. The paper at hand attempts to answer the following research questions:

RQ1. Can the phases of human sense-making be detected during HRE processes? If so, which are the crucial phases?

RQ2. Which strategies distinguish more and less efficient reverse engineers?

RQ3. Which cognitive prerequisites play a role for the success of HRE?

RQ4. How can those insights be used to derive hypotheses for cognitive obfuscation?

Our exploratory study opens many venues for further in-depth research on the challenging interdisciplinary problem of understanding HRE.

3 Methodology

In the following section, we first describe the research environment enabling our study. Second, we outline the details of our user study, including our participants and all study related measures and processes. Last, we explain the behavioral analyses providing the underlying data for our exploration of human factors in HRE.

3.1 Research Environment

This section outlines the research environment consisting of an extensive HRE training, the HRE tool HAL, and a practical HRE task under study.

3.1.1 HRE Training

The contents of the 14-week HRE training were developed based on the comprehensive body of technical research and industry practices and on educational guidelines facilitating learning HRE with input from experts from academia and practice [43]. Subsequently, it was shown that the training successfully promotes HRE skill acquisition [42]. During the first six weeks, the instructors conveyed necessary theoretical backgrounds, before students solved four training tasks with the HRE tool HAL in the 8-week practical part. The training was followed by a two-week study, where participants solved a realistic HRE task. Crucially, neither the proposed HRE phase model, nor the concrete scenario of, or any solution strategies for the Study Task were part of the training.

3.1.2 HRE Tool HAL

Given today's integration density, which commonly results in very complex netlists, it is virtually impossible to reverse even moderately sized IC or FPGA designs without tool support. Therefore, it is safe to assume that professional reverse engineers, for example, in government agencies and specialized companies, have access to such (internally developed) tools, cf. [20, 34]. Recently, the open-source framework HAL [15, 38] has become available on GitHub, which is the first tool specifically designed to facilitate HRE. By using HAL in our study, we create an environment that is, thus, similar to one encountered in real-world HRE situations.

HAL operates on gate-level netlists. It has a modular and extendable design and supports both, static analysis and cycle-accurate simulation of netlists. There are several references describing semi-automated reverse engineering and manipulation tasks using HAL [14, 15, 39]. Crucially for our study, HAL can capture all user interactions and therefore enables the investigation of the many sub-steps that take place during HRE, including the study of human factors. Below is a description of HAL features that were particularly useful for our study:

- HAL offers an *interactive GUI* allowing detailed manual inspection and exploration of netlist components as well as module grouping functionalities.
- HAL natively implements a *Python interface* enabling script-based interactions with the netlist.
- HAL comes with a variety of *accompanying materials*, most importantly a detailed documentation of HAL-specific Python commands and a coding guide providing many examples for common use cases.

3.1.3 Tasks and Materials

In the practical part of the HRE training and in the study, participants worked on five reverse engineering tasks based on flat netlists synthesized for FPGAs. Those netlists do not contain any high-level information about the design such as component names, module boundaries or hierarchy elements. Netlist components are composed of Look-up tables (LUTs)

with up to six inputs, which realize the circuit's Boolean functions, multiplexers, Flip Flops (FFs), and their respective logic interconnections.

All five tasks are based on ideas drawn from recent literature on HRE attacks and countermeasures [1, 6, 14]. Thus, they represent a number of different HRE settings ranging from sub-circuit detection over obfuscation circumvention up to malicious manipulation of cryptographic cores. The tasks are designed with increasing difficulty, that is, the problem itself increased in complexity, the level of guidance decreased, and the size and complexity of the target netlist increased, cf. Table 1. The level of guidance includes task-related support by instructors, which participants received only during the training tasks, as well as the amounts of accompanying materials, e.g., (excerpts of) scientific papers, or relevant examples from the coding guide for the task.

Table 1: Difficulty, level of guidance, and netlist complexity for the HRE tasks on a scale from **+** (low) to **+++** (high).

	Difficulty	**Guidance**	**Complexity**
Training Task 1	+	+++	+
Training Task 2	++	+++	+
Training Task 3	++	++	+
Training Task 4	+++	++	++
Study Task	+++	+	++

In the following, the last and most difficult task, which provides the data for our study, is introduced.

Study Task: Breaking Watermarkings. In this task, the IP protection mechanism—a so-called watermarking scheme—proposed by Schmid et al. [31] is embedded in a hardware design, enabling the detection of IP theft. The analyst has the objective to clone the underlying netlist without copying its watermark, and therefore thwarting detection of IP theft in the cloned circuit. While materials such as the original paper on the functionality of the watermarking scheme were provided, the analysts had to develop the strategies for detection and removal of the watermarking themselves.

To enable analysis of HRE strategies employed by the participants while solving the Study Task, netlist components are pre-annotated in relevant and irrelevant ones. Out of 4653 total netlist components, 54 were marked as relevant since they are associated with the watermark. 50 of those components are *candidates*, thereof 30 actually implement the watermark and are therefore *targets*. The remaining 4 components are important anchor points for the watermark detection. Naturally, the annotation was invisible to the participants. All netlist components have a unique identifier in HAL, which allows their tracing during the behavioral analyses described in Section 3.3.

3.2 User Study and Variables

This section describes our user study in detail. This includes the specification of our study participants and the expert, ethical considerations concerning this research, as well as an outline of the study procedures and variables.

3.2.1 Study Participants

We conducted our quasi-experimental study with a within-subject design, where every participant had to solve the Study Task using HAL. Since experts were generally unavailable for this research, we approximated experienced reverse engineers by recruiting senior-level and MSc-level students with a (target) degree in cyber security. The recruited participants acquired relevant HRE knowledge and skills during the extensive training phase and were selected based on their performance in the four training tasks described in Section 3.1.1. Furthermore, we approximated the situation of experienced reverse engineers by providing a coding guide containing relevant code snippets, instructions, and programming methods which were used throughout the study and the training tasks.

Overall, 22 participants volunteered to participate in the study. All participants were enrolled in either the last year of a three-year bachelor's or in a master's program in cyber security. Five of the original 22 participants decided to drop out. Furthermore, three participants were excluded due to incomplete datasets. Out of the remaining 14, eight participants (mean age: 24 years; $SD = 4$ years; one graduate student) were selected based on their performance in the training tasks, i.e., the average solution probability of all training tasks was above 95% per selected participant. Solution probabilities were evaluated on a scale from 0% to 100% by three teaching assistants based on a detailed gradebook with sample solutions.

3.2.2 Sanity Check with Expert

Despite the difficulty of recruiting HRE experts as study participants, we were able to recruit one expert through the professional network of one of the authors. The expert also performed the Study Task. There was a two-fold objective to engaging the expert. First, the comparison of the reversing behavior between the expert and the study participants allowed us to explore if the students' approximated level as experienced reverse engineers was an adequate assumption. Second, we could evaluate the difficulty of the Study Task. For assessing the status as HRE expert, we followed the criteria of Votipka et al. [37]: Our expert had 5 years of experience and a self-assessed skill-level of 4 on a 5-point Likert-Scale (with 1 being a beginner and 5 being an expert). The expert gave written informed consent on using the collected socio-demographic and behavioral data in the context of this research project. Cognitive abilities were not tested for the expert to protect the person's privacy in case of a potential de-anonymization.

The expert received the underlying netlist of the Study Task and corresponding materials (virtual machine running HAL, coding guide, documentation of Python commands, and paper on the implemented watermarking) and was asked to remove the IP protection mechanism from the circuit. Moreover, the expert was already experienced in working with HAL. Subsequently, the behavioral data collected while the expert was working on the Study Task was analyzed in order to compare the observations made for our student sample with respect to solution time, as well as phases of and applied strategies during HRE with the expert's problem solving behavior.

3.2.3 Control Variables

We asked participants to provide information about their socio-demographics (age, major, target degree, number of semesters enrolled) in a self-developed questionnaire (cf. Appendix A).

3.2.4 Cognitive Abilities

Based on the definition of reverse engineering as a specific type of problem solving [19], we transferred the concept of problem solving research into the domain of HRE. Therefore, as problem solving performances can depend on cognitive abilities [17], we measured sub factors of general intelligence and their correlations to HRE problem solving performances. As a measure for problem solving performance, we correlate the variable time on task, which is a traditional measure in cognitive psychology, with levels of cognitive abilities. In order to assess the participants' levels of cognitive abilities, sub tests of a valid test instrument, the Wechsler Adult Intelligence Scale (WAIS-IV) [40], were used. We integrated the following three sub scores in our study: Perceptual Reasoning (PR), Working Memory (WM), and Processing Speed (PS). The fourth test of the WAIS-IV on Verbal Comprehension (VC), which measures verbal reasoning and verbal expression, was not included in the study since participants had different native languages. PR measures the ability to accurately interpret and work with visual information. The sub score WM assessed the ability to store information and to perform mental operations on that stored information. The third score PS reflected the ability of processing visual information quickly and efficiently.

3.2.5 Ethical Considerations

Our institute does not have an ethics board or IRB, but the study protocols were reviewed and approved by the universities' data protection officer. Before entering the HRE training, all 22 participants gave written informed consent. They received monetary compensation well above minimum wage levels for time spent on materials related to our study. We informed participants that they can withdraw from our study at any time and that all partial data will not be analyzed or stored. Privacy was ensured by randomly assigning pseudonyms to the participants, which were used instead of their actual names throughout all materials related to our study.

3.2.6 Study Procedures

The WAIS-IV was conducted in a 60 to 90 minutes face-to-face session with each participant prior to the study. Before starting the Study Task, participants were asked to answer the questionnaire on socio-demographics via the online survey provider Soscisurvey. After finishing the Study Task, participants uploaded log files recorded while solving the Study Task onto a server located at the university.

3.3 Behavioral Analyses

In order to explore human processes in HRE, we analyzed log files automatically generated by HAL. We collected one log file — containing several thousands up to tens of thousands of log entries — for each participant solving the Study Task. Every log entry consists of a timestamp and one of the following events in HAL: (i) content and terminal output of the executed Python script; (ii) manual selection of netlist component (unique identifier) via GUI; or (iii) additional system-level entries such as indications for user (in)activity. The following parameters were extracted from the collected logfiles and serve as the basis for the behavioral analyses.

Total Solution Time. Since participants solved the task over a period of two weeks, relative solution time was calculated, where loading the inspected netlist for the first time in HAL marks the starting time. All periods lasting more than 10 minutes without log entries were manually reviewed. Based on observed wall clock time and script changes during those periods, a threshold of 60 minutes was identified for periods of inactivity, which were subsequently excluded from time on task (cf. Table 2 in Appendix B).

Executed Scripts. Executed scripts were saved as standalone versions together with their associated execution time and accessed netlist components. Each script execution represents an intermediate state of an iteratively developed solution script.

Manual Component Selections. Manual selections of netlist components together with the associated time were extracted.

Time per Phase. The phases of human sense-making as described in Section 2.2 were detected by checking the following conditions for the completion of each phase:
Phase 1 is completed when all 50 candidates are identified in the netlist. Thus, they have to be accessed by a script.
Phase 2 is finalized when all 30 targets which actually implement a watermark are identified and the watermarking signatures are read out. Therefore, data from those components has to be extracted and processed via script.
Phase 3 is finished after removing all watermark signatures from the circuit. This implies that all 30 targets have to be manipulated. Consequently, the script solving Phase 3 has to conduct a manipulation effectively removing the watermarking on those targets.

Fulfillment of those conditions was checked through a combination of automated script analysis, for example, when and

how relevant components are accessed, as well as further manual reviews of scripts. Manual reviews were conducted collaboratively by two researchers familiar with the task under study.

Progress Metric. As described in Section 2.1, reverse engineers use the two principal actions of semi-automated scripts and manual analysis. We assessed progress during the Study Task with respect to both actions.

Script Progress Score. Regarding semi-automated reverse engineering, we employed automated script analysis, e.g, the observation of persistent lines of code with respect to the final solution script, and further manual reviews of single script iterations. For each executed script, a progress score between 0 and 3 was assigned according to the rules below:

0: no progress made: no or only obsolete code added,
1: few relevant line(s) of code implemented; small subproblem solved,
2: relevant block of code or single significant line of code implemented; important subproblem solved,
3: significant idea or significant block of code implemented.

Manual Progress Score. With respect to manual analysis, groups of manual interactions were assessed on the same scale by first pooling consecutive component selections and then allocating the following progress scores between 0 and 3:

0: selection of irrelevant component(s),
1: repeated selection of a single relevant component,
2: repeated selection of multiple relevant components,
3: first selection of relevant component(s).

Progress Visualization. To enable visual representations of progress during HRE processes, we allocated weights to the progress scores. Progress Scores 0 and 1 were weighted as is, while Scores 2 and 3, which represent the main progress and occur only sparsely, were allocated a weight of 3 respectively 7. Those weighted scores served as foundation for the graphical progress visualization as follows: For each participant and phase, all assigned weighted scores were summed up. Progress made is represented as the fraction of the weighted progress score divided by the aforementioned sum.

4 Results

In the following section, we briefly illustrate the results of the expert and classify them in relation to the results of our eight participants before we examine the behavioral data and cognitive abilities of the participants with respect to the research questions formulated in Section 2.4. Section 4.2 provides insights for RQ1 by analyzing occurrence and relevance of phases from the three-phase model (cf. Section 2.2). With regard to RQ2, the respective strategies of the overall most and least efficient participants are evaluated in Section 4.3. Section 4.4 presents first indications about cognitive factors and HRE regarding RQ3.

4.1 Sanity Check with HRE Expert

The total solution time of the recruited expert for correctly solving the Study Task in HAL was 162 minutes, which is comparable to the fastest participant of our study (169 minutes). The occurrence of all three phases could clearly be detected via analysis of the behavioral data as described in Section 3.3 (cf. Figure 1). This is notable because the expert has an entirely different training background than the study participants. In Phase 1, the expert identified the watermark candidates and saved them in a data structure, before writing a function to successfully extract and functionally verify the watermark signatures, therefore identifying the targets. In the last phase, the expert removed the watermark signatures and generated the watermark-free netlist.

With 164 executed scripts, the expert lies within the upper range of our participants, who executed between 62 and 175 scripts (cf. Appendix B, Table 3). Also, the number of 204 manual component selections is comparable to our participants' (cf. Appendix B, Table 4).

Furthermore, the in-depth analysis of HRE strategies revealed that the expert applied strategies similar to those of the fastest participant (e.g., divide-and-conquer), and only rarely ran into periods of stagnation (cf. Figure 3 in Appendix B).

In summary, the expert's results and behavior are in line with the study participants, and therefore allow the assumption of classifying them as experienced reverse engineers. At the same time, the HRE expert solved the Study Task in a time comparable to the participants' time on task, which indicates the adequate difficulty of the task.

4.2 Empirical Observation of Phases (RQ1)

All participants were able to solve the Study Task correctly as indicated by solution probabilities between 97% and 100%. Based on the behavioral analysis described in Section 3.3, the occurrence of the three phases proposed in Section 2.2 could be observed for every participant. First, participants saved the watermark candidates in (different) data structures in Phase 1, before iterating over those candidates to apply their specific methods of functional verification to identify the targets in Phase 2. Only after successful verification, the participants began to remove the watermarking and to subsequently generate the watermark-free netlist in Phase 3. In the following passages, we report our results regarding spent time, executed scripts, and manually inspected netlist components per phase.

4.2.1 Solution Time per Phase

Figure 1 shows the time each participant required for solving the Study Task, together with a break down of the times they spent on each of the three phases. On average, participants spent 254 minutes ($SD = 74$ minutes) on task. We also calculated the group average of relative time spent on each of the three phases, by averaging the relative times of all eight participants: On average, they spent 12% ($SD = 4\%$) of their

time in Phase 1, Phase 2 consumes 58% ($SD = 11\%$), and Phase 3 took 30% ($SD = 10\%$) of participants' time solving the task.

4.2.2 Executed Scripts per Phase

On average, participants executed 117 scripts ($SD = 32$) while solving the Study Task. We calculated the relative number of executed scripts per phase by averaging relative number of scripts per participant: 15% ($SD = 9\%$) of scripts were executed in Phase 1, 55% ($SD = 13\%$) in Phase 2, and 30% ($SD = 12\%$) in Phase 3. The number of executed scripts per phase is represented in Appendix B, Table 3.

4.2.3 Manual Component Inspections per Phase

On average, participants manually inspected 123 ($SD = 82$) netlist components while working on the Study Task. We calculated relative numbers of inspected components per phase by averaging relative numbers per participant. Participants conducted 48% ($SD = 35\%$) of their manual inspections in Phase 1, 42% ($SD = 38\%$) in Phase 2, and 10% ($SD = 8\%$) in Phase 3. Table 4 in Appendix B shows detailed statistics on manual component inspections for all participants.

4.3 HRE Strategies (RQ2)

In this section, we perform an in-depth exploration of HRE strategies per phase for Participants 1 (P1) and 8 (P8), the least and most efficient overall participants with respect to solution time. While P1 required a total time of 398 minutes, P8 solved the task in 169 minutes. Both participants spent a similar amount of relative time in Phases 1 to 3 (cf. Figure 1), and had comparable total amounts of manual netlist interactions (P1: 127, P8: 90). P8 executed the least (62) and P1 the third-most (132) scripts. We want to illustrate that this was not a problem of poor programming skills: Whereas only 47% of the scripts written by P8 were syntactically correct, P1 executed 73% syntactically correct scripts. A detailed overview of the assigned progress scores per phase as introduced in Section 3.3 is shown in Appendix B, Table 5. In the case of executed scripts, a progress score of 0 — which indicates stagnation — dominates all phases but Phase 1 of P8. Manual analyses have crucial impact (score of 3) on progress in Phase 1 for both participants, and on Phase 2 for P8. Overall, script-based progress dominates progress made through manual analyses. A visualization of participants' P1 and P8 progress over time is shown in Figure 3 in Appendix B. The figure visually complements the exploration of strategies applied by P1 and P8 in Phases 1, 2 and 3 as presented below.

4.3.1 Phase 1 (Candidate Identification)

P1 required 54 minutes to solve Phase 1, while P8 identified the correct candidates in 20 minutes. Both participants started with manual analysis and were able to achieve initial progress by selecting relevant components: P1 detected the first relevant components after 7 minutes, and P8 found the first relevant components after 4 minutes. After initial progress had been made, P8 implemented his concept of structural candidate identification in the very first script (minute 18), which then only needed minor adjustments in order to finish Phase 1. P1 similarly implemented his crucial idea for Phase 1 in the first script (minute 26) but needed several script iterations in order to successfully identify candidates. Despite their similar solution strategies with respect to interactions of manual and script-based analyses, our manual script review revealed that P1 chose a less efficient and more complex implementation to detect the candidates as indicated by a larger search space and a deeper nesting of the algorithm.

4.3.2 Phase 2 (Candidate Verification)

Participants 1 and 8 spent 111 respectively 228 minutes in order to functionally verify the candidates. P8 made immediate progress through manual inspection of relevant components. Afterwards, P8 progressed towards the verification of candidates via the development of scripts and single manual inspections without significant periods of stagnation. Only between minutes 89 and 99, the participant could not advance the problem of data extraction from netlist components via script. In our manual reviews of scripts, we observed that P8 applied a divide-and-conquer strategy, and considered the overall objective — removing the watermark signatures — already in his approach of extracting them in Phase 2.

On the contrary, P1 started with an initial stagnation period (minute 54 to 78), where three scripts were executed without noticeable direction. His initial progress in Phase 2 was sparked through the manual inspection of relevant components, although there were still irrelevant components among the inspected. The next significant stagnation period (minute 131 to 203) concerned the same problem causing the 10-minute stagnation of P8. Another period of stagnation lasted from minute 210 to 242, where P1 inspected irrelevant components. Manual review of scripts revealed that P1 developed an efficient but complex algorithm for Phase 2 including bitwise binary processing depending on several conditions.

4.3.3 Phase 3 (Realization)

The realization of the underlying objective took 116 minutes for P1 and 37 minutes in the case of P8. Consequently, P8 moved quickly towards the solution without manual netlist inspections. In the stagnation period from minute 153 to 168, he merely tried to fix printing methods. P1 had a significant stagnation period from minute 305 to 352, where the manipulation of binary strings represented a significant hurdle. After this stagnation period, manual component inspections served as impulse for further progress. In the last stagnation period, P1 went back to the algorithm from Phase 1 to replace hard-coded data with variables.

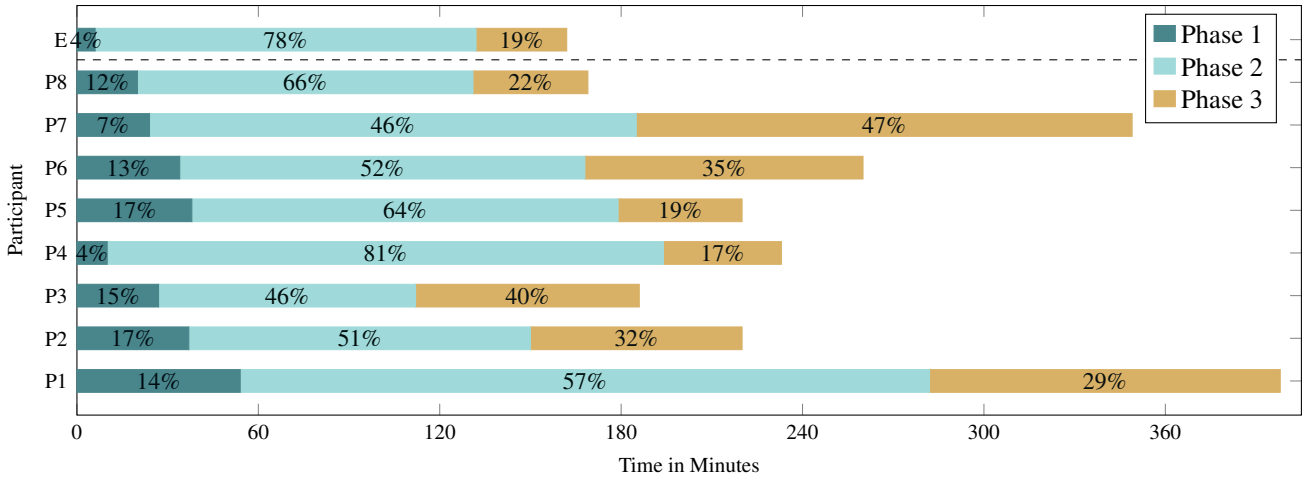


Figure 1: Absolute and relative solution times per participant and phase. For comparison, solution times of the expert are represented above the dotted line.

4.4 Cognitive Factors and HRE (RQ3)

The small sample size rendered statistical analyses of the influences of cognitive factors on HRE performance impossible. We do, however, want to briefly describe one interesting finding which can be discussed as an incentive for future research on cognitive factors in HRE. Our data suggest a potential negative correlation between Working Memory (WM) scores and the overall solution time of the Study Task. The descriptive data shows that participants with higher scores in WM tend to solve the task quicker than participants with lower WM scores. P8 had the highest WM score (126) and achieved the shortest time on task with 169 minutes. Meanwhile, the least efficient participant (P1) with a time on task of 398 minutes had a lower WM score of 108. There appears to be one outlier in the data (P6) which will be discussed in Section 5.3. The descriptive and visual analyses by scatter plots for both cognitive factors Processing Speed (PS) and Perceptual Reasoning (PR) did not show comparable results. Appendix B, Table 6 summarizes the descriptive data of cognitive factors and solution time.

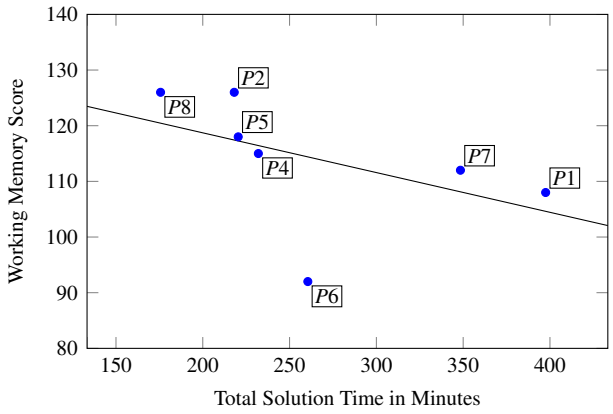


Figure 2: Scatter plot of solution time (x-axis) and Working Memory scores (y-axis) with trend line. P3 did not participate in the cognitive tests.

5 Discussion

In this section, we discuss implications of the results presented in Section 4 with respect to our four research questions.

5.1 HRE Phase Model (RQ1)

We were able to detect the proposed three-phase model for netlist reverse engineering for all participants of our study (cf. Figure 1). Thus, we suggest that passing through all phases in their respective order is essential to solve the underlying task. Our data revealed differences in absolute and relative times spent per phase, implicating their respective levels of difficulty. Moreover, we observed HRE as an interwoven process of manual and script-based netlist interactions. The observed proportions of manual component inspections per phase imply a high relevance of visual inspections for Phases 1 and 2. Even though amounts of manual interactions varied between participants as indicated by high standard deviations, we could observe purposeful usage for every participant. In the following, we briefly discuss our results with respect to each phase.

5.1.1 Phase 1 (Candidate Identification)

The relatively short solution times observed in this phase indicate that the identification of candidates, that is, relevant components, could be solved efficiently via structural analyses as described in Section 2.2. Since most manual netlist interactions were conducted in Phase 1, we deduce that they play an important role in order to detect and inspect starting

points (e.g., global in- and outputs, or nets with a great number of sinks) for reverse engineering. The identification of candidates in Phase 1 is a necessary precondition to conduct targeted candidate verification in Phase 2. It is therefore — despite its relatively short duration as observed in this exploratory study — crucial for the overall HRE process.

5.1.2 Phase 2 (Candidate Verification)

The significant portion of total solution time spent on Phase 2 implies that candidate verification was the most challenging subtask for our participants. In order to verify candidates, participants developed algorithms which functionally analyzed single candidates. We assume that the development of such an algorithm is a challenging problem. The results suggest that manual analyses in Phase 2 supported most participants in verifying candidates and developing their algorithms. A sound solution of Phase 2 is very important as preparation for the specific HRE objectives realized in Phase 3.

5.1.3 Phase 3 (Realization)

In Phase 3, reverse engineers conduct goal-oriented actions on previously identified targets by means such as interpretation, manipulation, simulation, or a combination thereof. In our case, watermarks had to be removed from the netlist via manipulation. Although netlist manipulation was an essential part of the training tasks, it was difficult to transfer those skills to the application of removing watermarks as indicated by a portion of 30% of total solution time. Participants applied manual interactions only sparsely in Phase 3, suggesting that they have less impact here compared to Phases 1 and 2.

5.2 HRE Strategies (RQ2)

In this section we discuss our results from the in-depth exploration of HRE strategies of the most and least efficient participants (cf. Section 4.3).

The comparable amounts of manual netlist interactions for P1 and P8 indicate their importance for HRE regardless of the reverse engineers' efficiency. Different amounts of executed scripts between both participants are an evidence that the faster participant made more progress per script. However, our data suggests that P8 was not the better programmer than P1, implying that other factors contribute to a higher level of progress per script.

With respect to our progress score assignments (cf. Table 5), we note that scores of 0 indicating stagnation dominate the overall HRE progress. This supports the assumed high complexity of netlist reverse engineering even for medium-sized netlists as in the underlying Study Task. The observation that script-based progress dominates progress made by manual analyses implies the nature of manual analysis methods as supporting factor for HRE.

In the following passages, we discuss the progress over time of P1 and P8 with special emphasis on periods of stagnation (cf. Figure 3 in Appendix B).

5.2.1 Phase 1 (Candidate Identification)

In Phase 1, both participants followed similar strategies based on structural netlist exploration via semi-automated scripts. The fact that initial progress was made through manual component inspections by both participants further supports the indication from Section 5.1 that manual inspections play an important role at early stages of netlist reverse engineering. Longer duration of Phase 1 for P1 was caused by a more complex algorithmic approach for candidate identification which is harder to implement.

5.2.2 Phase 2 (Candidate Verification)

Both participants followed different strategies in Phase 2 in order to verify candidates identified in Phase 1. P1 started with manual inspection of several relevant components, which then were manually divided into subgroups of candidates. Subsequently, P8 started — supported by further manual analysis — the development of an algorithm applying a divide-and-conquer strategy. Thus, we assume that manual analysis is a crucial factor for choosing this strategy and therefore to solve Phase 2 in a shorter amount of time. Towards the end of Phase 2, P8 applied his algorithm in a slightly adapted version to the other subproblems. Consequently, the divide-and-conquer strategy seems to be an efficient approach for solving HRE problems in this phase.

On the contrary, no clear strategy is identifiable for P1 initially. We observed that P1 tried to continue developing the candidate extraction algorithm from Phase 1 for this phase. We assume that P1 did not develop such a concrete plan as P8 based on the long-lasting stagnation periods observed throughout Phase 2.

Additionally, we monitored a significant stagnation period of 70 minutes, where P1 tried to read-out data from candidates; P8 solved the same problem within 10 minutes. We assume that P8 was able to apply prior knowledge from training tasks faster, since the they taught all necessary methods to solve this problem. A further stagnation period during a manual component selection could be observed for P1. We hypothesize that P1 tried to explore new points of interest in order to continue the verification of his candidates. This is a further hint that P1 applied a sub-optimum strategy.

5.2.3 Phase 3 (Realization)

Forward-thinking of P8 enabled the fast solution time in Phase 3 because the participant could adapt the existing algorithms from Phase 2 in order to remove the watermark. In contrast, P1 struggled as indicated by a long stagnation period in which the participant tried to remove the watermarking on a binary level. Additionally, P1 still needed visual input by manual netlist interactions in order to execute the planned strategy, which is a time-consuming process. In the last stagnation period, P1 tried to improve the algorithm for candidate identification originally implemented in Phase 1.

5.3 Cognitive Factors and HRE (RQ3)

In the light of RQ3, the conduction of statistical analyses was infeasible based on the small sample size. Nevertheless, we reported a promising result concerning a possible negative correlation between scores in Working Memory (WM) and solution time. The term WM is defined as a brain system which enables the storage of sensory input (e.g., visual input) in immediate awareness and the manipulation of that input in order to solve complex cognitive tasks like problem solving [2]. It consists of several parts, for example, the central executive (attentional-controlling system) [3]. Based on a descriptive analysis the data suggested that participants with higher WM scores tend to solve the task faster than participants with lower WM scores. Due to missing statistical analyses, we can only hypothesize that the WM might play a role in solving HRE problem tasks and future studies should statistically investigate the interplay between WM and the efficiency of solving HRE tasks.

Nevertheless, we try to explain differences in problem solving strategies of the participants by referring to the functionalities of the WM. Baddeley and Hitch postulated that the WM is central for solving cognitively difficult tasks like problem solving by performing mental operations on temporally stored information [4]. Against this background, it might be possible that P8 with the highest WM score achieved a more efficient solution than P1 who had a lower WM score. P8's strong WM might have supported the participant in mentally dividing the main problem into several sub problems, and helped in selecting further actions in order to achieve sub goals. Moreover, we assume that P8's stronger WM might have helped to conclude a stagnation period in Phase 2 more efficiently (10 minutes) by quicker activating relevant prior knowledge compared to P1 who needed over 70 minutes. Both participants struggled with the same problem in reading out relevant information from candidates. Although both participants acquired the same amount of HRE knowledge and HRE skills (e.g., methods to read out data from the identified candidates) during the training phase, P8 was quicker in solving that problem. We hypothesize, that the stronger WM enabled P8 to activate stored information from the long-term memory in order to analyze and work with inputs the participant stored in immediate awareness in the WM. This ability of activating prior knowledge is described by the term chunks [7], which postulates that the working load of the WM can be reduced by activation of knowledge structures. Those chunks can boost the immediate memory of the WM and are connected to the level of expertise. A stronger WM could have been advantageous in producing a faster and more efficient solution.

Moreover, the scatter plot also revealed an outlier (P6) with the lowest WM score. A possible explanation for this outlier is that the measurements of WM abilities could have been influenced by uncontrolled variables. Prior work showed that the performances in WM tasks can be impaired by several variables, such as chronic psychological stress [22], acute psychological stress [26], negative emotions like anxiety [24], or neural disorders [44].

5.4 Hypotheses for Cognitive Obfuscation (RQ4)

Hardware obfuscation is understood as design method that impedes reverse engineering. Even though it should not be considered a silver bullet for hardware security, it can be a useful tool, for example, for improving IP protection or increasing the cost factor of an attack. Two important observations of our exploratory study are that (i) all participants progressed through a unique phase model and that (ii) the principal methods with respect to manual and script-based analyses employed by the participants were similar. This allows us to derive hypotheses for hardware obfuscation that take this process steps into account.

5.4.1 Obfuscation delaying HRE Phases

Our data showed that Phase 1, the necessary first step of netlist reverse engineering, can be solved efficiently. In order to increase time on Phase 1, we suggest the following approach for cognitive obfuscation: The netlist should be composed of many equally-looking regular structures, for example, by employing dummy wires or camouflaged gates [36]. Only few of those structures should contain the actual HRE targets. Thus, Phase 1 search techniques will return a considerable number of candidates. As a result Phase 1 will be slowed down through the amount of candidates found, as well as Phase 2 since all candidates have to be verified functionally.

5.4.2 Obfuscation impeding HRE Strategies

Hardware reverse engineers have to develop strategies that are tailored to their specific objective. From a problem solving perspective, Dörner and Funke concluded that a universal problem solving strategy which can be applied to solve different problems does not exist [10]. Against this background, it seems promising to apply different cognitive obfuscation methods for a given netlist which are similar in appearance but require individual solution strategies. This will force the reverse engineer to develop and apply various strategies rather than using a universal one.

5.4.3 Obfuscation against Cognitive Abilities

Based on our initial findings on the role of Working Memory (WM), it is possible that the performance in HRE might be correlated to levels of WM. If future studies can support this hypothesis with statistical data, the development of countermeasures could aim to overload cognitive capabilities of a reverse engineer. One important aspect in this context is the capacity of the WM. Even though the commonly assumed limit of seven items [21] can be extended by activating knowledge structures (chunks) stored in the long-term memory [7,11,16],

the capacity of items that can be stored will always be a rather moderate number. We hypothesize that cognitive obfuscation could overload the capacity of the WM. A possible approach might be the development of structures which force the analyst to apply a combination of many different strategies simultaneously (e.g., simulation in Phase 3 in combination with structural and functional analyses in Phases 1 and 2) which might overload the capacity of the WM.

5.5 Limitations and Future Work

While our exploratory study provides important initial insights into the technical and cognitive processes underlying HRE, there are certainly limitations. First, this research is limited by the small sample size which did not allow us to conduct any statistical analyses. With respect to cognitive factors, future research with larger sample sizes could quantify the influences of WM on HRE processes more accurately. In this context, it would also be interesting to explore to what extent prior knowledge and chunks are activated during HRE.

Second, our results are based on the behavioral analyses of cyber security students who acquired relevant HRE skills and knowledge during an extensive training phase. Even though we worked with students, we took precautions to compare our sample and the Study Task by recruiting an HRE expert. The descriptive analyses of this expert served as a sanity check for our findings, the approximation of experienced revers engineers with our student sample, and the appropriate difficulty of the Study Task. Our results, discussion and takeaways regarding strategies were obtained analyzing two participants. Further detailed analyses of additional participants' and of the expert's HRE problem solving strategies would have been worthwhile. However, given that the primary goal was to provide first insights into the technical and cognitive processes of HRE, we feel that our analysis is warranted and led to interesting results worth reporting. We hope that our exploratory study will trigger follow-up work which focuses on the in-depth analysis of problem solving strategies in HRE. In the future, qualitative data such as interview data may complement the comprehension of technical and cognitive processes in HRE and could potentially lead to a more nuanced view of the strategies the participants followed.

Third, we developed our HRE model based on the behavioral analyses of reverse engineers for a single, medium-complex HRE task and it remains unclear whether the phases transfer to different HRE settings. Nevertheless, it seems plausible that the three phases can be found in other HRE tasks too, but more complex real-world tasks could potentially lead to an extension of the phase model, for example, by discovering further sub phases that are universally observed.

Finally, we proposed first hypotheses for cognitive obfuscation techniques. Future work should evaluate which cognitively difficult tasks are suitable to raise time on HRE tasks and how they can be implemented in netlists.

6 Conclusion

The motivation behind this work is to take a first step towards a better understanding of technical and cognitive processes in HRE, an area with little prior published work despite its importance in commercial and national security contexts. We conducted an exploratory behavioral study with eight participants who solved a realistic HRE task. We examined the approximation of the participants as experienced reverse engineers and the appropriate difficulty of the Study Task through behavioral analyses of an HRE expert. We found that the three-phase model for HRE, which we had postulated, was in fact executed by all participants (as well as by the expert). The analysis of behavioral data showed that the two principal types of actions, manual and automatic analyses, are closely interwoven during a given HRE task. The analysis of cognitive prerequisites of the participants indicates that the Working Memory (WM) might play an important role in solving HRE problems.

Our study suggests several new research directions that can be worthwhile to explore. It seems promising to extend the study to a larger population of participants and, in particular, to observe true HRE experts using a similar methodology as presented here. Another especially interesting research direction lies in the design of a novel class of countermeasures against HRE, which will take cognitive limitations of reverse engineers into account.

Acknowledgements

We would like to thank Yasemin Acar, Maximillian Golla, as well as our anonymous shepherd and all reviewers for their constructive feedback on this work. We are also grateful to Oliver Gebhard and René Walendy for supporting us with data analysis. Finally, we thank Carl for contributing the "Carlgorithm" at an early stage of this research.

This work was supported in part by DFG Excellence Strategy grant 39078197 (EXC 2092, CASA), through ERC grant 695022 and NSF grant CNS-1563829.

References

[1] Leonid Azriel, Ran Ginosar, and Avi Mendelson. SoK: An Overview of Algorithmic Methods in IC Reverse Engineering. In *Proceedings of the 3rd ACM Workshop on Attacks and Solutions in Hardware Security Workshop*, pages 65–74, 2019.

[2] Alan Baddeley. Working Memory. *Science*, 255(5044):556–559, 1992.

[3] Alan Baddeley. Working Memory: Theories, Models, and Controversies. *Annu. Review of Psychology.*, 63:1–29, 2012.

[4] Alan Baddeley and Graham Hitch. Working Memory. In *Psychology of Learning and Motivation*, volume 8, pages 47–89. Elsevier, 1974.

[5] Jens F Beckmann and Natassia Goode. The Benefit of Being Naïve and Knowing It: The Unfavourable Impact of Perceived Context Familiarity on Learning in Complex Problem Solving Tasks. *Instructional Science*, 42(2):271–290, 2014.

[6] R. S. Chakraborty et al. HARPOON: An Obfuscation-Based SoC Design Methodology for Hardware Protection. *IEEE Trans. CAD of Integrated Circuits and Systems*, 28(10):1493–1502, 2009.

[7] William G Chase and Herbert A Simon. Perception in Chess. *Cognitive Psychology*, 4(1):55–81, 1973.

[8] Gregory H Chisholm, Steven T Eckmann, Christopher M Lain, and Robert L Veroff. Understanding Integrated Circuits. *IEEE Design & Test of Computers*, 16(2):26–37, 1999.

[9] Z. Ding et al. Deriving an NCD file from an FPGA bitstream: Methodology, architecture and evaluation. *Microprocessors and Microsystems*, 37(3):299–312, 2013.

[10] Dietrich Dörner and Joachim Funke. Complex Problem Solving: What It Is and What It Is Not. *Frontiers in Psychology*, 8, 2017.

[11] K Anders Ericsson and Walter Kintsch. Long-Term Working Memory. *Psychological Review*, 102(2):211–245, 1995.

[12] Andreas Fischer, Samuel Greiff, and Joachim Funke. The Process of Solving Complex Problems. *Journal of Problem Solving*, 4(1):19–42, 2011.

[13] Marc Fyrbiak, Sebastian Strauß, Christian Kison, Sebastian Wallat, Malte Elson, Nikol Rummel, and Christof Paar. Hardware Reverse Engineering: Overview and Open Challenges. In *2017 IEEE 2nd International Verification and Security Workshop*, pages 88–94. IEEE, 2017.

[14] Marc Fyrbiak, Sebastian Wallat, Jonathan Déchelotte, Nils Albartus, Sinan Böcker, Russell Tessier, and Christof Paar. On the Difficulty of FSM-based Hardware Obfuscation. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, pages 293–330, 2018.

[15] Marc Fyrbiak, Sebastian Wallat, Pawel Swierczynski, Max Hoffmann, Sebastian Hoppach, Matthias Wilhelm, Tobias Weidlich, Russell Tessier, and Christof Paar. HAL – The Missing Piece of the Puzzle for Hardware Reverse Engineering, Trojan Detection and Insertion. *IEEE Transactions on Dependable and Secure Computing*, 16(3):498–510, 2018.

[16] Fernand Gobet and Herbert A Simon. Templates in Chess Memory: A Mechanism for Recalling Several Boards. *Cognitive Psychology*, 31(1):1–40, 1996.

[17] David Z Hambrick and Randall W Engle. The Role of Working Memory in Problem Solving. *The Psychology of Problem Solving*, pages 176–206, 2003.

[18] Jill Larkin, John McDermott, Dorothea P Simon, and Herbert A Simon. Expert and Novice Performance in Solving Physics Problems. *Science*, 208(4450):1335–1342, 1980.

[19] NY Louis Lee and PN Johnson-Laird. A theory of reverse engineering and its application to Boolean systems. *Journal of Cognitive Psychology*, 25(4):365–389, 2013.

[20] Ewen MacAskill and Gabriel Dance. NSA Files: Decoded. https://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded, November 2013. [Online; accessed 2020-February-26].

[21] George A Miller. The Magical Number Seven, Plus or Minus Two Some Limits on Our Capacity for Processing Information. *Psychological Review*, 63(2):81, 1956.

[22] Kazushige Mizoguchi, Mitsutoshi Yuzurihara, Atsushi Ishige, Hiroshi Sasaki, De-Hua Chui, and Takeshi Tabira. Chronic Stress Induces Impairment of Spatial Working Memory Because of Prefrontal Dopaminergic Dysfunction. *Journal of Neuroscience*, 20(4):1568–1574, 2000.

[23] Amir Moradi, Alessandro Barenghi, Timo Kasper, and Christof Paar. On the Vulnerability of FPGA Bitstream Encryption against Power Analysis Attacks: Extracting Keys from Xilinx Virtex-II FPGAs. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, pages 111–124, 2011.

[24] Tim P Moran. Anxiety and Working Memory Capacity: A Meta-Analysis and Narrative Review. *Psychological Bulletin*, 142(8):831, 2016.

[25] Allen Newell and Herbert Alexander Simon. *Human Problem Solving*, volume 104. Prentice Hall Englewood Cliffs, NJ, 1972.

[26] Shaozheng Qin, Erno J Hermans, Hein JF van Marle, Jing Luo, and Guillén Fernández. Acute Psychological Stress Reduces Working Memory-Related Activity in the Dorsolateral Prefrontal Cortex. *Biological Psychiatry*, 66(1):25–32, 2009.

[27] Shahed E Quadir, Junlin Chen, Domenic Forte, Navid Asadizanjani, Sina Shahbazmohamadi, Lei Wang, John Chandy, and Mark Tehranipoor. A Survey on Chip to System Reverse Engineering. *ACM Journal on Emerging Technologies in Computing Systems*, 13(1):6, 2016.

[28] Michael G Rekoff. On Reverse Engineering. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 244–252, 1985.

[29] Jordan Robertson and Michael Riley. The Big Hack: How China Used a Tiny Chip to Infiltrate U.S. Companies. https://www.bloomberg.com/news/features/2018-10-04/the-big-hack-how-china-used-a-tiny-chip-to-infiltrate-america-s-top-companies, October 2018. [Online; accessed 2020-February-26].

[30] Adam Satariano. Huawei Tells Parliament It's No Security Threat, Aiming to Avoid a Ban. https://www.nytimes.com/2019/06/10/technology/huawei-britain-parliament-ban.html, Jun 2019. [Online; accessed 2020-February-26].

[31] Moritz Schmid, Daniel Ziener, and Jürgen Teich. Netlist-Level IP Protection by Watermarking for LUT-Based FPGAs. In *2008 International Conference on Field-Programmable Technology*, pages 209–216. IEEE, 2008.

[32] Pramod Subramanyan, Nestan Tsiskaridze, Wenchao Li, Adria Gascon, Wei Yang Tan, Ashish Tiwari, Natarajan Shankar, Sanjit A Seshia, and Sharad Malik. Reverse Engineering Digital Circuits Using Structural and Functional Analyses. *IEEE Transactions on Emerging Topics in Computing*, 2(1):63–80, 2014.

[33] M Tehranipoor and F Koushanfar. A Survey of Hardware Trojan Taxonomy and Detection. *IEEE Design & Test of Computers*, 27(1):10–25, 2010.

[34] Texplained. Chipjuice IC Reverse Engineering Software. https://www.texplained.com/about-us/chipjuice-software/. [Online; accessed 2020-February-26].

[35] R. Torrance and D. James. The State-of-the-Art in IC Reverse Engineering. In *International Workshop on Cryptographic Hardware and Embedded Systems*, pages 363–381. Springer, 2009.

[36] Arunkumar Vijayakumar, Vinay C Patil, Daniel E Holcomb, Christof Paar, and Sandip Kundu. Physical Design Obfuscation of Hardware: A Comprehensive Investigation of Device-and Logic-Level Techniques. *IEEE Transactions on Information Forensics and Security*, 12(1):64–77, 2016.

[37] Daniel Votipka, Seth M Rabin, Kristopher Micinski, Jeffrey S Foster, and Michelle L Mazurek. An Observational Investigation of Reverse Engineers' Processes. In *29th USENIX Security Symposium (USENIX Security 20)*, Boston, MA, August 2020. USENIX Association.

[38] Sebastian Wallat, Nils Albartus, Steffen Becker, Max Hoffmann, Maik Ender, Marc Fyrbiak, Adrian Drees, Sebastian Maaßen, and Christof Paar. Highway to HAL: open-sourcing the first extendable gate-level netlist reverse engineering framework. In *Proceedings of the 16th ACM International Conference on Computing Frontiers - CF '19*, pages 392–397. ACM Press, 2019.

[39] Sebastian Wallat, Marc Fyrbiak, Moritz Schlögel, and Christof Paar. A Look at the Dark Side of Hardware Reverse Engineering – A Case Study. In *2017 IEEE 2nd International Verification and Security Workshop*, pages 95–100. IEEE, 2017.

[40] David Wechsler. Wechsler Adult Intelligence Scale Fourth Edition (WAIS–IV). *San Antonio, TX: NCS Pearson*, 22:498, 2008.

[41] Carina Wiesen, Nils Albartus, Max Hoffmann, Steffen Becker, Sebastian Wallat, Marc Fyrbiak, Nikol Rummel, and Christof Paar. Towards Cognitive Obfuscation: Impeding Hardware Reverse Engineering Based on Psychological Insights. In *Proceedings of the 24th Asia and South Pacific Design Automation Conference*, pages 104–111. ACM, 2019.

[42] Carina Wiesen, Steffen Becker, Nils Albartus, Christof Paar, and Nikol Rummel. Promoting the Acquisition of Hardware Reverse Engineering Skills. *IEEE Frontiers in Education*, 2019.

[43] Carina Wiesen, Steffen Becker, Marc Fyrbiak, Nils Albartus, Malte Elson, Nikol Rummel, and Christof Paar. Teaching Hardware Reverse Engineering: Educational Guidelines and Practical Insights. In *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering*, pages 438–445. IEEE, 2018.

[44] Erik G Willcutt, Alysa E Doyle, Joel T Nigg, Stephen V Faraone, and Bruce F Pennington. Validity of the Executive Function Theory of Attention-Deficit/Hyperactivity Disorder: A Meta-Analytic Review. *Biological Psychiatry*, 57(11):1336–1346, 2005.

Appendix

A Survey Instruments

Participant Questionnaire

1. Please enter your pseudonym: _____

2. How old are you? Please indicate your age in years. _____
3. What is your target degree? Please select one answer.
   - ○ Bachelor of Science
   - ○ Master of Science
   - ○ Other: _____
   - ○ Prefer not to answer
4. What is your major? Please select one answer.
   - ○ Cyber Security
   - ○ Electrical Engineering
   - ○ Computer Science
   - ○ Other: _____
   - ○ Prefer not to answer
5. Please indicate the number of semesters you have studied so far. _____

Expert Questionnaire

1. How old are you? Please indicate your age in years: _____
2. What is your highest level of education? Please select one answer.
   - ○ Bachelor of Science
   - ○ Master of Science
   - ○ Bachelor of Arts
   - ○ Master of Arts
   - ○ Ph.D.
   - ○ Other: _____
3. What is your current job position? What are your responsibilities? _____
4. On a scale 1 to 5, how would you asses your hardware reverse engineering skill level? Please indicate your skill level on the 5-point scale from 1 (being a beginner) and 5 (being an expert) by choosing one answer.
   ○ 1 (Beginner) ○ 2 ○ 3 ○ 4 ○ 5 (Expert)
5. How many total years of experience do you have with hardware reverse engineering? Please indicate your answer in years. _____

B Full Results

Table 2: Amount of excluded inactivity periods over 60 minutes per participant and phase.

	P1	P2	P3	P4	P5	P6	P7	**P8**	**E**
Phase 1	0	1	0	0	0	0	0	0	0
Phase 2	2	0	0	2	1	0	1	0	1
Phase 3	3	0	3	0	0	1	1	1	0
Total	**5**	1	3	2	1	1	2	**1**	**1**

Table 3: Amount of executed scripts per participant and phase. For comparison, the experts' amount of executed scripts is represented in the rightmost column.

	P1	P2	P3	P4	P5	P6	P7	**P8**	**E**
Phase 1	28	21	33	3	42	20	4	3	5
Phase 2	60	51	72	111	103	35	49	39	122
Phase 3	44	27	21	23	30	40	57	16	37
Total	**132**	99	126	137	175	95	110	**62**	**164**

Table 4: Amount of manual component inspections per participant and phase. For comparison, the experts' amount of manual inspections is shown in the rightmost column.

	P1	P2	P3	P4	P5	P6	P7	**P8**	**E**
Phase 1	93	66	0	4	217	16	47	23	8
Phase 2	19	4	37	162	8	0	93	67	194
Phase 3	15	0	1	21	59	4	24	0	2
Total	**127**	70	38	187	284	20	164	**90**	**204**

Table 5: Overview of assigned progress scores for executed scripts and manual analysis assigned to P1, P8, and the expert as explained in Section 3.3. Dashes indicate zero occurrences.

		P1				P8				Expert			
	Score	0	1	2	3	0	1	2	3	0	1	2	3
Phase 1	Script	22	3	3	1	1	-	1	1	1	2	1	1
	Manual	1	-	-	1	1	-	-	1	8	-	-	-
Phase 2	Script	48	7	4	-	29	5	2	2	92	26	5	1
	Manual	7	2	-	-	1	3	-	1	5	6	2	1
Phase 3	Script	36	5	2	-	12	3	1	-	25	7	4	1
	Manual	2	3	-	-	-	-	-	-	-	1	-	-

Table 6: Scores of the cognitive factors Working Memory (WM), Processing Speed (PS) and Perceptual Reasoning (PR) per participant and time spent on the Study Task in minutes (P3 did not participate in the cognitive tests).

	P1	P2	P3	P4	P5	P6	P7	P8
WM	**108**	126		115	118	92	112	**126**
PS	106	146		146	119	109	119	117
PR	104	129		100	115	100	104	106
Time	**398**	218	186	232	220	261	348	**176**

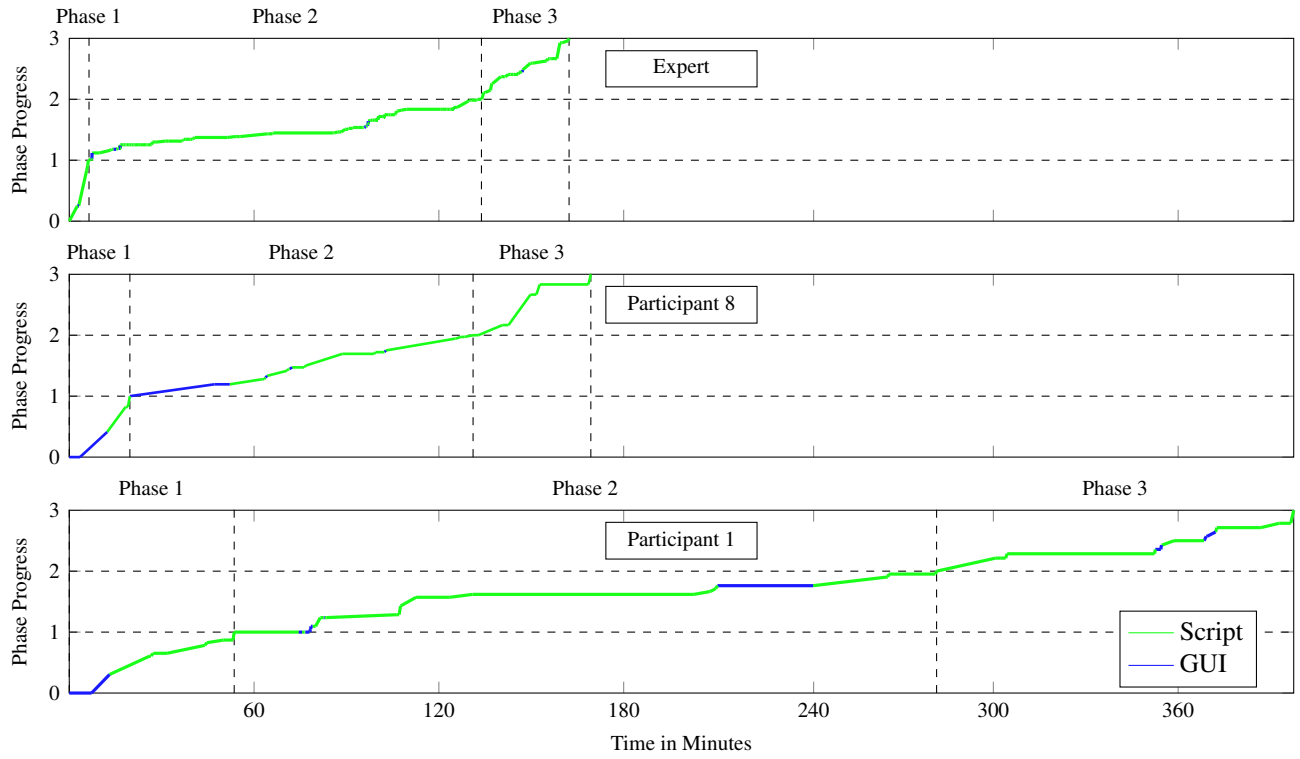


Figure 3: Progress visualization of the expert (top), P8 (middle), and P1 (bottom). The *x*-axis shows time in minutes, and the *y*-axis measures progress based on the progress metric described in Section 3.3. Blue lines indicate manual interactions with the netlist, while green lines indicate interactions with the netlist via semi-automated scripts. Reaching the horizontal lines 1, 2, 3 marks completion of the respective phases, which are also separated by vertical dotted lines.

Industry Responses to the European Directive on Security of Network and Information Systems (NIS): Understanding policy implementation practices across critical infrastructures

Ola Michalec
University of Bristol

Dirk van der Linden
University of Bristol

Sveta Milyaeva
University of Bristol

Awais Rashid
University of Bristol

Abstract

As traditional legacy systems that run critical national infrastructures (CNI) are increasingly digitized for performance monitoring and efficiency, significant attention has been brought to improving their cyber security. Network and Information Systems Security (NIS) Directive is the first European-scale attempt to establish a high standard of cyber security among CNIs. NIS raises questions about defining scope, providing evidence or mobilizing funding. Most importantly, there is the fundamental question whether it would become a tick-box exercise or lead to long-term improvements in security practices. We interviewed 30 cyber security practitioners in the UK to gather an in-depth understanding of NIS implementation and its probable futures. Our analysis found that the emerging field of Operational Technology Security is yet to formulate norms, standards and career trajectories. We are, therefore, at a critical junction, where the scope of the profession is shaping together with the need for evidence-based policy advice. Our findings are twofold: (1) a number of security tropes (e.g., "security solutions are the same across the sectors"), which may drive implementation of regulations such as NIS; (2) a classification of cyber security practices mapping the diversity of policy interpretations. We conclude with recommendations for policymakers and CNI operators.

1 Introduction

Critical National Infrastructures (CNIs) are facilities and systems essential for a country to function. While the exact scope is up for a political decision, sectors like water, energy, transport, defense, health, emergency services would typically be designated as CNIs. However, not all components of these sectors are deemed critical, for example, current policy advice focuses on protecting systems where a major detrimental impact on the delivery of services could pose serious threats to human life or compromise national security [31].

Critical infrastructures are supported by the Operational Technology (OT) and Information Technology (IT) systems. Some examples of IT systems are billing software, staff intranet or data servers, however, their criticality depends on the organizational context. Meanwhile, OT is defined as hardware and software used to monitor and control physical equipment like pumps or valves [24]. Some example systems and devices used in OT environments include: *1) Supervisory Control and Data Acquisition (SCADA)*: control network architecture used for monitoring assets over large geographical areas; 2) *Programmable Logic Controllers (PLC)*: industrial computers built to endure harsh conditions and provide strong safety and real-time properties; 3) *Telemetry, or sensors communicating using radio, infrared or cellular networks*; 4) *Actuators*, components which drive the actual physical process based on commands from PLCs. Experts working in the Operational Technology environment would typically have an engineering background, with specialization in control, networks or safety processes. In the absence of university education, OT specialists would progress in their careers starting from blue-collar roles like technician and plant operator [48].

For decades, OT systems have been limited to basic functionalities, however, with increased digitization and the advent of the Industrial Internet of Things (IIoT), they modernize at an unprecedented scale. One of the reasons for the slow pace of technological advancement in Operational Technology is the strict regulatory environment of the critical infrastructure operators which prioritizes safety. However, the most recent generation of OT devices promises not only improvements in safety but also efficiency and monitoring, e.g., automatic leak detection in water systems or increasing flexibility of energy networks [15, 18, 23]. Yet, reports of recent attacks on critical infrastructures show that securing OT systems from cyber

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

attacks still remains a challenge (see analyses of Triton [29] and the Ukrainian power grid attack [21]).

OT security is much younger than its IT counterpart, and its concerns, traditions and feasible solutions cannot exactly translate from IT security due to the differences in material and regulatory arrangements. Furthermore, a significant knowledge gap exists in terms of OT cyber security best practices [74]. Although the differences between OT and IT security are still poorly understood, protecting Operational Technology environments is a priority for nation-states due to their strategic importance and role in delivering essential services [19]. Therefore, the regulations and standards informing the design of OT systems require special attention from security researchers and practitioners.

In order to investigate the interactions between regulations and innovation, we turn to the Network and Information Systems Security Directive (NIS). NIS is the first European-scale attempt to regulate and stimulate the development of cyber security in CNIs [2]. The directive asked each complying government to identify sectors in scope of the policy. The designated industries are considered essential to human life and exposed to cyber security incidents. As such, there is a number of Operational Technologies within the scope of NIS, which have not been previously dealt with by other digital regulations (i.e. General Data Protection Regulations).

The development of new regulations in the field of emerging technologies raises questions about defining scope, providing appropriate evidence, mobilizing funding and, finally, implementing policy [25]. Consequently, two questions about the possible futures of NIS need to be answered:

- What responses to NIS are likely to bring about meaningful organizational change?
- How can NIS avoid being reduced to a *tick-box exercise*?

While we cannot predict the future, we can anticipate a range of potential outcomes by investigating how OT security practitioners gain their expertise, and how they then apply it to policy interpretation and implementation. We interviewed 30 cyber security practitioners based in the United Kingdom, asking about their experiences of the evolving cyber security policy landscape in the context of critical infrastructures. The research project was guided by the following questions:

RQ1. How is the knowledge of OT security created?

RQ2. How do CNI operators interpret and implement cyber security regulations?

RQ3. What OT security practices can be observed as a result of these regulations?

We addressed the above questions in the context of NIS Regulations, as implemented in our case study country, the United Kingdom. Our questions were motivated by the interest in the emerging profession and the act of collective "sense-making" of the unprecedented policy. We argue that only by understanding the practitioners and practices on the ground we can establish whether and how security may be advanced as a result of regulations such as NIS. Our work is among the first to investigate organizational and practitioner responses to NIS. The novel contributions of our work are as follows:

We identify four Operational Technology (OT) *security tropes* which could influence the implementation of OT security regulations: "separation means security", "IIoT is inevitable", "security solutions are the same across the sectors", "raising awareness leads to security". In these four tropes encountered, the notion of cyber security was equated with solely individual or technological capability. We propose that organizational and political dimensions of security should receive its due regard in the debate. Our analysis of these tropes acts as a call to shape the trajectory of professionalization in this field. We recommend that practitioners who encounter OT security tropes should seek robust evidence and ensure that these statements are appropriately translated to the OT and sector-specific context before being circulated.

We propose a classification of Operational Technology cyber security practices which maps the diversity of policy responses and interpretations. We analyze these practices in conjunction with OT security tropes to indicate how practitioners' understanding of NIS could lead to more security or more insecurity. In doing so, we provide a set of recommendations for policymakers and critical infrastructure operators. Finally, we set the agenda for further research on the emerging field of OT security.

2 Related Work

2.1 (Supra)national policy or legislation on CNI cyber protection

Directive (EU) 2016/1148 [2] on establishing a high level of security of Network and Information Systems (NIS), commonly referred to as the "NIS Directive", is the European Parliament's effort to improve network and information systems security across the European Union (EU). The NIS Directive has been mentioned as a motivating factor for organizations to improve their cyber security processes (cf. [5,47]). At the time of writing (June 2020), most governments identified the policy scope, outlined implementation road maps and suggested penalties for non-compliance [10,51].

Significant research has been done on other European efforts in unifying law addressing digital infrastructures, such as the General Data Protection Regulation (GDPR) [28,39]. However, despite the importance of critical infrastructures to society, organizational and practitioner response to NIS have so far not been investigated in depth. Maglaras et al. [49],

for example, gave a detailed overview of challenges for the implementation of the NIS Directive in Greece, but little work since then has explored practitioners' experiences of policy implementation since different member states transposed the directive into their respective legislations.

Meanwhile, in our case study country, the UK, Shukla, Johnson, and Jones [64] discussed how NIS implementation strategy addresses critical infrastructure security risks in the UK, giving a set of ten recommendations to bridge gaps identified in the NIS framework. Their suggestions centered around holistic security governance, outcome-based audit approach, and progressive road map to improve cyber capabilities of the critical infrastructure operators.

Overall, the UK is considered to be fairly mature in terms of IT cyber security, and less so when it comes to the OT systems [64]. The government's ambition is for the UK to grow domestic cyber security sector and become the global cyber security leader [20]. The potential to realize these ambitions will depend on the governance arrangements across critical infrastructures. Carr [19] called for sustained attention on the emerging public-private partnerships between the operators and the government's regulatory bodies. Each of these partnerships could impact the trajectory of NIS implementation due to varying relationships and funding arrangements between the (energy, water, transport) operators and regulatory bodies overseeing equipment safety and pricing regimes.

Finally, in contrast to Shukla et al., our work presents one of the first empirical works performed during the implementation of NIS, going into much more granular detail to the challenges faced by those affected by NIS, allowing us to reframe what advice remains most urgent to practitioners.

2.2 Security Practices and Behaviors

While researching policy documents is crucial to investigating cyber security regulations, implementation is ultimately a social activity. After all, people create shared understanding of secure behaviors and practices. Previous research on human factors in cyber security focused on systematizing the types and kinds of security behaviors. For example, there is the mapping of employees' information security behavior to various levels of information security knowledge [3]. A typology of end-user security behavior triggers is suggested, where social triggers (i.e. interacting with, or observing other people) are the most common types, and social interactions in the context of security are essential to our understanding of security-related behaviors [26, 27]. Risk perception may also play a role in security practices, as incorrect perceptions have been noted to play a significant part in past attacks on CNIs [57]. Such incorrect perceptions may arise due to latent design conditions, or improper specification of system qualities, borders, observability and controllability [33], making it difficult to reduce blame to the level of the individual. Furthermore, it is necessary to investigate the motivators and barriers of employees' security behaviors, paying attention to responsibility, personal and work boundaries and how these differ across various contexts [12]. To stimulate security behaviors, people need to be positively motivated, e.g., by overcoming negative perceptions of security through establishing trust with audiences and addressing concerns in an honest, transparent way [37]. Moreover, it would seem that one-size-fits-all solutions to improving security are not necessarily realistic, as different authentication methods place burdens on their users, leading to great variations among participants' security approaches and implementations [50].

We argue that the response to cyber security regulations is a result of mutual shaping between policy interpretation, capabilities of the stakeholders and material resources available [63]. Positioning the research in the social rather than individualistic framework requires a shift from behavioral theories to the theories of practice [72]. Situated practices are "routinized and hierarchically organized human activities which take into account material resources" [62]. Although they are widely studied in IT and engineering [30, 71], Cavelty's interdisciplinary literature review [22] shows that situated practices have received limited attention in cyber security. The practice lens encourages to trace how tacit knowledge, circulation of norms and evolving technological capabilities influence each other to shape "what people do" and "how people are". Finally, this focus emphasizes that "security best practices" found in industry guidelines and regulations, when performed in the real-world context, can sometimes lead to the unexpected instances of insecurity [22]. For this reason, it is crucial to understand the difference between "best practices" on paper (i.e. in NIS guidelines) and situated practices found during policy implementation process.

2.3 Differences between IT and OT security

As typical information technology (IT) and operational technology (OT) solutions differ in hardware and software, securing them necessarily does so too. The protection of OT systems from cyber attacks is increasingly important [38, 69]; below we outline how the typical concerns and favored solutions in OT systems differ from the IT (Table 1). It is worth noting that OT security measures are not as established as IT measures, therefore, their efficacy is still under debate [46, 56].

Security behaviors in OT systems may thus also be an entirely different beast from IT systems, as the varying demands of different stakeholders represent many complexities that place OT security into a gray area, with security workers having to balance competing and complex demands [74]. CNIs which fall under NIS operate both OT and IT systems. This further emphasizes the need to study differences between the security practices in IT and OT context, so that we understand what support practitioners require.

Table 1: Differences between IT and OT systems and typical security measures ([46, 56]).

IT SYSTEMS	OT SYSTEMS
• State of the art technology • Usually private enterprises • Priorities are: confidentiality, integrity and availability • Operated by office-based IT professionals	• Legacy systems • Highly regulated for safety, mix of private and state-owned organizations • Priorities are: safety, reliability, robustness, maintainability, integrity and availability • Operated by engineers and manual laborer's
IT SECURITY	**OT SECURITY**
• Relevant standards: ISO270001, NIST Cyber Security Framework • Common solutions: pen testing, firewall, antivirus, insurance, access management • Behaviors and practices well documented	• Relevant standards: IEC 62443 • Potential solutions: patching, access management, firewall • Behaviors and practices poorly understood

3 Study Design

We performed a qualitative study using key informant semi-structured interviews [44] between November 2019 and January 2020, interviewing 30 people across professions and sectors to provide opinions and experiences of security in critical national infrastructures (CNIs). The study was approved by our institutional review board.

3.1 Recruitment

We used a combination of snowball sampling [55] and purposive maximum variability selection [52] attending industry events to establish contacts (10 participants), and through there expanding our search to mutual contacts (17 participants). Besides this, we identified a small number of participants (3 participants) through online cold calling. Our informants were cyber security practitioners who are currently working in various CNI sectors. We stopped recruiting when we reached a sufficient variation of sectors (e.g. energy, water, transport) and roles (e.g. regulators, security consultants, CNI operators) as well as data saturation, a point where consecutive interviews cease to provide novel insights [32]. A basic overview of the participants we recruited is given in Table 2.

3.2 Interview design and methodology

We used a common topic guide for the interviews, shown in Table 3. We designed the topic guide to allow gradually building a rapport and make participants comfortable (interview questions 1-3 in Table 3). In particular, interview questions 2 and 4 pertain to RQ1; interview questions 5 and 6 relate to RQ2 and interview questions 3 and 7 are relevant to RQ3. Questions were tailored to each participant to account for differences in sectors and professions. Interviews took place either at the participant's organization, our institution or via online calls. One primary researcher conducted all interviews,

Table 2: Demographic data.

#	Role	Sector(s)
P01	Security consultant	Oil and Gas
P02	Regulator	Energy
P03	Regulator	Energy
P04	Security working group co-ordinator	Energy
P05	Engineering consultant	Energy
P05	Engineering consultant	Energy (Civil Nuclear)
P07	Director in Engineering consultancy	Water, Energy (Civil Nuclear)
P08	Security Manager at the CNI Operator	Energy
P09	Security Trainer	Defence
P10	Incident Response Director	IT, Finance
P11	Security consultant	IT, Finance
P12	Vendor of security product	IT, Finance
P13	Lawyer	IT, Finance
P14	Working group coordinator	Telecoms
P15	IIOT Manufacturer	Across all
P16	Security consultant	Across all
P17	Business Development at IIOT R&D	Across all
P18	Project Manager in Engineering Consultancy	Across all
P19	Project manager at IIOT R&D	Across all
P20	Security consultant	Transport (Rail)
P21	Safety Engineer	Transport (Rail)
P22	Human factors expert in Engineering Consultancy	Transport
P23	Incident response for a manufacturer	Transport
P24	Security Consultant	Water
P25	Security Consultant	Water
P26	Security manager at the CNI operator	Water
P27	Security manager at the CNI operator	Water
P28	Regulator	Water
P29	Regulator	Water
P30	Regulator	Water

which was necessary to build up rapport and trust given the sensitive nature of participants' work, and to allow for snowballing recruitment. Each interview lasted approximately 60 mins; all conversations were recorded with the interviewees' consent. No reimbursement was given for participation.

3.3 Data coding and analysis

Interviews were transcribed using a professional service. Transcripts were subsequently coded using NVivo software. The coding was based on thematic analysis according to Braun and Clarke [16], taking an inductive, open approach, meaning that we established our themes and sub-themes based

Table 3: Topic guide.

Topic guide for semi-structured interviews
1. *Story of OT security in your sector/organization*: How did OT security look like before NIS, what has NIS changed in comparison?
2. *Story of your career*: How did you get into security? How did your previous roles influence your current job?
3. *Situating security in a sector-specific context*: What are the typical security concerns, regulations, technologies and procedures in your sector/organization?
4. *Standards and regulations*: How do you apply industry standards and security policies in your organization? How to ensure they are applied successfully?
5. *Experiences from NIS*: How do you understand NIS guidelines? What are your opinions on it? How have you been implementing NIS so far?
6. *Investments and innovations in your organization/sector*: What are your plans for the next few months/years with regards to improving security? How will NIS influence your future investments?
7. *Communicating OT security in your organization/sector*: How should we communicate across IT-OT divide? How should we communicate between the board members and the technical experts?

on the most frequent and novel responses rather than fitting participants' opinions to pre-designed categories. We iteratively discussed the transcripts and the developing coding schemata, where we focused on fostering discussion among the authors and building towards a shared understanding of the coded data to ensure coding quality [42]. As Barbour [7] notes: "the degree of concordance between researchers is not really important; what is ultimately of value is the content of disagreements and the insights that discussion can provide for refining coding frames." While some categories were descriptive (e.g., notes on participants' backgrounds and sectoral differences in NIS implementation), in most cases, the act of coding reflected our analytic efforts (e.g., we coded participants' ways of talking which reflect IT, engineer's or regulator's "worldviews"). The resulting codebook is given in Appendix A. We grouped repeating and widely occurring responses to represent major themes (security tropes and policy implementation practices) as well as their sub-themes. For example, "Security tropes" were referenced 102 times across 23 participants, while policy practices were referenced 246 times across 25 participants. We noted that "security tropes" was a worthy category of analysis, as it reflected the emerging state of Operational Technology expertise. Similarly, we classified "policy implementation practices" to establish an exploratory categorization system so further researchers of NIS and other security policies could verify it against their empirical findings.

3.4 Limitations

While we built a sample to ensure the validity of our research (e.g., aiming at our sample being representative in terms of sectors, roles and professions), we neither intend, nor claim to offer highly generalizable results. Although the sample includes a range of people, sectors, and professions, the insights derived from the coding do not necessarily generalize to any of those variables, nor to the wider critical infrastructures context. Nevertheless, the paper provided unique insights into NIS and its implementation. Due to the sensitive nature of the field, an important and time-consuming part of the recruitment process focused on gaining trust and access to participant's honest thoughts and reflections. This was partially enabled by the authors' participation in an industrial-academic research consortium, which may have influenced the topics participants were willing or prioritized to talk about.

4 Results and Discussion

In summary, our research found that the emerging field of Operational Technology Security is yet to formulate professional norms, qualification certifications and career trajectories. We are, therefore, at a critical junction, where the future of the profession is shaping together with the urgent need for evidence-based policy advice, such as NIS directive. Our findings are twofold: (1) a number of security tropes, common practitioners' generalizations about the best OT security measures (e.g., "security solutions are the same across the sectors"); (2) a classification of cyber security practices mapping the diversity of policy implementation (e.g. negotiation, workarounds) . We analyze these practices through the lens of security tropes to highlight whether they are likely to bring about or hinder organizational change with regards to security best practices. In doing so, we provide a set of recommendations for policymakers and CNI operators.

The following section analyzes the results in detail. We first provide a demographic description of participants and outline our findings on the professionalization of Operational Technology security. Then, we analyze the four security tropes widely discussed in the context of NIS implementation. Finally, we propose a classification of policy implementation practices and illustrate how they could lead to security or insecurity. Combining results and discussion demonstrates how a direct dialogue between empirics and theory can advance the field of Operational Technology security as a whole.

4.1 New regulations, new roles

NIS itself mobilized the OT security expertise, with new roles created to enable implementation of cyber security policies. For example, all 5 regulators joined their organizations no earlier than 6 months before the interview date. We observed that some participants anticipated that future recruitment might prove difficult as "*one of the biggest challenges is to scale up resources and to go to the market, especially in OT. There is a really, really, really, really limited number of specialists*" [P03]. Therefore, we argue that NIS both *creates the skills gap*, by prioritizing the need for a new type of expertise as well as *fills the gap* by unlocking investment in new staff.

According to our participants, OT cyber security is a relatively new concept: during the interviews participants frequently remarked that they *"worked in cyber security back*

when there was no such thing as cyber security" [P27]. To understand how participants' professional experience built their present expertise, we charted the diversity of informants' backgrounds in Figures 1 (length of relevant experience) and 2 (education). In the absence of OT security-specific degrees, participants often studied computer science, information security or engineering. In terms of professional experience, interviewees' past roles were largely technical, with only 4 participants having at least 3 years of experience in human and social aspects of technology.

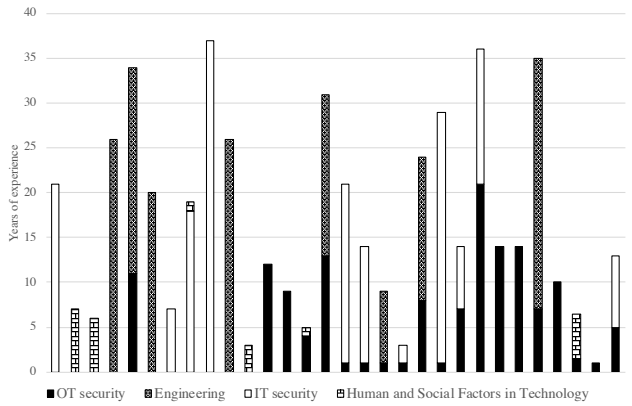


Figure 1: Length of relevant experience – each bar on X axis represents a participant. Y axis corresponds to the cumulative years of relevant experience. X axis is anonymized, and participants are represented in a random order.

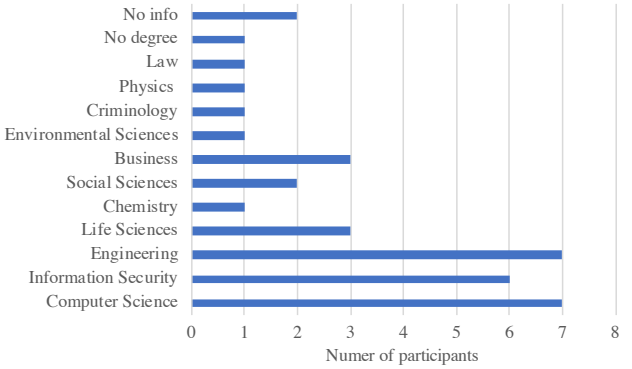


Figure 2: Participants' education. N.B. Some participants obtained more than 1 qualification.

Despite the burgeoning of accreditations and certifications, some participants remain skeptical about their value: *"You look at an individual course and that might be $6,000, and then you realize that's only a small part of it and you need another seven modules, so that ends up $40,000 for a certificate. Should you be spending that much money on certification or should you be going out and helping the industry in doing better and learning more?"* [P16]. Drawing from participants' skepticism about the current cyber security education, we contend that the pathway to professional recognition is not clear.

Across the participants' pool, we observed considerable lateral movement across CNI sectors (e.g., from water to energy), with 24 participants working for more than one CNI sector in their careers. In particular, the defense sector was a notable employer of security experts, as it features in the working history of 8 interviewees. More than half of participants (18) moved across organization types (e.g., from the private sector to the public sector). This could partially explain reports on the OT professionals shortages [40], where the skills gap is shifted from one sector to another rather than tackled with a systematic effort to train new professionals.

4.2 OT Security Tropes

Our main finding suggests that the state of OT security professionalization observed at the time of data collection is characterized by a combination of the following factors:

1. Increasing pressure to recruit experts;
2. Lack of established and "typical" career trajectory;
3. A need for professional education and guidelines.

As the question of NIS implementation is positioned in the center of this "trilemma", we risk that poorly evidenced and OT-inappropriate advice will be circulated to influence key security decisions.

We examined the instances where participants discussed OT security tropes. We define them as widely held beliefs which require a further level of detail before they can be successfully applied to the OT context. Due to the combination of rhetorical qualities like generalization, ambiguity and strong normativity, they lead to the creation of advice which can be easily marketed at mass scale [9]. As they're quite vague, they can appeal to professionals from diverse backgrounds. We argue that participants held a variety of opinions on "the best OT security practices". This reflects the diverse levels of sophistication when it comes to practitioners' understanding of organizational, social and political contingencies of NIS implementation. As the social science studies of expertise suggest, it is crucial to understand these tropes in order to aid professionalization of the industry and effective policy implementation [61].

As previous research on cyber security expertise demonstrated, "translating" security knowledge form IT to OT is not only a matter of IT experts learning engineering [70]. We first ought to enquire: who believes in these tropes and why? Then, we ought to pay attention to people and organizations

benefiting from the circulation of OT security tropes by asking: who makes a profit of it? Whose security "solutions" fit the stereotypes? Finally, we should bear in mind: what other measures are overlooked as a result?

4.2.1 Separation means security

Interviewees discussed the feasibility of security measures. In particular, "air-gapping" received considerable (both positive and negative) attention. Air-gapping employs physical separation of secure computer networks from the unsecured ones (e.g. public Internet, local area network). Traditionally, air-gapping has been applied to critical infrastructures due to the low level of digitization prevalent among their OT systems.

However, the current state-of-the-art attack methods are sophisticated enough to deal with air-gapped systems, with the most well-known (at least to the security practitioners) example being Stuxnet [59]. Yet, outside the OT security bubble, the conversations could look different, e.g., one of our participants recalled difficulties when convincing senior management that "air gapping" does not ensure security: *"When you're talking to the board and they say, "We don't need to worry about security because our production facility is air-gapped", there is only one place which is air gapped and that is Battlestar Galactica!"* [P01].

Debunking air-gapping is justified not only with the advancements in threats, but also with participants frequently predicting that OT systems will continue to become more IT-like, for example through implementing IT-standard network protocols in OT devices or migrating data to the Cloud. Although technologically feasible, IT-OT blending is not adopted across CNI sectors at the same rate due to the current regulatory constraints, such as the requirement for the OT equipment to comply with industry *safety* standards. For example, the regulator for the water industry debunked an assumption of widespread digitalization across all critical infrastructure sectors: *"There are a small number of products that have to be approved because they're in contact with drinking water. For instance, a valve with a computer on the back of it. It's not worth approving this valve with the new computer so you have to use the old computer"* [P28]. In the same interview, he later argued that if the safety regulations remain stringent, OT systems will likely stay, to some extent, air-gapped, meaning *"ultra-safe old-school where you don't connect anything"* [P28]. We, therefore, recommend that before falling for the most technologically advanced (and the most expensive!) security advice, CNI operators ought to sense-check it against the organization-specific conditions. Figure 3 demonstrates how this "trope" relates to NIS.

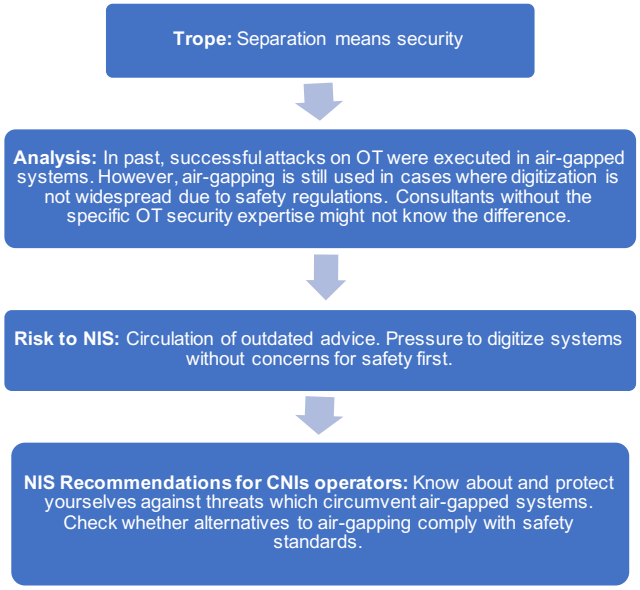


Figure 3: Analysis of the OT security trope: Separation means security.

4.2.2 IIoT is inevitable

Tracing the course of innovation further, we observed a paradox where as many as 11 interviewees would express worry about IIoT, yet at the same time would say it's "inevitable". We observed various ways to contribute to pervading the discourse of "the IIoT inevitability": from framing OT cyber security as a challenge to be overcome solely with technical solutions to treating the socio-political complexities of IIoT as irrelevant to the participant's job. Boyd and Holton [14], in their analysis of innovation discourses, critiqued the assurances of "inevitability". They called for an alternative perspective emphasizing complexity, uncertainty and the role of power relations. As such, we recommend that an alternative look at the future of IIoT would ensure that concerns about security, privacy, affordability, sustainability and labor losses are jointly addressed before deciding whether and how IIoT will be present in critical infrastructures.

In terms of NIS implementation, participants flagged a misalignment between the timescales of IIoT innovation and policy development: *"We're facing the problem of IIoT arriving. When we did the self-assessment, everyone was using very traditional industrial control systems. In that time in the last six months, we've all started adopting IIoT and it's going to get worse. So, it's a big change and it's one that's very much on everyone's radar including mine"* [P26]. We argue that although this is certainly a concern to the industry, concerns about IIoT also opens up a space to generate critical inputs into the evolving OT security regulation landscape. Given that innovation is not technologically determined but

it's a result of co-production, mutual shaping between the society and technology [54, 63], security practitioners are key stakeholders in the process of IIoT co-production as they have the agency to raise, publicize and prioritize their concerns. Figure 4 demonstrates how this trope relates to NIS.

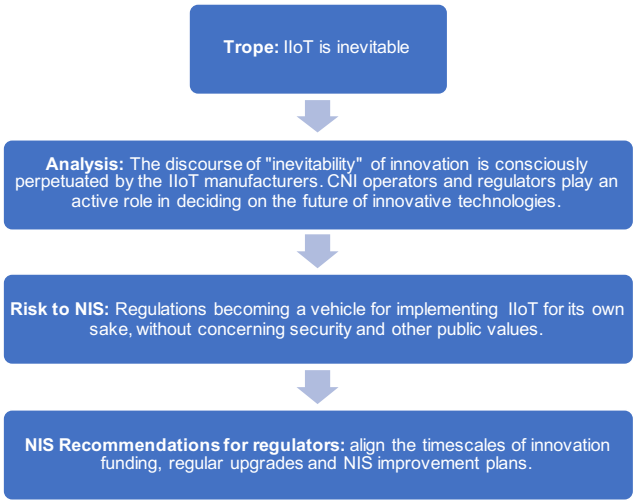


Figure 4: Analysis of the OT security trope: IIoT is inevitable.

4.2.3 Security solutions are the same across the sectors

Another techno-deterministic understanding we observed among participants is that security solutions do not differ across the CNI organizations because *"the tech basis of cyber is the same across the sectors"* [P09]. However, a closer look at the CNI operators' arrangements reveals cross-sectoral differences which can be explained by physical constraints and governance traditions, e.g., *"In oil and gas is, the production facilities, be that an offshore oil platform, are in a small geographical location, you can't get onto an offshore oil platform without getting on to a helicopter. Oil or gas pipelines, on the other hand, are more like the electricity grid, but they are run and owned and operated by completely separate companies. In the water industry, we are unique in so far as we operate both the production sites and the distribution network, and the security model is very different for the two"* [P26].

The diverse ways practitioners understand the application of security measures in new contexts raises questions about the biases they might carry when working across CNI sectors or across IT and OT systems. If security is a subject to the material [4] and regulatory [53] constraints, what is the efficacy of sharing "best practices" or even a cross-sectoral top-down directive like NIS? We recommend that initiatives focused on sharing "best practices" should go beyond talking about security measures and take time to explain unique organizational contexts. We hope that, by turning to the diversity of participants' experiences, we will able to exemplify the need for contextuality in OT security policy development. Figure 5 demonstrates how this trope relates to NIS.

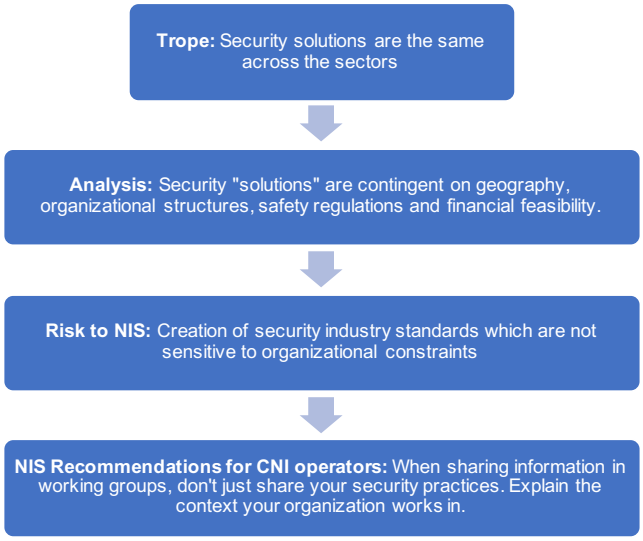


Figure 5: Analysis of the OT security trope: Security solutions are the same across the sectors.

4.2.4 Raising awareness leads to security

Mentions of "awareness raising" were present in 15 interviews. Participants frequently acknowledged lack of awareness as a key issue in OT Security. They had diverse understanding of what constitutes good "awareness" and the likelihood of awareness leading to improved protection. While staff training is one of the deliverables of NIS, we problematize "raising awareness" as an effective educational activity and encourage practitioners to gain a more nuanced understanding of human and organizational factors. The concept of "awareness raising" has gained popularity through the application of the "information deficit model" since the 1980s [65]. The core tenets of the information deficit model are: 1) Ignorance is the reason for a lack of support for various issues in science and technology; 2) Better awareness of science and technology will lead to the desired societal outcomes (ibid.). The framework has been discussed in several empirical studies of cyber security [6, 36, 75]. Information deficit model received criticisms for proposing that rational and individual agency is the key determinant for a given action. However, in the words of an OT Security Consultant, "raising awareness" might not only fail to bring about positive results, but also unwittingly deteriorate the state of affairs: *"P24: I believe is that OT cyber security is such a new thing. It's in the minds of academia and it's in the minds of certain people within essential operating companies. It is not disseminated into the public awareness either as an employee or as a member of the public. And it shouldn't be because general population*

can't rationalize. They can worry. So why would you want to worry a population? Is it beneficial for society to know of all the things that bad people want?"

Going back to the notion of staff training as "raising awareness" of NIS, we found that OT experts had a range of opinions in terms of evidencing that training would work: from colleagues' anecdotes, own reviews, to, finally, an increase in security tickets. In particular, the third notion deserves more attention: one CNI operator reported that although staff training successfully raised awareness about security, it looked like a failure from the outside as it led to an increase in security incidents being reported. We recommend that for OT security, the goal of "raising awareness" ought to be reframed to consider the following questions. First, how do you evidence awareness and security? Second, who should be aware: IT experts, engineers, board members, manual workers, policy makers or the general public? Third, what should they be aware of: technologies, politics, human factors, organizational issues? Finally, we recommend that "awareness raising" should be combined with other training methods, e.g. linking security measures to personal values or communicating operational benefits of security (e.g., improved monitoring, and asset management). Figure 6 demonstrates how this trope relates to NIS.

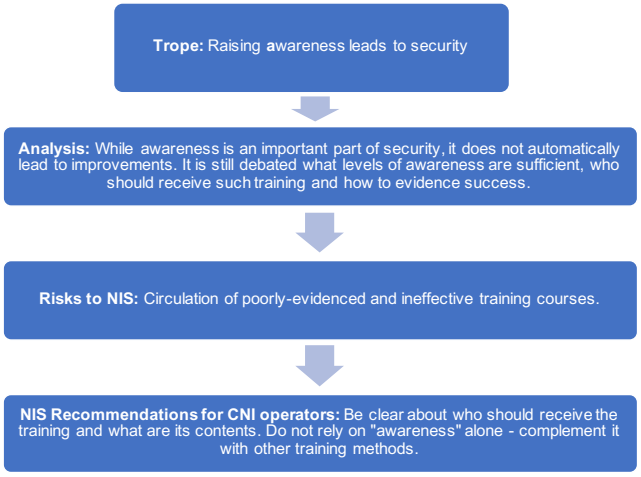


Figure 6: Analysis of the OT Security trope: Raising awareness leads to security.

4.3 Social practices observed

The following section will turn to social practices of NIS implementation, an uncovered through our qualitative interviews. If behaviors are analyzed through the lens of individual perceptions, motivations and attitudes; both OT security tropes and practices are examined in relation to power, competences, meanings and materials [62]. We show that a nuanced understanding of OT security tropes could contribute to the construction of well-evidenced and context-specific expertise and, ultimately, to the adoption of secure practices. However, as OT security is in the early stage of professionalization, some of the observed practices could, in fact, undermine the case for improvements in security through regulations.

Drawing from previous studies on policy classifications [13, 67]we argue that an empirically based exploratory classification system could be of use to fellow researchers and practitioners. We classified the policies under high-level themes to avoid falling into the trap of a formalistic understanding of law; instead showing that NIS does not happen in vacuum and is shaped by the environment in which it is implemented.

4.3.1 Classification of practices

We propose an exploratory classification of practices enacted as a response to cyber security regulations across critical infrastructures (Table 4). The aim of our exploratory classification is to establish an empirically grounded investigation into the 'actually existing' activities rather than idealized types [68]. As such, we identified four main categories: compliance, workaround, going above and beyond policy remit, and negotiation. We recognize that this is not an exhaustive list but, rather, a call for closer examination of the relations between the stakeholders involved in NIS. We intend that the classification could be further utilized by security researchers and practitioners; i.e. they could verify our categories and add new ones to reflect their experiences.

In creating our classification, we acknowledged previous work conducted in the field. For example, compliance and workarounds are well-researched in cyber security [8, 11, 35, 41, 45, 58, 73], however, rarely from the lens of situated practices [22]. The latter two categories ("going above and beyond policy remit" and "negotiation") received less attention in cyber security scholarship (with a notable exception of Slayton and Clark-Ginsberg [66], Shane [60] and Carr [19]).

Overall, past literature conceptualized compliance and workarounds as one-way transactions, where a policymaker sets the rules and a policy recipient responds to them. Meanwhile, we question this static configuration, demonstrating that in all four categories critical infrastructure operators are not passive recipients of the policy but its active co-creators. Consequently, our analysis is focused on relations which make the acts of policy interpretation and implementation "happen".

4.3.2 Compliance

We understand compliance as "*the act of obeying a formal cyber security policy*". The topic was extensively studied by the security community. For example, Gerber et al. [35] investigated the effectiveness of goal setting and rewards, whereas Safa et al. [58] and Bauer and Bernroider [8] examined employees' attitudes to compliance drawing from the social bond

Table 4: Practices of policy interpretation and implementation in the field of OT security.

	Compliance	Workaround	Going above and beyond policy remit	Negotiation
Security Example	Completing asset discovery as an essential basis for further cyber improvements	OT experts implementing their own security measures, using policy as a "sanity check"	Intelligence sharing through a working group set on the basis of trust and shared terms of reference	Operators giving feedback on Infosec-biased language of self-assessment forms
Insecurity example	Interpreting the scope of NIS self-assessment framework to own advantage, while excluding key OT assets	Senior executives ignoring the need for improvements	Overreliance on the latest 'buzzword' technologies when basic knowledge is missing	Prioritizing business values over public values in policy interpretation (e.g., security at cost of privacy)

theory and the theory of reasoned action. Yet, we observe that compliance is not limited to a mere acceptance of the policy (expressed by, for example, a positive attitude to it), but it involves a degree of interpretation.

The NIS Regulations are written in a top-down manner; therefore, they are not specific to CNI sectors. At the time of writing, the only sector-specific documents were regulators' guidelines on completing self-assessment forms. In the eyes of critical infrastructure operators, NIS requires the operators to manage assets which not only previously lacked regulations in terms of security, but also lacked adequate procedures in other operational areas. As such, some participants admitted that they need to *"get the basics right before thinking about expensive silver bullets"* [P01]. In order to improve the basics of security, CNI operators ought to record their assets, decide which ones to deem critical and, then, establish procedures for management and monitoring: *"CNIs don't necessarily understand their assets. So, the water industry, for instance, might have tens of thousands of assets distributed over, 200, 300 square km. Do they know everything about every one of those assets? Not necessarily because some of them might have been put in 50, 60 years ago. They might have dropped off an asset list sometime. They might have come back on. It might have been refreshed but left there. Who knows? And they are finding out that the work, that a discovery piece in their asset management is pretty huge"* [P01]

How does one know what counts as a critical asset? We observed that this decision usually depends on the security manager's competences: what if an IT-trained practitioner included only enterprise IT, excluding cyber-physical infrastructures from the scope? Furthermore, our analysis shows that the processes also differ depending on the organization type, e.g., participants argued that asset discovery is more challenging in CNIs with geographically dispersed assets, e.g., in the water sector. Finally, we argue that understanding of *asset criticality* can be constructed as a way to advance own career: a security manager could be interpreting the scope of NIS self-assessment framework to achieve a good score, while excluding assets which cannot be easily secured. One participant remarked: *"As the self-assessment form is subjective, it is a reflection of mindsets rather than cyber maturity, e.g. some companies are adding physical security*[1] *stuff in their scope and, therefore scoring themselves higher."* [P27].

We link these compliance practices to the trope that "awareness leads to security". We suggest asking: who should be aware of what? What assumptions are made about the current awareness of policymakers and CNI operators? Should policymakers be aware of sectoral specifics so they can write better guidelines? Should security managers (especially if their role or a whole field is new) be assumed to correctly assess the policy scope? We recommend that practitioners pay continuous attention to the idea of "translation" across IT and OT as well as across the sectors to improve their capabilities of policy formulation and interpretation.

4.3.3 Workarounds

Workarounds are "*circumventions of a cyber security policy, which do not explicitly address its problems*". Like compliance, workarounds received considerable attention from the researchers, especially as they tie into the idea of usability [11, 73] Kirlappos et al. [41] introduced the term "shadow security" to describe employees' unofficial security measures (some of them of questionable efficacy) devised to ensure their day-to-day work goals are achieved. Koppel et al. [43] argued that understanding workarounds requires not only an analysis of technical rules, but also interviews and observations of key informants. In our classification of practices as workarounds, we drew from social scientific understanding of the term [17], identifying actions which evade security policies in a non-confrontational manner. Through avoiding confrontation and applying their own definitions of "appropriate and proportionate measures", NIS stakeholders are devoid of an opportunity to negotiate the scope of the policy.

In the context of NIS, some operators were known to implement their own security improvement plans, using NIS as a "sanity check". One of the energy regulators argued that this is a welcome practice since the policy was written in a basic and generic way: *"There are people who are more confident, or have a different attitude to risk perhaps, and will have their own views about what is the right thing to do in their organization, and they might use NIS as a kind of sanity check, a checklist to see how they compare with it. But their real logic, decision-making will be based on their expert knowledge of what they think is the best thing to do in their circumstances, and they won't blindly follow NIS"* [P02].

Operators could also use workarounds to avoid implementing security measures without the need to openly criticize

[1] CNI employees traditionally differentiate between cyber security (related to the protection of personnel and equipment from malicious incidents using digital technologies), physical security (related to the protection of personnel and equipment from malicious incidents) and safety (related to the protection of personnel and equipment from accidents).

policy, e.g.: *"I didn't enjoy being a CISO because it was always going into the board saying: "You need to spend money", and the board said "Well, why? Prove it, show me metrics, show me reasons", and I can scare them with regulations, but it was never a very scientific question, it was always a bit finger in the air and, "This is what happens if we don't do it, but it might not because we might not get hit""* [P10]. We note that as the core requirement of NIS is responding in an "appropriate and proportionate" manner, the policy does not prescribe risk assessment methodologies or security budgets, leaving the decision what (not) to secure to the operators.

We link the above "workaround" to the trope "separations means security". Here we reiterate the point, that "appropriate and proportionate" interpretation of NIS cannot be simply assumed as senior decision makers in CNIs might not have the expertise in the state-of-the-art attack methods. We also state that relying on outdated measured is not only linked to the lack of knowledge, but also to organizational hierarchies: board members might circumvent the policy if it is tied to activities, they are not willing to take part in. We cannot ignore the dimensions of power relations present in workaround practices. To better understand how practitioners, construct the notions of "risk" and "appropriate measures", we recommend further research asking the following questions: How is this knowledge negotiated between the board, a CISO and other employees? How do practitioners know what is "proportionate and appropriate"?

4.3.4 Going above and beyond policy remit

Across the responses to NIS, we noticed that the ambitions of some critical infrastructure operators exceeded the policy requirements. We call this category "going above and beyond the policy remit". As OT cyber security policies are new, we did not identify any relevant past research related to this concept. However, we noted that there is a considerable amount of grey literature on "industry best practices" (e.g., [1, 34]). While these reports might be detailed and informative, they lack rigor to be treated as evidence for policymaking.

We identified that certain CNI operators formed working groups to share progress on NIS and establish a whole-sector benchmark. We claim that these practices are examples of reconfigurations of power and competence, where the operators are able to elevate their status through co-operation and sharing capabilities. The existence of working groups questions the notion of policy implementation as a one-way, passive activity. Their evolution will be interesting to observe, especially as many of CNIs for-profit companies and competitors.

As participants reported, working groups were not required by the regulators, therefore they tend to be industry-led: *"So, the working group is something has been going on for years now. We [regulators] are not permanent members. We were invited, of course, to be part, but this is a closed forum for operators that is running for years for them to share and it's not only about cyber but also about other topics as well to share and to experience"* [P03]. Nevertheless, working groups are not necessarily uniformly effective. Participants remarked that the key to success are: the basis of trust, shared terms of reference and secure storage of confidential data.

An example of a high ambition which does not lead to security improvements is overreliance on the latest 'buzzword' technologies when basic knowledge of security is missing. We noted participants' fears that some operators can be tempted to neglect basic improvements in favor of asset upgrades which they argue as "due to be replaced". Regulators reported that the key part of their role to differentiate between legitimate security improvements and costly innovations for their own sake: *"We want a highly resilient network, so that implies that you replace these assets before they stop working, and there's some subjectivity when that should be. So, there's an argument that operators put forward is, we should replace those assets a bit sooner than previously forecasted and at the same time we can upgrade the cyber security. So that potentially saves them some money, but it's hard to draw out the separation sometimes between the cyber security arguments and the physical lifetime of the assets, but there are big sums of money, hundreds of millions."* [P02].

Consequently, we identify that the practices of "going above and beyond policies" are at risk of falling for the "IIOT is inevitable" trope. We recommend that security practitioners are cautious of the promises made by the manufacturers of innovative technologies. IIoT, machine learning, and "essential upgrades" are not necessarily inevitable. Furthermore, the funding mechanisms of critical infrastructures innovations in the field of security ought to receive a closer scrutiny.

4.3.5 Negotiation

Our final category of Operational Techonology cyber security policy response is negotiation. We define it as a *"collaborative process, where cyber security stakeholders co-create the interpretation and implementation of the policy"*. Negotiations often involve compromising on conflicting priorities. Co-production of cyber security expertise was described by Slayton and Clark-Ginsberg [66] who argued that cyber security expertise is value-laden and it can contribute to "regulatory capture", a situation where regulation serves private interests rather than the public good. Furthermore, Carr [19] analyzed the private-public partnerships forming in the UK and the U.S. Her paper identified a disconnect between the expectations of the public and private sector stakeholders in terms of roles, responsibilities and power. These influential papers are one of the few high-level empirical works in the field of cyber security governance. Here we complement their findings by outlining the details of stakeholders' configurations and practices on the ground.

One example of negotiation is the practice of the operators and regulators working together to improve the language of

NIS guidelines and self-assessment forms. In the following quote, an OT security manager busts the myth that "security measures are the same across the sectors". He reports: *"one of my big bugbears with the OT cyber security policies is that they are info sec driven. I prefer to use the term "cybersecurity" because information isn't the asset. Traditionally, forget the computer systems. I start with something I say to everybody. There are two fundamental differences between information security and OT security. Number one is that the computer system is just another component in the mechanical plant in my world. So, it has no more importance than a pump or a valve. If it breaks, the plant stops working (. . .) So the regulator looked for two water companies to work with them last year to develop NIS into something workable for the water industry. They brought out draft guidance and I sat down with them for a day, and I went through some of the things which just don't work and there's a big difference."* [P26]

In negotiating how to improve security through NIS through the emerging working groups and public-private partnerships, there is a risk that a significant influence of private sector over policy interpretation could lead to prioritizing business values over public values. As the most cyber-mature operators establish close relationships with the regulators, policymakers ought to remain objective and avoid biases during audits and funding competitions. To complement that, CNI operators ought to improve their capabilities in the areas of human and social factors of technology, so they while implementing NIS, they do not compromise on other public values such as privacy, sustainability or equity.

4.4 POLICY RECOMMENDATIONS FOR NIS DIRECTIVE

Based on the analysis of OT security tropes in conjunction with our mapping of policy implementation practices, we proved a set of five policy recommendations for NIS stakeholders in the UK and other European countries. We outline these below mirroring the order of research findings in earlier sections and specifying the target audience for each.

Recommendation 1 (for CNIs operators deciding on improvement plans): Know about and protect yourselves against threats which circumvent air-gapped systems. Check whether alternatives to air-gapping comply with safety standards.

Recommendation 2 (for regulatory bodies overseeing NIS): Align the timescales of innovation funding, regular upgrades and NIS improvement plans. When approving price reviews for network upgrades, seek robust evidence for the claims on the operational benefits of proposed innovations.

Recommendation 3 (for CNI operators responsible for cyber security training): Tie the training to employees' personal concerns to make it relatable and interesting. Do not rely on "awareness" alone - complement it with other training methods. Above all, place "awareness" in the usability context of daily work; i.e. plant supervisors have different concerns to admin staff.

Recommendation 4 (for all stakeholders) Practitioners should pay continuous attention to the idea of "translation" across IT and OT as well as across the sectors to improve their capabilities of policy formulation and interpretation. This will ensure that the scope and latter auditing of NIS pertains both OT and IT and that the improvements are tailored to each system.

Recommendation 5 (for all stakeholders) We recommend that security practitioners ought to improve their capabilities in the areas of human and social factors of technology, so they while implementing NIS, they do not compromise on other public values such as privacy, sustainability or equity.

5 Conclusions

The Network and Information Systems Security (NIS) Directive is the first European-scale attempt to establish a high standard of cybersecurity among the operators of critical infrastructures. In order to understand whether it is likely to bring about a meaningful organizational change and avoid becoming a tick-box exercise, we interviewed 30 UK-based cyber security practitioners inquiring about their experiences of policy implementation.

Our analysis found that the emerging field of Operational Technology (OT) Security is yet to formulate norms, standards and career trajectories. We ought to be careful that "OT security tropes" are appropriately scrutinized before serving as policy advice. Furthermore, our study proposed a classification of cyber security practices which maps the diversity of policy responses to NIS. We analyzed OT security practices through the lens of OT security tropes to indicate whether they could lead to more security or insecurity. As such, we observed that the process of NIS implementation is a two-way exchange. We provided empirical examples of the practices of negotiation and going above and beyond policy remit which illustrate this point. Similarly, we argue that compliance with NIS might not be reduced to a tick-box exercise if practitioners are reflective about "OT security tropes" circulating in the field. Practices most likely to bring about meaningful organizational change are appropriately situated in the OT and sector-specific context and aligned with cross-cutting public values (e.g. security, privacy, sustainability and equity).

From a research point of view, our findings recommend that the funding mechanisms of CNI innovations in the field of security ought to receive closer scrutiny, and that further research should be done to better understand how practitioners construct the notions of "risk" and "appropriate and proportionate measures".

Acknowledgments

This work was supported by RITICS grant "How many shades of NIS? Understanding Organizational Cybersecurity Cultures and Sectoral Differences".

References

[1] 12 Best Cybersecurity Practices in 2020. https://www.ekransystem.com/en/blog/best-cyber-security-practices. Accessed: 26-02-2020.

[2] Directive (EU) 2016/1148 of the european parliament and of the council of 6 july 2016 concerning measures for a high common level of security of network and information systems across the union. *OJ*, L 194, 19-07-2016.

[3] Zauwiyah Ahmad, Mariati Norhashim, Ong Thian Song, and Liew Tze Hui. A typology of employees' information security behaviour. In *2016 4th International Conference on Information and Communication Technology (ICoICT)*, pages 1–4. IEEE, 2016.

[4] Claudia Aradau. Security that matters: Critical infrastructure and objects of protection. *Security dialogue*, 41(5):491–514, 2010.

[5] Helder Aranha, Massimiliano Masi, Tanja Pavleska, and Giovanni Paolo Sellitto. Enabling security-by-design in smart grids: An architecture-based approach. In *2019 15th European Dependable Computing Conference (EDCC)*, pages 177–179. IEEE, 2019.

[6] Maria Bada, Angela M Sasse, and Jason RC Nurse. Cyber security awareness campaigns: Why do they fail to change behaviour? *arXiv preprint arXiv:1901.02672*, 2019.

[7] Rosaline S Barbour. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *Bmj*, 322(7294):1115–1117, 2001.

[8] Stefan Bauer and Edward WN Bernroider. From information security awareness to reasoned compliant action: analyzing information security policy compliance in a large banking organization. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 48(3):44–68, 2017.

[9] Bernadette Bensaude Vincent. The politics of buzzwords at the interface of technoscience, market and society: The case of 'public engagement in science'. *Public understanding of science*, 23(3):238–253, 2014.

[10] Bird & Bird. Developments on nis directive in eu member states. Technical report, 2020.

[11] Jim Blythe, Ross Koppel, and Sean W Smith. Circumvention of security: Good users do bad things. *IEEE Security & Privacy*, 11(5):80–83, 2013.

[12] John M Blythe, Lynne Coventry, and Linda Little. Unpacking security policy compliance: The motivators and barriers of employees' security behaviors. In *Eleventh Symposium On Usable Privacy and Security (SOUPS) 2015)*, pages 103–122, 2015.

[13] Geoffrey C Bowker and Susan Leigh Star. *Sorting things out: Classification and its consequences*. MIT press, 1999.

[14] Ross Boyd and Robert J Holton. Technology, innovation, employment and power: Does robotics and artificial intelligence really mean social transformation? *Journal of Sociology*, 54(3):331–345, 2018.

[15] Hugh Boyes, Bil Hallaq, Joe Cunningham, and Tim Watson. The industrial internet of things (iiot): An analysis framework. *Computers in Industry*, 101:1–12, 2018.

[16] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[17] David Campbell. Policy workaround stories are valuable evaluative indicators: but should they be told? *American Journal of Evaluation*, 32(3):408–417, 2011.

[18] Catarina Araya Cardoso, Jacopo Torriti, and Mate Lorincz. Making demand side response happen: A review of barriers in commercial and public organisations. *Energy Research & Social Science*, 64:101443, 2020.

[19] Madeline Carr. Public–private partnerships in national cyber-security strategies. *International Affairs*, 92(1):43–62, 2016.

[20] Madeline Carr and Leonie Maria Tanczer. Uk cybersecurity industrial policy: an analysis of drivers, market failures and interventions. *Journal of Cyber Policy*, 3(3):430–444, 2018.

[21] Defense Use Case. Analysis of the cyber attack on the ukrainian power grid. *Electricity Information Sharing and Analysis Center (E-ISAC)*, 388, 2016.

[22] Myriam Dunn Cavelty. Cybersecurity research meets science and technology studies. *Politics and Governance*, 6(2):22–30, 2018.

[23] Weiping Cheng, Hongji Fang, Gang Xu, and Meijun Chen. Using scada to detect and locate bursts in a long-distance water pipeline. *Water*, 10(12):1727, 2018.

[24] Yulia Cherdantseva, Pete Burnap, Andrew Blyth, Peter Eden, Kevin Jones, Hugh Soulsby, and Kristan Stoddart. A review of cyber security risk assessment methods for scada systems. *Computers & security*, 56:1–27, 2016.

[25] Jennifer Chubb, Jasper Montana, Jack Stilgoe1, Andy Stirling, and James Wilsdon. A review of recent evidence on the governance of emerging science and technology. Technical report, Wellcome Trust, 11 2018.

[26] Sauvik Das, Laura A Dabbish, and Jason I Hong. A typology of perceived triggers for end-user security and privacy behaviors. In *Fifteenth Symposium on Usable Privacy and Security ({SOUPS} 2019)*, 2019.

[27] Sauvik Das, Tiffany Hyun-Jin Kim, Laura A Dabbish, and Jason I Hong. The effect of social influence on security sensitivity. In *10th Symposium On Usable Privacy and Security (SOUPS) 2014)*, pages 143–157, 2014.

[28] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We value your privacy... now take some cookies: Measuring the gdpr's impact on web privacy. *arXiv preprint arXiv:1808.05096*, 2018.

[29] AC Alessandro Di Pinto, Younes Dragoni, and Andrea Carcano. Triton: The first ics cyber attack on safety instrument systems. In *Proc. Black Hat USA*, pages 1–26, 2018.

[30] Yvonne Dittrich. What does it mean to use a method? towards a practice theory for software engineering. *Information and Software Technology*, 70:220–231, 2016.

[31] Jens Ivo Engels. *Key Concepts for Critical Infrastructure Research*. Springer, 2018.

[32] Sandra L Faulkner and Stormy P Trotter. Data saturation. *The international encyclopedia of communication research methods*, pages 1–2, 2017.

[33] Sylvain Frey, Awais Rashid, Alberto Zanutto, Jerry Busby, and Karolina Follis. On the role of latent design conditions in cyber-physical systems security. In *Proceedings of the 2nd International Workshop on Software Engineering for Smart Cyber-Physical Systems*, pages 43–46, 2016.

[34] GE. An executive guide to cyber security for operational technology. Technical report, 2017.

[35] Nina Gerber, Ronja McDermott, Melanie Volkamer, and Joachim Vogt. Understanding information security compliance-why goal setting and rewards might be a bad idea. In *HAISA*, pages 145–155, 2016.

[36] Antonios Gouglidis, Benjamin Green, Jeremy Busby, Mark Rouncefield, David Hutchison, and Stefan Schauer. Threat awareness for critical infrastructures resilience. In *2016 8th International Workshop on Resilient Networks Design and Modeling (RNDM)*, pages 196–202. IEEE, 2016.

[37] Julie M Haney and Wayne G Lutters. "it's scary... it's confusing...it's dull": How cybersecurity advocates overcome negative perceptions of security. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS) 2018)*, pages 411–425, 2018.

[38] Wayne Harrop and Ashley Matteson. Cyber resilience: A review of critical national infrastructure and cybersecurity protection measures applied in the uk and usa. In *Current and Emerging Trends in Cyber Operations*, pages 149–166. Springer, 2015.

[39] Chris Jay Hoofnagle, Bart van der Sloot, and Frederik Zuiderveen Borgesius. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98, 2019.

[40] $(ISC)^2$. Strategies for building and growing strong cybersecurity teams. Technical report, 2019.

[41] Iacovos Kirlappos, Simon Parkin, and M Angela Sasse. " shadow security" as a tool for the learning organization. *ACM SIGCAS Computers and Society*, 45(1):29–37, 2015.

[42] Simon C Kitto, Janice Chesters, and Carol Grbich. Quality in qualitative research. *Medical journal of Australia*, 188(4):243–246, 2008.

[43] Ross Koppel, Sean W Smith, Jim Blythe, and Vijay H Kothari. Workarounds to computer access in healthcare organizations: you want my password or a dead patient? *ITCH*, 15(4):215–220, 2015.

[44] PJ Lavrakas. Key informant. *En Encyclopedia of Survey Research Methods*, page 2455, 2008.

[45] Chunghun Lee, Choong C Lee, and Suhyun Kim. Understanding information security stress: Focusing on the type of information security compliance activity. *Computers & Security*, 59:60–70, 2016.

[46] John Lindström, Petra Viklund, Tideman Fredrik, Hällgren Berndt, and Elvelin Jonny. Oh, no–not another policy! oh, yes-an ot-policy! In *52nd CIRP Conference on Manufacturing Systems (CMS), Ljubljana, Slovenia, June 12-14, 2019*, volume 81, pages 582–587, 2019.

[47] Martin Lundgren and Erik Bergström. Security-related stress: A perspective on information security risk management. In *2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, pages 1–8. IEEE, 2019.

[48] Erik Skov Madsen, Arne Bilberg, and D Grube Hansen. Industry 4.0 and digitalization call for vocational skills, applied industrial engineering, and less for pure academics. In *Proceedings of the 5th P&OM World Conference, Production and Operations Management, P&OM*, 2016.

[49] Leandros Maglaras, George Drivas, Kleanthis Noou, and Stylianos Rallis. Nis directive: The case of greece. *EAI Endorsed Transactions on Security and Safety*, 4(14), 2018.

[50] Shrirang Mare, Mary Baker, and Jeremy Gummeson. A study of authentication in daily life. In *Twelfth Symposium on Usable Privacy and Security (SOUPS) 2016)*, pages 189–206, 2016.

[51] Dimitra Markopoulou, Vagelis Papakonstantinou, and Paul de Hert. The new eu cybersecurity framework: The nis directive, enisa's role and the general data protection regulation. *Computer Law & Security Review*, 35(6):105336, 2019.

[52] Bryan Marshall, Peter Cardon, Amit Poddar, and Renee Fontenot. Does sample size matter in qualitative research?: A review of qualitative interviews in is research. *Journal of computer information systems*, 54(1):11–22, 2013.

[53] Fabio Massacci, Raminder Ruprai, Matthew Collinson, and Julian Williams. Economic impacts of rules-versus risk-based cybersecurity regulations for critical infrastructure providers. *IEEE Security & Privacy*, 14(3):52–60, 2016.

[54] Clark A Miller and Sheila Jasanoff. States of knowledge: The co-production of science and social order, 2004.

[55] Chaim Noy. Sampling knowledge: The hermeneutics of snowball sampling in qualitative research. *International Journal of social research methodology*, 11(4):327–344, 2008.

[56] Qais Saif Qassim, Norziana Jamil, Maslina Daud, Ahmed Patel, and Norhamadi Ja'affar. A review of security assessment methodologies in industrial control systems. *Information & Computer Security*, 2019.

[57] Awais Rashid, Joseph Gardiner, Benjamin Green, and Barnaby Craggs. Everything is awesome! or is it? cyber security risks in critical infrastructure. In *International Conference on Critical Information Infrastructures Security*, pages 3–17. Springer, 2019.

[58] Nader Sohrabi Safa, Rossouw Von Solms, and Steven Furnell. Information security policy compliance model in organizations. *computers & security*, 56:70–82, 2016.

[59] P Shakarian, J Shakarian, and A Ruef. Attacking iranian nuclear facilities: Stuxnet. *Introduction to cyberwarfare: A multidisciplinary approach*, pages 223–239, 2013.

[60] Peter M Shane. Cybersecurity policy as if ordinary citizens mattered: The case for public participation in cyber policy making. *Isjlp*, 8:433, 2012.

[61] James Shires. Enacting expertise: Ritual and risk in cybersecurity. *Politics and Governance*, 6(2):31–40, 2018.

[62] Elizabeth Shove, Mika Pantzar, and Matt Watson. *The dynamics of social practice: Everyday life and how it changes*. Sage, 2012.

[63] Elizabeth Shove and Frank Trentmann. *Infrastructures in practice: the dynamics of demand in networked societies*. Routledge, 2018.

[64] Meha Shukla, Shane D Johnson, and Peter Jones. Does the NIS implementation strategy effectively address cyber security risks in the uk? In *2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, pages 1–11. IEEE, 2019.

[65] Molly J Simis, Haley Madden, Michael A Cacciatore, and Sara K Yeo. The lure of rationality: Why does the deficit model persist in science communication? *Public Understanding of Science*, 25(4):400–414, 2016.

[66] Rebecca Slayton and Aaron Clark-Ginsberg. Beyond regulatory capture: Coproducing expertise for critical infrastructure protection. *Regulation & Governance*, 12(1):115–130, 2018.

[67] Kevin B Smith. Typologies, taxonomies, and the benefits of policy classification. *Policy Studies Journal*, 30(3):379–395, 2002.

[68] Kevin B Smith. Typologies, taxonomies, and the benefits of policy classification. *Policy Studies Journal*, 30(3):379–395, 2002.

[69] Kristan Stoddart. Uk cyber security and critical national infrastructure protection. *International Affairs*, 92(5):1079–1105, 2016.

[70] Holger Stritzel. Security, the translation. *Security Dialogue*, 42(4-5):343–355, 2011.

[71] Lucy Suchman. Embodied practices of engineering work. *Mind, Culture, and activity*, 7(1-2):4–18, 2000.

[72] Lucy Suchman, Dominik Gerst, and Hannes Krämer. " if you want to understand the big issues, you need to understand the everyday practices that constitute them." lucy suchman in conversation with dominik gerst & hannes krämer. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, volume 20, 2019.

[73] Rogier Woltjer. Workarounds and trade-offs in information security–an exploratory study. *Information & Computer Security*, 2017.

[74] Alberto Zanutto, Benjamin Oliver Shreeve, Karolina Follis, Jeremy Simon Busby, and Awais Rashid. The shadow warriors: In the no man's land between industrial control systems and enterprise it systems. 2017.

[75] Moti Zwilling, Galit Klien, Dušan Lesjak, Łukasz Wiechetek, Fatih Cetin, and Hamdullah Nejat Basim. Cyber security awareness, knowledge and behavior: A comparative study. *Journal of Computer Information Systems*, pages 1–16, 2020.

A Codebook

Below are given per theme, the individual codes reported on in this paper, together with a description of their meaning, and exemplary quotes.

A.1 Practices

Compliance, secure

Description: The act of obeying a formal cyber security policy likely to improve security.

Example: "CNIs don't necessarily understand their assets, per se. So, the water industry, for instance, might have tens of thousands of assets distributed over, 200, 300 square km. Do they know everything about every one of those assets? Not necessarily because some of them might have been put in 50, 60 years ago. They might have dropped off an asset list sometime. They might have come back on. It might have been refreshed but left there. Who knows? And they are finding out that the work, that a discovery piece in their asset management is pretty huge"

Compliance, insecure

Description: The act of obeying a formal cyber security policy likely to deteriorate security.

Example: "As the self-assessment form is subjective, it is a reflection of mindsets rather than actual cyber maturity, for example some companies are adding physical security stuff in their scope and, therefore scoring themselves higher

Workaround, secure

Description: Circumventions of a cyber security policy which do not explicitly address its problems; likely to improve security.

Example: ""There are people who are more confident, or have a different attitude to risk perhaps, and will have their own views about what is the right thing to do in their organization, and they might use NIS as a kind of sanity check, a checklist to see how they compare with it. But their real logic, decision-making will be based on their expert knowledge of what they think is the best thing to do in their circumstances, and they won't blindly follow NIS""

Workaround, insecure

Description: Circumventions of a cyber security policy which do not explicitly address its problems; likely to deteriorate security.

Example: "I didn't enjoy being a CISO because it was always going into the board saying: "You need to spend money", and the board were, like, "Well, why? Prove it, show me metrics, show me reasons", and I can scare them with regulations, but it was never a very scientific question, so to speak, it was always a bit finger in the air and, "This is what happens if we don't do it, but it might not because we might not get hit"

Above and Beyond, secure

Description: The act of exceeding the policy requirements; likely to improve security.

Example: "So, the working group is something has been going on for years now. We [regulators] are not permanent members. We were invited, of course, to be part, but this is a closed forum for operators that is running for years for them to share and it's not only about cyber but also about other topics as well to share and to experience"

Above and Beyond, insecure

Description: The act of exceeding the policy requirements; likely to deteriorate security.

Example: ""We want a highly resilient, highly available network, so that implies that you replace these assets before they stop working, and there's some subjectivity when that should be. So, there's an argument that operators put forward is, well, perhaps we should replace those assets a bit sooner than previously forecasted and at the same time we can upgrade the cybersecurity of the sensors or activators within these devices. So that potentially saves

them some money, but it's hard to draw out the separation sometimes between the cybersecurity arguments and the physical lifetime of the assets, but there are big sums of money"" hundreds of millions.""

Negotiation, secure

Description: Collaborative process, where cyber security stakeholders co-create the interpretation and implementation of the policy; likely to improve security.

Example: "one of my big bugbears with the OT cybersecurity policies is that they are very info sec driven. That's why I prefer to use the term "cybersecurity" because information isn't the asset. Traditionally, forget the computer systems. I start with something I say to everybody. There are two fundamental differences between information security and cybersecurity, OT security. Number one is that the computer system is just another component in the mechanical plant in my world. So, it has no more importance than a pump or a valve or whatever. If it breaks, the plant stops working (. . .) So the regulator looked for two water companies to work with them last year to develop NIS into something workable for the water industry. They brought out draft guidance and I sat down with them for a day, and I went through some of the things which just don't work and there's a big difference."

Negotiation, insecure

Description: Collaborative process, where cyber security stakeholders co-create the interpretation and implementation of the policy; likely to deteriorate security.

Example: "if you have several suppliers, you can actually have competitive discussions with the suppliers. If you actually use one supplier only, you don't have that anymore. One of the problems we're having in the railways on the signalling side is that we're having less and less suppliers and so the costs are spiralling because the supplier knows we don't have any other options."

A.2 Clichés

Separation means security

Description: Comments on air-gapping.

Example: "When you're talking to the board and they say, "We don't need to worry about security because our production facility is airgapped", there is only one place which is air gapped and that is Battlestar Galactica!"

IIOT is inevitable

Description: Comments on the trajectory and pace of innovation.

Example: "We're facing the problem of IIOT arriving. When we did the self-assessment, everyone was using very traditional industrial control systems. In that time in the last six months, we've all started adopting IIOT and it's going to get worse. So, it's a big change and it's one that's very much on everyone's radar including mine"

Solutions are the same across sectors

Description: Comments on sectoral differences and similarities

Example: "the tech basis of cyber is the same across the sectors"

Raising awareness leads to security

Description: Comments on the value of awareness raising activities.

Example: "General population can't rationalize. They can worry. So why would you want to worry a population? Is it beneficial for society to know of all the things that bad people want?"

A.3 Participants

Background

Description: Participants' education and previous roles.

Example: "worked in cyber security back when there was no such thing as cyber security"

Talk

Description: Discourses of participants.

Example: "Virtually, you could look at every outcome in the self assessment framework and have a debate about whether you've met it or not, purely on the basis of what sort of attacks do we expect to defend against. The framework doesn't take a position on that. So when everything's subjective, the decisions are made, are all largely made ultimately on personal opinions and it's hard to summarize, hard to understand from that just what attacks they are designed to be resilient against and what they aren't. Whereas if you have a standards based approach, standards on their own aren't the full solution, but I think they're part of the solution. Things like the [government cyber security] standard, which government has promoted. I don't know if you're aware of that."

"You've Got Your Nice List of Bugs, Now What?" Vulnerability Discovery and Management Processes in the Wild

Noura Alomar[1], Primal Wijesekera[1,2], Edward Qiu[1], and Serge Egelman[1,2]
[1]*University of California, Berkeley*
[2]*International Computer Science Institute (ICSI)*
{nnalomar, primal, edwardqiu, egelman}@berkeley.edu

Abstract

Organizational security teams have begun to specialize, and as a result, the existence of red, blue, and purple teams have been used as signals for an organization's security maturity. There is also now a rise in the use of third-party contractors who offer services such as incident response or penetration testing. Additionally, bug bounty programs are not only gaining popularity, but also are perceived as cost-effective replacements for internal security teams. Due to the many strategies to secure organizations, determining which strategy is best suited for a given situation may be a difficult task. To understand how these varying strategies are applied in practice and to understand non-technical challenges faced by professionals, we conducted 53 interviews with security practitioners in technical and managerial roles tasked with vulnerability discovery or management. We found that organizations often struggle with vulnerability remediation and that vulnerability discovery efforts are hindered by significant trust, communication, funding, and staffing issues. Based on our findings, we offer recommendations for how organizations can better apply these strategies.

1 Introduction

Security vulnerabilities have caused substantial financial and reputational damage to organizations and users over the years and continue to do so at an alarming rate, despite the availability of advanced tools for hunting vulnerabilities down [4, 6]. Manual code inspection, black- and white-box vulnerability scanners and penetration testing are only a few examples of techniques commonly used for security vulnerability detection before releasing software to users or after putting it into production [12, 18, 26, 28, 30, 33]. We have also witnessed the proliferation of bug bounty and vulnerability disclosure programs that encourage security researchers to assist organizations with vulnerability discovery. Additionally, regulatory bodies have released standards, such as ISO/IEC 29147 and ISO/IEC 30111, that guide organizations on how to design their vulnerability disclosure policies and what to consider after receiving vulnerability reports from external researchers [1, 2]. Despite all these tremendous efforts by the security and regulatory communities, the prevalence of cyber intrusions and leaks suggest that processes involving vulnerability management require urgent improvements.

We shed light on the current state of vulnerability detection and management processes by conducting 53 semi-structured interviews with managerial and technical security practitioners operating in the US and other countries. We contribute to the understanding of interactions, inter-group issues, and the challenges security teams face working with each other. Our findings are informed by industry experiences of red, blue, purple, and penetration testing teams, working as security practitioners in internal security teams or as third-party contractors, who are involved in vulnerability discovery or remediation on a daily basis. We reveal insights into how different roles fit into organizational security strategies and how human factors affect vulnerability management processes.

We interviewed personnel with experience managing and implementing the security posture of their organizations. We identify challenges managerial personnel face and the factors they consider when they make decisions on vulnerability discovery and management (e.g., creating bug bounty programs, outsourcing tasks, etc.). We also identify the challenges professionals from red, blue, purple, pen testing, and bug bounty teams face in practice and how management decisions affect their work. Our study attempts to piece together the perspectives of different actors involved in vulnerability discovery and management processes and identify challenges that require immediate attention by the security community.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

Vulnerability discovery and remediation processes are hindered by a host of behavioral and organizational challenges. Our results also show:

- an inherent *trust* problem between organizations and testers or bug bounty hunters.
- a lack of effective communication channels between testers who discover vulnerabilities and those responsible for taking proper remediation actions.
- a lack of clarity around who is responsible for fixing discovered vulnerabilities, which turns vulnerability management into a "blame management" problem.
- a misalignment between security and business priorities, organizations represented by our participants often took reactive rather than proactive approaches to vulnerability management.
- a tendency to adopt *compliance-oriented* approaches to security rather than grounding vulnerability handling processes in an understanding of potential risks.

We believe that our findings can propel further research to better understand the human factors impacting vulnerability discovery and management processes in the wild.

2 Related Work

Howard and Lipner [23] studied the security development lifecycle in organizations and observed that implementing security process improvements requires commitment from upper management, which is often hard to get due to the difficulty of associating security improvements with concrete business benefits. Another line of research has focused on the economics of bug bounty programs, the role of white-hat hackers in vulnerability discovery, and how to encourage participation in such programs [13, 17, 19, 22, 25, 27, 34, 36, 37]. Finifter et al. highlighted vulnerability reward programs' potential for helping organizations discover more security vulnerabilities than internal teams [19]. Zhao et al. found that the diversity of contributions in bug bounty platforms correlates positively with the number of discovered vulnerabilities [36]. Maillart et al. found that current incentive structures on bug bounty programs attract security researchers to newly released programs [27]. Researchers also found that the number of vulnerabilities reported by bug bounty hunters correlates positively with the monetary payouts disclosed in program policies [37]. Chatfield and Reddick reported on the Pentagon's successful experience with outsourcing security testing through bug bounty programs [17]. Votipka et al. found that hackers and testers follow similar approaches when searching for vulnerabilities [34]. Building on this work, we combine the perspectives of bug bounty hunters and security managers to shed light on the factors that organizations consider before deciding to create bug bounty programs, and the expectations of bug bounty hunters from bug bounty program managers.

Haney and Paul [21] conducted 14 interviews with members of the blue and red teams at one public organization in the US and identified a number of challenges that stem from a lack of effective information sharing between these teams. Botta et al. found that most organizations follow a decentralized approach to managing cybersecurity and highlighted the need for approaches that facilitate reporting and information sharing across security teams [11]. Björck et al. interviewed security practitioners in Sweden to explore how to address the cost-inefficiencies associated with security management [10]. Werlinger et al. examined incident response processes and emphasized the need for tools that facilitate active collaboration between security teams whenever a response to a security incident is needed [35]. Consistent with our findings, Thomas et al. demonstrated the communication challenges that security and development teams face in practice, and emphasized the role of automation in helping security teams address security issues earlier in the Software Development Lifecycle (SDLC) [32]. Ceccato et al. conducted an experiment to evaluate the usefulness of reports generated by static and dynamic vulnerability detection tools for maintenance teams who are tasked with developing security patches [15]. Assal and Sonia highlighted the need for lightweight approaches to security that can help organizations adopt sustainable approaches to managing security throughout the SDLC [3, 32].

Other research has focused on privacy processes in organizations (e.g., [5]); however, we consider this orthogonal to our study, as the strategic decisions made throughout vulnerability management processes are of higher relevance to securing an organization's entire technical infrastructure, while taking into account human and procedural aspects. Several studies have also addressed the problem of how to improve employees' compliance with security policies, and help security managers assess the security culture in their organizations [7–9, 29]. Furthermore, Stevens et al. [31] conducted a thorough investigation of three security compliance standards and identified 148 security concerns that might render an organization that is considered compliant vulnerable to high severity risks.

While prior work has examined certain individual aspects of vulnerability management processes, our focus is on gaining a holistic understanding of the pipeline, highlighting issues that relate to organizations' strategic planning of vulnerability management, and describing how different teams' efforts help organizations improve their security postures.

3 Methodology

From April to June 2019, we conducted semi-structured interviews with 53 security professionals in different security testing roles (described in the next section). We also recruited current and previous managers who had made important decisions on vulnerability discovery processes in their organizations. We used several channels for recruitment, including advertising on Twitter, and two Slack workspaces (*bugboun-*

tyworld, and *bugbountyforum*), which are used by testers and bug bounty hunters to share resources, collaborate, and communicate with bug bounty platform employees. We also sent emails to several pen testing companies, inviting their security testers and managers to participate in the study.

We conducted six pilot interviews to validate our interview process. Participants recruited for the pilot interviews were contacted via Twitter or through professional connections. One pilot participant did not consent to use his interview data; hence, for the remainder of the paper, we report on the data from 5 pilot interviews and 48 recruited participants. Our study was approved by the relevant Institutional Review Board (IRB). As a token of appreciation, we offered a lottery drawing for five $100 Amazon gift cards.

We asked participants to complete a pre-interview questionnaire, which asked about their security testing experience and their current and prior roles. Next, we conducted a 45-70 minute interview via online video chat software and recorded the audio for transcription purposes (see Appendix A). The lead researcher led every interview, and another researcher participated in an observatory capacity. We conducted interviews until the themes for each category reached saturation [16], which was independently determined by two researchers.

We tailored our semi-structured interviews to participants' roles and prior security testing experiences. The interview process for non-managerial participants evolved around the discovery processes they follow, tools they use, how they communicate with the rest of the teams, challenges they face that might hinder their work, nature of the vulnerabilities they work on, and their expectations. For those with managerial experience, the interview covered their experience working with different teams, reasons for deploying specific strategies in their managerial capacity, and their expectations when deploying different security testing strategies.

Our sample included blue, red, and purple teamers, penetration testers working in internal security teams, security consultants, bug bounty hunters, and security managers (e.g., CTOs and CISOs). At the time of our interviews, a sizable portion of engineers and managers that we interviewed were employed by some of the world's largest tech companies, serving billions of end-users. Twenty of the participants had more than ten years of experience doing security testing, eight participants had between 6 to 10 years of experience, 21 participants had 1 to 5 years of experience, and 4 participants had less than one year of experience doing security testing. Many of our participants held multiple previous positions. For example, some had industry experience in red teaming and pen testing at the same time or held managerial positions after spending some time working as security engineers (see Appendix C). In particular:

- 38 worked as internal or external penetration testers;
- 28 had experience doing bug bounty hunting;
- 23 had held internal or external red teaming positions;
- 17 reported experience working in blue teams;
- 17 were CISOs, CTOs, or engineering managers;
- 11 had experience working as purple teamers; and
- 5 reported working as quality assurance engineers.

The majority of our participants came from the US, but we also had participants from Canada, Germany, UK, India, Israel, Singapore, Serbia, Brazil, and Bangladesh. Some participants managed their own companies, offering external pen testing services. We asked each participant about the number of employees in their organizations and their security teams. For around 40% of our participants, the number of employees in their organizations were in the thousands; eight worked in organizations with 100 to 5,000 employees, while the rest had 1 to 70 employees in their organizations.

Three external transcribers transcribed the interviews. Next, the lead researcher and two other researchers iteratively built the codebook. We used basic thematic coding, wherein each code mapped to an emerging theme (e.g., *blue-red-conflict*, *bugbounty-trust*). Our codebook has 32 major code categories based on the role of the speaker, and those categories are further divided into 88 sub-codes (see Appendix B).

Once the codebook was finalized, the lead researcher and another researcher independently re-coded the interviews and merged the codes to calculate the inter-rater agreement. Disagreements in coding were discussed and resolved after reaching consensus among all the coders. Our recorded inter-rater agreement was 85.36% (Cohens' $\kappa = 0.71$), indicating substantial agreement. We arranged codes into the three phases of the pipeline: strategic decision making, vulnerability discovery, and vulnerability management.

4 Security Teams

There are a number of security roles that engineers might be tasked with in organizations. Many of our participants seemed to be confused as to how to define their work and how their role contributes to vulnerability discovery. For instance, a director of security who manages several security teams and is responsible for product security at a tech company that serves hundreds of millions of users responded:

> *"Yeah pen testing, red teaming, those are very confusing terms. What would you say is the meaning of pen testing?"* (P52).

A CTO of a security consultancy offering vulnerability discovery services to organizations mentioned that many of his clients think that they should do red teaming or pen testing, when what they need is an incident-response team:

> *"I have had clients that have suffered a breach in the past and immediately they are like, 'Hey, we want to get a penetration test or a red team to figure out how the attacker broke in'"* (P37).

We argue that not being able to distinguish between the roles different teams are expected to play might lead to making sub-optimal decisions and ultimately rendering vulnerability discovery processes ineffective. Below, we present a

brief definition for each of the different security testing strategies based on themes found in participant responses and our review of the related literature (e.g., [24]).

Internal Testing. We define this role as part of a quality assurance process in which people with security backgrounds either do testing or work with developers to build secure code:

"We have kind of check points along the development process. We also have kind of programmatic deployment gates that a handful of these tests must be met before things can even be deployed" (P10, security manager tasked with managing internal testing teams at a large organization).

Pen Testing. The goal of penetration testing is to find as many vulnerabilities as possible, focusing on a finished product or a service within a pre-defined period:

"I see penetration testing as audit. So it is assessing for all vulnerabilities, at every single point in time, within a defined scope" (P40, a full-time bug bounty hunter with extensive experience as a pen tester and a red teamer).

Red Teaming. Red teaming is goal-oriented and is about going deep (compared to breadth coverage in pen testing); that is, infiltrating the network in question and exploiting any vulnerabilities in the process to reach a pre-defined goal. It is commonly understood that red teaming emulates real-world hackers in the most realistic way possible:

"In a red team exercise, you are working with no knowledge of the product. It is as if you are an external hacker kind of thing, trying to hack without access or insider knowledge" (P31, security engineer tasked with doing internal red and blue team exercises within a large public organization).

"The red teaming is really more, hey, we are not interested in identifying every stage of vulnerability in the network, we are interested in seeing how far can we get, what type of data can we access, can we actually exploit the things that are most important to the organization" (P30, a security manager who has extensive experience leading red teams).

Blue Teaming. Blue teaming is about being defensive—in contrast to red teaming—and concerns monitoring target systems for abnormal behaviors that might indicate the presence of an adversary in the system or an intrusion attempt. Blue teamers are expected to collaborate with red teamers in order to make sure that their efforts complement each other and that they are improving their detection over time:

"[A] blue teamer is exactly the opposite, is the defending side. This means finding solutions, techniques, procedures to stop hackers, stop red teamers from succeeding in their attacks" (P39, an experienced blue teamer).

Purple Teaming. Purple teams might consist of members from red teams and blue teams. The most significant advantage of a purple team is to have both offensive and defensive teams work together to achieve a common goal:

"The purple team is the one that navigates between the red team and the blue team and sits with them together as a sort of mediator" (P39, an experienced blue teamer).

Bug Bounty Programs. Bug bounty programs can be seen as ways to crowdsource pen testing on publicly-facing systems without any time restrictions. Bug bounty programs invite the public or a group of hackers selected by an organization to test its systems according to a pre-defined scope and policy. HackerOne and BugCrowd are examples of bug bounty platforms that act as intermediaries between organizations and external security testers [14, 20].

5 The Vulnerability Management Pipeline

Based on the themes that emerged from the interviews, we identified three phases of an organization's vulnerability management pipeline: from the moment the organization's decision makers realize that they need security testing ("strategic decision making"), to the methods chosen to perform that testing ("vulnerability discovery"), to when discovered vulnerabilities are fixed ("vulnerability remediation").

Strategic Decision Making. We wanted to understand what decision makers expect from the various security testing strategies and what drives them to deploy a particular one. Our goal was to understand whether management are making correct decisions that serve their organizations' goals, whether those organizations are ready for a chosen testing strategy, and whether there are any cultural and/or organizational barriers.

Vulnerability Discovery. In this phase, the chosen security testing teams begin searching for vulnerabilities in target systems. Different security testing teams have different goals, expectations, and time restrictions. Further, some teams could be better suited for specific categories of companies. We believe that it is important to understand the activities carried out by each team and the expectations for each role.

Vulnerability Remediation. This phase concerns deciding how to fix the uncovered vulnerabilities in a timely manner to mitigate potential risks. An effective vulnerability management process entails having a systematic approach to prioritizing discovered vulnerabilities according to their potential impact and having clear communication lines between security testing teams and any other stakeholders.

6 Strategic Decision Making

The potential success of a vulnerability management pipeline hinges upon utilizing the correct vulnerability discovery strategy at the correct time. In this section, we present findings on participants' stated motives and other factors that go into their decision processes. We focus on understanding: (1) when to hire a red, blue, purple, or pen testing team, (2) when to create a bug bounty program, and (3) what considerations are being made before deciding to outsource security testing.

6.1 Deciding Between Different Teams

Below, we present the factors that influence the managerial decision of which security testing strategy to implement.

Red Teaming vs. Pen Testing. We observed uncertainty among our participants about when to do red teaming versus pen testing. The majority of our participants who are considered decision makers did not see the need to do red teaming when their organizations have regular pen testing engagements. Participants expressed different motives for pen testing, a commonly recurring one was to meet various compliance or certification requirements:

"Pen tests are kind of a reproducible formula for getting the engagement done whether it is to get that compliance checkbox checked or they are just kind of going through the motions to make sure that basic security structures are in place"(P10, security manager tasked with vulnerability management operations at a large organization).

Red teaming involves going in depth and emulating a malicious actor. Participants mentioned that another way to look at it is as testing all of the monitoring capabilities installed by the blue team. Thus, to have a fruitful red teaming engagement, an organization should have a reasonably mature security posture with in-depth monitoring:

"If my goal is not to find as many vulnerabilities as I can, my goal is to validate my security controls that I have put in place, then I would love to hire a red team" (P41, a former red teamer and a pen tester who is currently managing application security programs at a large organization).

Regardless of the engagement exercise, participants commonly agreed that neither red teaming nor pen testing should be the first approach to testing in an organization or the first step in their vulnerability management pipeline:

"I think most organizations are not mature enough to have true red team exercises performed. I don't think having a red team or having a penetration testing team should come before having an effective vulnerability management program" (P41, a former red teamer and a pen tester who is currently managing application security programs at a large organization).

We observed that there is a common preference for doing pen tests before putting software products into production and whenever major changes to codebases are made. Red team engagements come at a later stage, when organizations have already tested their products internally and set up monitoring and detection, performed the required pen tests, and are willing to validate the security controls they have in place. However, our results suggest that concerns around exposure to legal liability might drive reluctance around the use of red teams by organizations. Some participants expressed that it would be risky to let external red teaming contractors test their assets without being able to have full visibility into their activities, as they had concerns that their sensitive data might be leaked as a result. This might also create a potential conflict between duty-to-report rules for red teamers and the actual practices of companies. A former red teamer mentioned:

"You don't know when someone is in the network, you don't know what he is reading. I know that there are NDAs, but sometimes there is a code of ethic, that says if you find something illegal during a penetration test, you have to report it. And in red teaming, let's be honest not many companies are in regulations with the law" (P24).

Blue and Purple Teaming. Our participants agreed that organizations should have a blue team that is continuously ready to detect intrusions and respond accordingly. A security director responsible for vulnerability management operations at a leading tech company mentioned:

"The first thing that you need to hire is the blue team. What are we defending? What are the state of our systems? What is our attack surface? That stuff you have to first understand" (P52).

Though, the majority of the participants mentioned that, in most cases, blue teams are not set up until a breach incident happens, which causes upper management to realize the severe consequences of cyber-attacks and the importance of proper security posture. Once an organization properly implements the necessary security controls, red teaming can be utilized to evaluate the extent to which such controls can be exploited by the adversary and help in strengthening blue teams over time. Purple teaming, on the other hand, could be beneficial once a blue team is perceived as not maturing and learning from the results of prior red teaming engagements. However, considering that it would not be realistic for small or medium size organizations to have red, blue, and purple teams, a considerable number of our participants mentioned that they would prefer to start by setting up a purple team, where its members have red team and blue team backgrounds.

Bug Bounty Programs. We found different motives for creating a bug bounty program. Many participants mentioned positive experiences; having many eyes looking into their systems and a large pool of testers with diverse backgrounds can yield interesting vulnerability reports. A security director who started a bug bounty program at a large company, and slowly expanded their scope to cover their vendors as part of their security boundary, pointed out the high costs of pen testing services as the primary motivator for their decision:

"One pen tester for one week is going to cost you a lot of money and you can get much higher coverage with just that money and many more people involved" (P52).

A few participants also had reasons for not creating bug bounty programs. Particularly for financial and governmental organizations, a common theme that emerged in our interviews was that they do not want to risk having unvetted "people from the Internet" accidentally gain unauthorized access to sensitive data stored in their systems. They mentioned that they are more comfortable working with established external security firms, where they can keep tabs on people who are allowed to test their systems for security vulnerabilities:

"...trying to convince any government agency to try to offer people a reward to try to break into their financial data, I think we would have a very bad reaction to that, whereas saying, 'it's an audit' would be a much easier sell" (P44, security leader of a consultancy company that works as a government contractor).

Some participants mentioned that it is often difficult to justify the costs of creating a bug bounty program to upper management, as the costs associated with running such programs are not fixed, relative to paying for an outside audit. Some participants also mentioned that they hire external contractors to conduct pen tests to satisfy regulatory requirements and to convince their customers that they are sufficiently secure because they have been tested by an outside entity:

"..the quality of that report you get from pen testing companies is much much higher. It can also hold people who did the pen test accountable" (P32, a security engineer tasked with red teaming and pen testing at a large security consultancy company).

Participants who had created bug bounty programs mentioned that before creating a bug bounty program, organizations should make sure that they have a robust process for handling reports received from bug hunters: assessing them for validity, reproducibility, severity and impact, and then responding to the bug hunter promptly. Participants said that starting a bug bounty program could easily overwhelm the organization with a lot of noise (false reports). There were also concerns as to whether organizations, especially management, clearly understood the prerequisites:

"My CISO likes to be able to tell customers we have a bug bounty program; therefore we are very secure" (P41, application security manager).

Due to trust-related concerns, several participants preferred to create private bug bounty programs, where the organization gets to pick the bug hunters who are allowed to work on that program after doing proper background checks. One participant also mentioned that his organization decided to do all their bug bounty work on their staging environment, where they do not risk bug hunters getting access to their customers' data, and can make sure that they can reset the testing environment whenever anything goes wrong. Other organizations might prefer to narrow the scope of their public bug bounty program to include a few of their assets and expand the program scope incrementally as they go forward.

6.2 Factors affecting decisions

Below, we summarize other factors that might potentially influence decisions on vulnerability management.

Internal Teams vs. External Vendors. We observed common concerns related to the costs and difficulty of building and maintaining internal teams, which were justifications for outsourcing security testing. The majority of our participants agreed that having internal teams is more beneficial in the long term, as testers accumulate institutional knowledge about their organizations' systems and build understanding of their business logic in more depth over time, as well as have conversations with developers and build relationships with other security team members. Having internal teams also helps organizations adopt a proactive approach to vulnerability discovery. An engineer with more than 10 years of experience in red teaming, pen testing, and managing security teams said:

"The internal folks come more or less over time, with a built-in understanding of the business objectives, and the business logic that needs to be expressed. The external folks by definition, don't understand the business as well. And so sometimes their findings are not going to be as relevant to the company" (P17).

"A lot of folks struggle with external bug bounties because folks outside the company will report something and there's no clear implication to the business. And you cannot automatically say this is not a problem. But then you have to decide am I going to spend time even figuring out if this is a problem, versus someone who comes to the table with that work already done for you, then you can immediately start remediating" (P17).

However, participants also mentioned that external testers can help discover vulnerabilities that were previously not noticed by developers or internal testers, since they often have a lot of exposure to different troubleshooting scenarios and clients. Furthermore, many participants highlighted that most security engagements with external testers are time-scoped and hence bear less cost and logistical overhead compared to maintaining internal teams. This was especially the case for small organizations and is the primary reason why organizations may hire external vendors rather than build and maintain internal teams. Said an experienced red teamer who now works as a consultant in blue and red teaming:

"So if I am a company the size of Walmart, I am going to have my own red team. If I'm a small organization, and I don't have the money and IT isn't my focus in the first place, I am not going to try to build a red team and then I would definitely contract with a third party to do it" (P30).

Furthermore, due to trust-related reasons, participants also mentioned that external testers normally do not get to access every resource they need in order to cover significant parts of their clients' attack surfaces.

Selecting and Vetting External Testing Firms. Participants expressed a tendency to rehire the same external firms:

"We, our blue team, prefers a particular pen testing team. We have good relationships with them, their diagnostics and write ups have been just superb" (P9).

Other explanations included:

- they had a positive experience with a particular firm and built relationships and trust with its members;
- they were satisfied with the level of diagnostic investigation and the quality of write-ups they received;
- they did not want to waste a lot of time vetting different external firms;
- they preferred to work with firms that have a good reputation in the industry; or
- they perceived difficulty in obtaining visibility into how they are improving their security over time once they switch vendors, as different vendors have different methodologies and approaches to security testing.

A few participants, however, mentioned that they prefer to rotate through a pool of vendors to:

- fulfill regulatory requirements;
- get specific testing expertise (e.g., testing cryptographic solutions); and
- get different sets of eyes to test the same codebase.

Regarding the latter, an experienced pen tester stated:

"I think it is good to switch from time to time because other people have other techniques and will most likely find other or new issues" (P15).

Staffing and Budget. Participants mentioned the difficulty of justifying security-related staffing or budget decisions. Participants with management responsibilities admitted that staffing is a big problem, and it is not a luxury that many can afford. In response to why they do not have a bug bounty program, the CISO of a large organization responded:

"We already know about more risks than we have the capacity to deal with" (P1).

This might lead to a lack of visibility into unpatched vulnerabilities and a false sense of security until an incident occurs. Participants also expressed the difficulty of convincing upper management to provide security engineers with the required freedom to do their assessments and find vulnerabilities. Others mentioned that there had been situations where their managers decided to build a security team very quickly to react to security incidents without proper planning and forethought:

"...out of the news, all of our [security] costs are basically viewed as not required and extravagant. And it is always a fight to get funding, even though when there is a problem suddenly funding is free and how many people do you need" (P9).

Participants mentioned that, in some organizations, monetary allocations might give upper management false assurances about the organization's security posture. In cases where an organization has an annual budget allocated for security, upper management might be under the impression that they are *secure* when a regular security assessment with an external contractor is only done as a formality to *'tick the boxes'* and have a report that says so. Said a pen tester:

"The CEO does not want to know if you have SQL injection, or XSS. He just wants to see in the report, that we have spent that amount of money on the security, and he always thinks like, 'oh we have spent a lot of money on security. We must be secure!' " (P24).

Many mentioned that they had to frame everything around a dollar value to get attention. That is, some security managers expressed the importance of conveying risks to the business and describing the impact of potential legal liabilities or monetary losses to upper management:

"When you say that this vulnerability has a CVSS score seven, they don't get it. When you can say, like, this risk represents an expected loss of a hundred fifty million dollars, and they're like, 'okay we know what to do with that' " (P17, security engineer).

Other managers also expressed difficulty in building security teams that have combinations of qualified people with different areas of security knowledge; e.g., network security, mobile security, and cryptography. Some security managers also mentioned the difficulty of making sure that discovered vulnerabilities are addressed promptly by their qualified security engineers, as they are worried they might lose their personnel once they are not satisfied with their workload.

Considering the costs of setting up teams of qualified personnel, several different approaches to managing security in small organizations emerged in our interviews. Outsourcing security to external contractors, using open source security testing tools to automate some tasks, leveraging the security teams of well-known cloud providers by putting most of their security services in the cloud, creating private (invite-only) bug bounty programs, and hiring a few security engineers with purple teaming experience are the main approaches that the participants brought to our attention. Particularly for bug bounty programs, decision-makers working in small organizations were concerned that qualified security researchers would not be attracted to work on their programs, given that they would not be able to compete with organizations that have mature bug bounty programs and provide high payouts. This observation was confirmed by several bug bounty hunters:

"I have got two or three programs that I have spent a lot of time on and I keep going back to them because I enjoy working with the team, I know that they pay fairly well and they will turn around my bugs pretty quickly" (P40, bug bounty hunter).

Some participants mentioned that they have a small team of security engineers wearing multiple hats (i.e., doing offensive, defensive, and quality assurance work at the same time).

Scoping Considerations. Our discussions with internal testers, security managers, bug bounty hunters, and external contractors revealed that there is often uncertainty in how to scope security testing. When defining the scope of a bug bounty program, many decision makers preferred to focus on detecting vulnerabilities that can be reached through publicly available services. Some bug bounty hunters also mentioned that they have accidentally stumbled upon critical vulnerabilities that are considered out of scope, and decided against reporting them, because they fear that the organization might take legal action against them. This fear was confirmed by some bug bounty program managers, who expressed that they would not welcome such discoveries, as they do not want to incentivize bounty hunters to go out of scope.

Other program managers expressed that they would welcome such submissions, as it signals that the bug bounty hunter cares about helping them improve their security posture. Other factors that might influence program managers' decisions include the budget for paying bounty hunters and the capacity of their internal triaging team to process incoming reports. A broader scope is likely to incur higher costs, just in terms of coping with more reports.

In contrast to bug bounty programs, scoping red teaming and pen testing engagements is mostly focused on organizations' internal assets. Although no-scope red teaming engagements are perceived to be common for large organizations, some participants mentioned that narrowing the scope of red teaming could be useful in cases where an organization wants to validate certain controls that they recently implemented. We also observed that some organizations prefer to focus on a specific area at a time; e.g., for the first couple of months they will focus on a specific set of systems and then test other systems later. For pen testing, the majority of our participants indicated that no-scope assessments are often expensive, and this might probably lead them to narrow the scope of systems to be tested in order to afford their costs.

Other security managers mentioned that they are not confident with letting external researchers test every system they have, leading to narrow scopes for pen testing or red teaming engagements. For the same reasons, some organizations resorted to creating duplicate environments, where they put all their code in a live environment, but with sensitive data removed to mitigate the impact of any potential data leakage. From the perspectives of pen testers, we noted that narrowing the scopes of pen testing or red teaming assessments could make the vulnerability discovery process very restrictive and completely counter-intuitive to what the vulnerability discovery process should achieve. Said a red teamer:

"...but like it gets to the point where they chop everything off to the point that you don't have a bite anymore and, you know, you want to try as hard as you can, just like any bad adversary" (P42).

7 Vulnerability Discovery

Vulnerability discovery can start from the earliest phases of the software development lifecycle. It can be part of each cycle or can be brought in after each major code release. An organization can have a dedicated security team testing all of the ongoing projects or assign a team member(s) to work with each development team(s). Below, we outline the processes security teams follow to discover security vulnerabilities based on a synthesis of common understandings in the industry and our observations from the interviews.

Pen Testing Activities. Pen testers aim for breadth by finding as many vulnerabilities as possible within a predefined scope and time frame. Participants also mentioned that they often get the help of external pen testing contractors to test their products before shipping them to users. Pen testers are not supposed to hide their activities from other members of the organization and are mainly expected to demonstrate the vulnerabilities they discover to prove that they exist, without exploiting them. However, one external pen tester mentioned that clients sometimes ask them to exploit identified vulnerabilities to demonstrate their impact. Furthermore, most pen testers' activities are supposed to be continuously monitored by members of the client organization to limit the consequences of any potential unintended exploitation.

Red Teaming Activities. One of the main goals of red teaming is to help blue teamers improve the defenses they have put in place. Pen testing and red teaming bring different benefits to an organization. Red teaming engagements normally take longer than pen testing engagements. Pen testers are usually given access to internal resources, whereas red teamers start their engagement with minimal to zero information. For instance, some participants mentioned that pen testers usually have internal network access, whereas for red teamers, gaining access to internal networks is a goal they should achieve. Furthermore, pen testing engagements are not stealthy, and many in an organization might be aware of them, whereas red teaming engagements are stealthy by nature to evade any monitoring or access control mechanisms placed in the organization and to effectively simulate the activities of malicious attackers. For instance, some of the tactics red teamers follow to decrease the risk of detection are:

- not installing custom software;
- trying to use system functionality as much as possible, rather than using exploits;
- sticking to regular user working hours, so a blue team would not suddenly start noticing loads of traffic;
- rotating their IP address or hiding them behind different VPN providers; and
- not scanning many ports at once.

Participants explained that there are basic steps that are usually followed by red teamers. At a very high level, the first phase is reconnaissance, which involves gathering as much public information about the target as possible. Based on that, red teamers can draw a plan outlining the potential targets that they can attempt to attack or utilize to exfiltrate data. For instance, they might send phishing emails in the hope that someone will click on a malicious attachment, which will then help them establish a foothold into the network. They could also examine whether lateral movement is possible and try to plant a backdoor in order to maintain persistence. To effectively simulate real-world adversaries, red teamers usually ask organizations not to inform their employees—including blue teams—that these offensive security engagements are taking place; once blue teamers detect them or they successfully evade a blue team's security controls, they communicate their findings and provide guidance on how to fix the identified vulnerabilities. Red teams need to have broader skill sets, including social engineering, physical security, and network security. Participants also mentioned that pen testing experience would be beneficial for red teaming.

Blue Teaming Activities. Blue teaming is concerned with fixing vulnerabilities found in the discovery phase or implementing proper monitoring capabilities to prevent intrusions. Some examples of blue teaming activities that were brought up in our interviews are installing firewalls and anti-virus software, monitoring systems for suspicious activities, responding to cyber incidents, trying to confuse potential adversaries once they are detected in order to gain an improved understanding of their activities and capabilities, and auditing and analyzing logs. Some blue teamers also mentioned that most of their work is concerned with fighting bad security culture, convincing people to install patches and teaching engineers the best security practices. It is also worth noting that, in contrast to red teams, blue teams usually have access to significant internal resources and are therefore expected to know the specifics of what they are defending. For this reason, asset management is perceived as a challenging task for internal blue teams. That is, they cannot protect systems that they do not know about.

Purple Teaming Activities. We observed common confusion surrounding what purple teams do in practice; some thought the term signals that an engineer has red and blue team experience, while others thought that applying offensive and defensive security practices at the same time can significantly contribute to hardening defenses in organizations. While many of our participants mentioned the difficulty of building a purple team, we noted that small organizations prefer to start with setting up purple teams without having separate blue and red teams due to decreased staffing costs.

In organizations with red and blue teams, purple teams can act as middlemen who facilitate communication between the two teams by making sure that they have regular meetings and regularly exchange information. Other organizations prefer to assign some members of the red team to work closely with blue teamers, to train them to detect the attacks mounted by red teamers or reflect on the regular feedback they get from red teamers. Another way of doing purple teaming exercises is to have blue and red teams work very closely, where all team members are involved in the same exercises. In such settings, once a red teamer manages to break defenses, the red teamer works with blue teamers to address the identified vulnerability. Afterwards, the two teams can resume their work and reflect on their progress.

Bug Bounty Hunters' Activities. Bug bounty hunters consider reconnaissance a critical phase in their vulnerability discovery processes, in which they attempt to collect as much information as possible about the target systems. This phase includes activities such as enumerating all sub-domains, using a service as a regular user to understand its functionality, parsing public datasets, understanding data flow, and identifying dependencies between different features or services. Some examples of the techniques used for vulnerability discovery that we identified in our interviews are disabling certificate pinning, reverse engineering apps, looking for default credentials, discovering old services, and watching for DNS changes.

Inter- and Cross-Team Communication. One of the main themes that emerged in the majority of our interviews with security engineers concerns the communication challenges security teams face with other internal teams. From the perspectives of red teamers and pen testers, they mentioned that their findings are often not welcomed by blue and development teams and that they are generally perceived as threats to other teams; red teamers' findings might make development/blue teams look bad to upper management or potentially increase their workload. Our results therefore suggest that establishing cooperative and collaborative relationships between blue and red teams seems to be a serious challenge.

In most cases, this might turn into a *"blame-management"* problem, in that each team thinks that fixing the discovered vulnerabilities is not part of their job. Several participants mentioned that they had had situations where they decided to report their discoveries to upper management to ensure that the identified vulnerabilities got fixed. From the perspectives of external contractors, they mentioned that they often report their findings to the person who hired them and might not get a chance to interact directly with internal teams who are in charge of fixing discovered vulnerabilities. Participants also noted that red teams are supposed to not disclose their activities or share much information about their testing strategies to make their simulations realistic, and this can be frustrating for blue teamers. From a managerial perspective, how security teams are organized and structured could introduce some communication problems. Said a security manager:

"..a lot of organizations diversify teams too quickly so where they have five or six people on a dozen different islands with different names. And they lose a lot of the collaborative potential of the team whenever they're segmented off that way" (P10).

To address these communication problems, participants mentioned some strategies followed by their organizations:

- situating security teams within engineering teams;
- letting different security engineers test the same product at different points in time over the year and comparing their findings to keep track of what vulnerabilities have been fixed and what still needs to be done;
- setting up a purple team; and
- holding regular security-related meetings with representatives from each team in the organization.

Some also raised the point that having clear objectives and increasing the level of transparency with security engineers could make vulnerability discovery processes more fruitful.

From the perspective of bug bounty hunters, participants have had a mix of positive and negative experiences when reporting their findings to internal teams. Many of the bounty hunters we interviewed mentioned that the bug bounty ecosystem has allowed them to build productive relationships with internal security teams at various companies, allowing them to collaborate on fixing reported vulnerabilities and receiving prompt feedback. Some mentioned that their work as bug bounty hunters increased their chances of getting hired at the companies to which they reported vulnerabilities. Several participants also mentioned that they had situations where internal teams ignored their vulnerability reports or considered the reports as out of scope because the internal teams did not triage the reports correctly.

8 Vulnerability Remediation

Below, we summarize our observations that relate to fixing the vulnerabilities uncovered during the discovery phase.

Triaging Vulnerability Reports. Internal teams have to triage vulnerability reports they receive in order to assess the severity level of each reported vulnerability and prioritize fixing the critical ones. Participants described their experiences with bug bounty programs that assigned non-technical people to the task of triaging vulnerability reports, which resulted in closing their reports as insignificant when the triaging team did not fully understand the impact of the findings.

Fixing Vulnerabilities. With regard to addressing vulnerabilities, one main theme emerged: internal teams' inabilities to fix reported vulnerabilities due to the lack of detailed information in the reports that allow reproducing the vulnerabilities and thus make informed decisions on how to fix them. For example, bug bounty hunters mentioned that patches applied by internal security teams could often be bypassed. For this reason, some security managers mentioned that they sometimes offer extra bounties or reputation points for bounty hunters willing to help with remediation. However, participants mentioned that providing extra incentives is not common and that the current bug bounty ecosystem does not incentivize external researchers to go beyond surface-level checks to ensure that reported vulnerabilities have been properly fixed:

"There is one program that I've worked on that specifically says you will receive a high bounty if you provide a remediation advice. So it is not something that I see across the board in having recommended fixes, I tend to think that the teams themselves are in a much better place to know how to fix it, that I'm here to explain the issue in a way that they can understand it" (P40, full-time bug bounty hunter).

Others thought it would be helpful to ask bug bounty hunters to retest the fixes released by companies:

"I think it's good to send it back to you after they fix it and ask you to check it because they might have put a patch that wasn't fully secure. So I think that is a great way to just get confirmation that it is fixed" (P25, bug bounty hunter).

A participant with extensive pen testing experience stated that external security testers usually lack the specifics needed to describe how to fix identified vulnerabilities. On the flip side, from internal teams' perspectives, some participants stated that they often do not have a clear idea of how a vulnerability reported by an external tester was discovered in the first place, which is one of the reasons why they might have no way to validate the fix. Other participants mentioned that they arrange for external pen testers to meet with internal teams to discuss possible remediation strategies. External pen testers might also be asked by the organization to do another pen test after a certain period to check whether the identified vulnerabilities still exist.

From security managers' perspectives, many stressed the importance of providing comprehensive and detailed mitigation strategies by security testers to allow internal teams to find an alternative approach to fixing a vulnerability, once they find it infeasible to fix it in a particular way.

When discussing the extent to which internal teams are capable of addressing vulnerabilities promptly, an experienced blue teamer explained:

" a lot of them don't know how to test whether or not the patch solved the problem. I think that the biggest reason is skill" (P9).

Other barriers hinder the remediation process. For example, internal teams might identify dependencies between different applications that might involve other internal departments or external vendors, or require hiring additional software developers with special expertise. This might decrease the likelihood of applying fixes promptly. Some security managers also mentioned that lack of transparency between all the internal and external parties involved in security processes is often an obstacle towards effective vulnerability remediation:

"We have trained our engineers to not come to us, but we have lost visibility of everything else that is happening, and now we are, back to a super reactive mode, where we want to be in a super proactive mode" (P43).

Compliance vs. Security. Depending on the industry, there are specific compliance requirements that must be met in order to make sure that the organization is doing the bare minimum to secure their users or customers. Such requirements have made organizations to adopt compliance-oriented rather than security-oriented mindsets. Most of the participants who have experience working as external pen testers described their experiences with organizations that ask them to downplay the severity rating of a critical vulnerability in the reports they plan to submit to regulators or to take some vulnerabilities out of a report when they do not have enough time to fix them.

This is made worse when pen testers are pressured by their managers to comply with these client requests to maintain relationships. These external pen tests are sometimes perceived as a way to shift the liability and accountability for security breaches from the client organization to the pen testing contractor. Said two pen testers:

"We've been told we don't want you to actually solve this problem, we just want you to make the check box go away" (P9).

"I think that turns pen tests into commodity and not clients that really want to understand the exposure and ultimately not necessarily interested in actually fixing things and getting a better security" (P38).

9 Discussion

In this section, we analyze the challenges that came up in the interviews, which we believe can be addressed by paying more attention to human factors.

Security Is a Reactive Measure. Security testers and engineers mentioned that convincing management that security should be a priority is hard in many organizations. Organizations that struggle to coordinate to place security as a high priority often spend more time reacting to incoming vulnerabilities than proactively searching for vulnerabilities. The reactive approach to vulnerability discovery and remediation may increase the chances of organizations falling victims to security attacks despite advances in vulnerability detection tools and techniques. Staffing and budgeting costs could also push organizations to be conservative in their investments in security, particularly in small organizations.

Insufficient Attention to Vulnerability Remediation. Participants mentioned that management might invest in security for reasons other than securing their systems (e.g., fulfilling compliance requirements). Although the byproduct of these intentions might still lead organizations to invest in vulnerability discovery, the downstream effects of finding a vulnerability may not be effective for fixing the root cause of the vulnerability. Compliance regulations often require organizations to establish and follow their own processes to handle vulnerabilities, with compliance often checked by periodic audits (e.g., every 6 months). Since compliance is checked at discrete intervals instead of continuously, organizations that invest in security for compliance purposes may only address discovered vulnerabilities once an audit is scheduled, so that they may develop a corpus of events for the purpose of passing the audit. This is likely to render vulnerability discovery efforts ineffective since discovered vulnerabilities might not be remediated correctly or promptly.

Trust. When working with external testers, one of the most significant challenges faced by organizations and internal testers is trusting external testers with their systems and data. This can discourage organizations from creating public bug bounty programs or working with external red teaming firms. Trust can also be a factor in deciding what components are in-scope for testing engagements, and this might increase the chances of leaving some vulnerabilities undetected. Our results suggest that many organizations tend to put rules in place that limit what red teamers can realistically do to simulate the adversary. Said the director of a pen testing firm:

"..everybody is talking about they want a red team, they want a red team, but at the end of the day, they want to put a bunch of rules around it, just like regular penetration tests" (P38).

For the same reason, many bug bounty hunters expressed that they often feel hesitant to report vulnerabilities they discovered to responsible entities, as they fear reprisal from organizations. This is likely to discourage external security researchers from reporting their discoveries or collaborating with organizations to improve their security posture.

Communication. Lack of communication had caused issues with the reproducibility of reported vulnerabilities when the person who reported a vulnerability does not provide sufficient detail (e.g., the feasibility of exploitation and whether the reports contain realistic exploits). That is, they lack the knowledge required to put the vulnerabilities in the context of the organization to allow assessing the severity and the impact of reported vulnerabilities correctly. This is a clear indication that the industry lacks a proper standard for communicating discovered vulnerabilities to relevant entities. One standard that does exist for assessing severity levels is the Common Vulnerability Scoring System (CVSS). However, participants stated it is not helpful because it lacks sufficient consideration for business requirements or organizational contexts, which can facilitate communication with testers, internal teams, and management as to whether a reported vulnerability is critical or not. Without an effective vulnerability rating mechanism, vulnerabilities with high impact can go unnoticed or issues

that require proper attention might not get it. Another important line of communication is between the security engineers and the management who decide on the budget and staffing. Participants mentioned that they had to put a monetary value on security issues that require immediate attention to garner attention from management. Furthermore, one CISO mentioned that finding a responsible point of contact for each independent system in the event of a breach or incident is especially challenging in large organizations.

Creating a security culture that is driven by a genuine interest in defending technical infrastructure can foster collaboration between developers and testers around resolving security issues. Surprisingly, many of the internal and external testers we interviewed raised the point that they often experience problems concerning receptiveness to their security feedback. One participant mentioned that developers might feel embarrassed or blamed once someone reports a vulnerability in their code. Such attitudes to security are likely to discourage developers from acknowledging the receipt of vulnerability reports and security teams from following up on whether a reported vulnerability has been fixed.

10 Recommendations

In light of our findings, we propose a set of recommendations that we expect to help organizations improve their vulnerability management processes.

First, it is essential to identify the technical and business owners of the various systems an organization has early in the process, so that remediation tasks can be assigned smoothly to their corresponding owners, who can take the lead in deciding how to plan subsequent remediation efforts. We would expect there to be guidelines to help identify which stakeholder to notify to improve the level of transparency between the teams involved and reduce the time from discovery to remediation.

Second, organizations should establish a clear set of risk priorities for upper management that communicates the risks an organization is willing to take, in order to guide decisions concerning vulnerability discovery and remediation. This would also allow security teams to communicate the relevance of discovered security vulnerabilities to the organization's priorities and business goals and therefore facilitate discussions that involve convincing upper management to allocate more resources for vulnerability discovery or remediation. Organizations should also have clear procedures for reporting potential risks to upper management and facilitating decisions that concern resource allocation, coordination among teams and stakeholders, and separation of responsibilities between security, development, and legal teams.

Third, we recommend having clear expectations of the maximum period of time that should elapse from the moment a vulnerability is discovered until it is completely remediated by its technical owners. Such expectations should be based on pre-defined criteria for assessing the severity and potential risks of reported vulnerabilities and assigning a timeline for implementing remediation plans that allow addressing critical vulnerabilities in a timely fashion.

Fourth, we recommend that organizations do not rush to create bounty programs, unless they have solid remediation processes in place. One of the prerequisite to creating a bug bounty program is performing internal security assessments (e.g., pen tests) and remediating all discovered vulnerabilities. A proper assessment of the costs of creating such programs, triaging vulnerability reports, assigning discovered vulnerabilities to their technical and business owners, and carrying out subsequent remediation efforts should also be performed before creating these programs.

Fifth, we recommend improving the current incentive structure implemented in bug bounty programs by providing higher bounties for researchers who include suggestions on how to address reported vulnerabilities or help with testing released fixes. We also recommend involving bounty hunters in vulnerability remediation and encourage internal testers to get in touch with them to explore the available remediation options and get confirmation that a released patch fully remediated the discovered vulnerability. It is also important to define what is in-scope or out-of-scope for bug bounty programs based on a pre-defined threat model, which considers the organization's security objectives and the risks that it aims to mitigate.

Sixth, we recommend customizing the language used in communications between different security teams or other stakeholders to the language that can be clearly understood by the target audience. Communications with upper management should be framed in terms of how a discovered vulnerability might introduce risks to the business to help them make informed decisions. On the other hand, blue teams and development teams might need a mix of technical and business-oriented discussions to guide their decisions on how to fix reported vulnerabilities and how to prioritize addressing them based on their potential impact to the business.

11 Conclusion

We present the results of an investigation of modern vulnerability management processes that focused on exploring the tensions between different security engineering and management roles in small, medium, and large organizations operating in the public and private sectors. We show that while the technical aspects of computer security are imperative for securing organizations' technical infrastructures, these efforts can be hindered by human factors such as a lack of trust, ineffective communication between security teams, and unwillingness to invest in security by upper management.

Acknowledgments

We would like to thank all of our study participants for participating, as well as Amit Elazari and all the anonymous reviewers for their insightful feedback. This research was supported by the Center for Long-Term Cybersecurity at U.C. Berkeley.

References

[1] ISO/IEC 29147: Information Technology – Security Techniques – Vulnerability Disclosure. https://www.iso.org/standard/45170.html, 2019. [Online; accessed: 27-June-2019].

[2] ISO/IEC 30111: Information Technology – Security Techniques – Vulnerability Handling Processes. https://www.iso.org/obp/ui/#iso:std:iso-iec:30111:ed-1:v1:en, 2019. [Online; accessed: 27-June-2019].

[3] Hala Assal and Sonia Chiasson. Security in the software development lifecycle. In *Fourteenth Symposium on Usable Privacy and Security*, pages 281–296, 2018.

[4] Andrew Austin and Laurie Williams. One technique is not enough: A comparison of vulnerability discovery techniques. In *International Symposium on Empirical Software Engineering and Measurement*, pages 97–106. IEEE, 2011.

[5] Kenneth A Bamberger and Deirdre K Mulligan. *Privacy on the ground: driving corporate behavior in the United States and Europe*. MIT Press, 2015.

[6] Jason Bau, Elie Bursztein, Divij Gupta, and John Mitchell. State of the art: Automated black-box web application vulnerability testing. In *IEEE Symposium on Security and Privacy*, pages 332–345, 2010.

[7] Adam Beautement, Ingolf Becker, Simon Parkin, Kat Krol, and Angela Sasse. Productive security: A scalable methodology for analysing employee security behaviours. In *Twelfth Symposium on Usable Privacy and Security*, pages 253–270, 2016.

[8] Ingolf Becker, Simon Parkin, and M Angela Sasse. Finding security champions in blends of organisational culture. *USEC*, 11, 2017.

[9] Ingolf Becker, Simon Parkin, and M Angela Sasse. The rewards and costs of stronger passwords in a university: linking password lifetime to strength. In *27th USENIX Security*, pages 239–253, 2018.

[10] Fredrik Björck. *Discovering information security management*. Department of Computer and Systems Sciences, Stockholm University, 2005.

[11] David Botta, Rodrigo Werlinger, André Gagné, Konstantin Beznosov, Lee Iverson, Sidney Fels, and Brian Fisher. Towards understanding it security professionals and their tools. In *Proceedings of the 3rd symposium on Usable privacy and security*, pages 100–111. ACM, 2007.

[12] Mehran Bozorgi, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker. Beyond heuristics: learning to classify vulnerabilities and predict exploits. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 105–114. ACM, 2010.

[13] Lorenz Breindenbach, Phil Daian, Florian Tramèr, and Ari Juels. Enter the hydra: Towards principled bug bounties and exploit-resistant smart contracts. In *27th USENIX Security Symposium*, pages 1335–1352, 2018.

[14] BugCrowd. https://www.bugcrowd.com/. Accessed: September 16, 2019.

[15] Mariano Ceccato and Riccardo Scandariato. Static analysis and penetration testing from the perspective of maintenance teams. In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, page 25. ACM, 2016.

[16] Kathy Charmaz. *Constructing grounded theory: A practical guide through qualitative analysis*. Sage, 2006.

[17] Akemi Takeoka Chatfield and Christopher G Reddick. Cybersecurity innovation in government: A case study of us pentagon's vulnerability reward program. In *Proceedings of the 18th Annual International Conference on Digital Government Research*, pages 64–73. ACM, 2017.

[18] Anne Edmundson, Brian Holtkamp, Emanuel Rivera, Matthew Finifter, Adrian Mettler, and David Wagner. An empirical study on the effectiveness of security code review. In *International Symposium on Engineering Secure Software and Systems*, pages 197–212. Springer, 2013.

[19] Matthew Finifter, Devdatta Akhawe, and David Wagner. An empirical study of vulnerability rewards programs. In *Presented as part of the 22nd USENIX Security Symposium*, pages 273–288, 2013.

[20] HackerOne. https://www.hackerone.com/. Accessed: September 16, 2019.

[21] Julie M Haney and Celeste Lyn Paul. Toward integrated tactical operations for red/blue cyber defense teams. In *Workshop on Security Information Workers at Symposium on Usable Privacy and Security*, 2018.

[22] Hideaki Hata, Mingyu Guo, and M Ali Babar. Understanding the heterogeneity of contributors in bug bounty programs. In *Proceedings of the 11th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 223–228, 2017.

[23] Michael Howard and Steve Lipner. *The security development lifecycle*, volume 8. Microsoft Press Redmond, 2006.

[24] Steven J Hutchison. Cybersecurity: Defending the new battlefield. Technical report, Defense Acquisition University, Fort Belvoir, VA, 2013.

[25] Aron Laszka, Mingyi Zhao, Akash Malbari, and Jens Grossklags. The rules of engagement for bug bounty programs. In *International Conference on Financial Cryptography and Data Security*, pages 138–159, 2018.

[26] V Benjamin Livshits and Monica S Lam. Finding security vulnerabilities in java applications with static analysis. In *USENIX Security Symposium*, volume 14, 2005.

[27] Thomas Maillart, Mingyi Zhao, Jens Grossklags, and John Chuang. Given enough eyeballs, all bugs are shallow? revisiting eric raymond with bug bounty programs. *Journal of Cybersecurity*, 3(2):81–90, 2017.

[28] Charles P Pfleeger, Shari Lawrence Pfleeger, and Mary Frances Theofanos. A methodology for penetration testing. *Computers & Security*, 8(7):613–620, 1989.

[29] Lena Reinfelder, Robert Landwirth, and Zinaida Benenson. Security managers are not the enemy either. In *Proceedings of the Conference on Human Factors in Computing Systems*, page 433. ACM, 2019.

[30] Eric Rescorla. Is finding security holes a good idea? *IEEE Security & Privacy*, 3(1):14–19, 2005.

[31] Rock Stevens, Josiah Dykstra, Wendy Knox Everette, James Chapman, Garrett Bladow, Alexander Farmer, Kevin Halliday, and Michelle L Mazurek. Compliance cautions: Investigating security issues associated with us digital-security standards. In *Network and Distributed System Security Symposium (NDSS)*, 2020.

[32] Tyler W Thomas, Madiha Tabassum, Bill Chu, and Heather Lipford. Security during application development: an application security expert perspective. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, page 262. ACM, 2018.

[33] Herbert H Thompson. Application penetration testing. *IEEE Security & Privacy*, 3(1):66–69, 2005.

[34] Daniel Votipka, Rock Stevens, Elissa Redmiles, Jeremy Hu, and Michelle Mazurek. Hackers vs. testers: A comparison of software vulnerability discovery processes. In *IEEE Symposium on Security and Privacy (SP)*, pages 374–391. IEEE, 2018.

[35] Rodrigo Werlinger, Kasia Muldner, Kirstie Hawkey, and Konstantin Beznosov. Preparation, detection, and analysis: the diagnostic work of it security incident response. *Information Management & Computer Security*, 18(1):26–42, 2010.

[36] Mingyi Zhao, Jens Grossklags, and Kai Chen. An exploratory study of white hat behaviors in a web vulnerability disclosure program. In *Proceedings of the ACM workshop on security information workers*, pages 51–58. ACM, 2014.

[37] Mingyi Zhao, Jens Grossklags, and Peng Liu. An empirical study of web vulnerability discovery ecosystems. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1105–1117. ACM, 2015.

A Interview Guide

The following subsections include the questions we used in our pre-interview questionnaire and the ones we asked the participants in our semi-structured interviews. Some of the questions were asked in all the interviews, whereas the remaining questions were chosen based on the current role/job title of the participant.

A.1 Pre-Interview Questionnaire

1. How long have you been working as a hacker or a tester? (Choices: Less than 1 year, 1 to 5 years, 6 to 10 years, More than 10 years)
2. How did you learn hacking/vulnerability discovery skills? (select all that apply) (Choices: Reading blog posts, Following hacker write-ups on bug bounty platforms, Following hacker write-ups on GitHub, I am majored in CS, Cybersecurity professional training, Other)
3. What is your job title (employment status)?
4. What other roles have you had in the past (select all that apply)? (Choices: Internal pentester, External pentester, Internal red teamer, External red teamer, Blue teamer, Quality assurance engineer, Purple teamer, Bug bounty hunter, CISO, other)
5. Briefly describe your career path (e.g., bug bounty hunter, pen tester and then red teamer).
6. Do you participate in bug bounty programs? (If yes, how many times have you been awarded a bug bounty? and how many years have you been working as a bug hunter?)

7. Have you received formal security education in computer security or software engineering? if yes, please list the degrees and security certifications you have.
8. In your opinion, what type of education or certifications you think helped you most with the type of software security testing you do?
9. How many people work at your company?
10. How many security professionals does your company have?
11. What is your company's main line of business?

A.2 Interview questions for all groups

1. How would you structure your testing process if you were explaining it to a university student?
2. Why do you think your role is important in testing/vulnerability discovery? (follow up question for managerial positions: How would you decide to use a given testing team and at what point? What sort of factors goes into that decision?
3. Based on your experience, what is the contribution of each team? And how would each of them fit into the testing pipeline?, How do you decide what software gets tested, how it gets tested and when it gets tested?, and How does the time frame affect your approach to testing?)
4. Are there any common mistakes or incorrect practices that you often notice people do when it comes to testing the target software?
5. Can you list a set of metrics that you find useful for quantifying your success/failure with regards to what you found in the software you are testing?
6. How do you define security-related test cases and based on what?
7. What happens after you submit the reports to other teams? (follow up question: In order to make the process more efficient, what are the changes that you want/expect to see after you report a vulnerability? (i.e., what are some ways the flow of the vulnerability reporting would be improved throughout an organization?))
8. How does the testing process work for projects that follow agile methodologies such as Scrum?
9. Briefly describe your recon methodology. (follow up questions: What is the type of information you find useful for the type of testing you do? and What are the main sources of internal and external information you normally rely on to complete your testing work including training and knowledge flows?)
10. What sorts of tools are you using? And what tools do you find useful for testing?
11. Do you prefer static analysis over dynamic analysis? And why?
12. What are the different testing phases that you use tools for?
13. How do you decide what tools to get to your teams? What factors can help you decide on that? Please describe your tool selection process and are there any restrictions in terms of what tools could your engineers use? (this question is for managerial positions such as CISOs)
14. Describe the role of 'automation' in terms of the types of tasks you do.
15. How would you define the scope for software/web application testing?
16. Do you think that scoping is limiting you in terms of what types of testing you can do? If so, why?
17. Are there any other organizational barriers/workplace policies or/and contractual limitations that could restrict your team from carrying out certain types of testing. If yes, please explain.
18. What are the vulnerability types you normally look for?
19. What are the vulnerability types that are commonly and consistently found in different projects?
20. Would you treat zero-days differently? If yes, why? And what is the process you follow for that?
21. Do you interact with other (red, blue, purple, pen testers, etc) teams on your day to day job?
22. How often do you communicate with other teams as part of your job vs how often do you collaborate with another team to get the job done? And, under what circumstances?
23. Have you ever had the chance to discuss the security vulnerabilities you found with software developers? If yes, please comment on the impact of such discussion on improving the security of the software you were testing.
24. Explain how did you acquire your current skill set?
25. How do you stay up-to-date with the latest discovered vulnerabilities and attacks? And how do you reflect this knowledge in your testing work?

A.3 Questions for red teams

1. What are the factors that drive companies to decide to create or contract with a red team?
2. How are red teams structured?
3. What is the frequency in which you do red teaming assessment in a year? What factors affect the number of assessments done in a year?
4. Describe the workflow of red teaming.
5. What is the most common ways you establish an initial foothold on an organization?
6. What are the tactics you employ to remain undetected by blue teams?
7. What are the rules of engagement that you usually make sure are followed?

8. What are the common mistakes that you observed blue teamers or other testing teams do? and what are your recommendations on how to make their job better?
9. Do you talk/teach/explain blue teams about your findings? If so, how regular are these meetings? and, what sorts of information do red teams/blue teams share with each other?
10. The basic idea behind red teaming is to challenge assumptions and identify blind spots, do you have any special standard techniques/checklists? And how do you normally operate? Where do you start?

A.4 Questions for blue teams

1. What are the factors that drive companies to decide to create a blue team?
2. How are blue teams structured?
3. Describe the workflow of blue teaming.
4. What information can help you differentiate between attacks mounted by red teams or real attackers?
5. How is your job different from red teamers' job? Can you describe red teamers' methodology to testing?
6. What are the common mistakes that you observed red teamers or other testing teams do? and what are your recommendations on how to make their job better?
7. Do you communicate with red teams? If so, how regular these meetings are? and, what sorts of information do red teams/blue teams share with each other?
8. If you get an alert, how do you respond? Do you have a process for that? And what bottlenecks do you normally experience throughout the process?

A.5 Questions for purple teams

1. What are the factors that drive companies to decide to create a purple team?
2. What is unique about your purple team from the blue team and the red team? and what is your team's role in the organization?
3. How are purple teams structured?
4. Describe the workflow of purple teaming.
5. Do you expect to have separate blue and red team when there is a purple team?
6. How to improve communication and workflow between red and blue teams?
7. What are the skillsets you share with blue teamers? and what are the skillsets you share with red teamers?
8. Are there differences in the sets of tools used by blue teams and red teams?
9. Do you see areas where improvements can be made in terms of elevating blue and red teamers' chances of finding security vulnerabilities more efficiently and improving the effectiveness of their testing work?

A.6 Questions for penetration testers

1. To what degree do you rely on automated vulnerability detection tools as opposed to manually learn about the systems of interest?
2. In your opinion, how to improve testers' vulnerability discovery capabilities?, how differently companies could use pen testers efficiently than the current practices? and what are the obstacles that could hinder your work?
3. Do you follow a systematic approach to finding vulnerabilities? Please explain.
4. Do you target your recon based on knowledge of an organization's business interests? If so, how?
5. How do you draw your penetration testing plans? is it based on the contract? and does the penetration testing team have the freedom to come up with its own testing plan?
6. How and when do you start to collect information about a target?
7. In your opinion, what are the reasons that drive companies to choose to outsource security testing to an external pen-testing company rather than doing the tests internally by their internal testing teams?
8. What are the types of sensitive details that companies normally disclose to third-party pen-testing services?
9. Can you comment on the nature of non-disclosure agreements that contractors must sign before starting the pen-testing work? and can you comment on whether these agreements are affecting the efficiency or effectiveness of penetration testing?
10. What mechanisms are used to assure accountability of actions taken by pen-testing teams? e.g., do companies keep logs that allow them to inspect what pen testers have done and investigate whether there is some degree of damage that resulted from pen testers actions?
11. Do companies archive information collected during pen testing? And how do they make sure that this information is not accessed by unauthorized people?

A.7 Questions for managerial positions such as CISOs

1. What are the factors that drive companies to decide to create an internal red/blue/purple team or involve an external red team in their pentesting processes?
2. In your opinion, when should an organization consider creating a bug bounty program?
3. Describe the dynamics of interaction that occur between different teams.
4. Do you have an internal red team, pentesting or application testing team? What is the main function of each team? And how do they integrate into the rest of the security ecosystem?
5. In your opinion, how can we maximize the effectiveness of all the testing teams in an organization and are there specific areas where you think that significant improvements can be made?
6. In your opinion, what are the challenges to unifying the efforts of all different teams?
7. Can you explain some lessons learned from previous pentesting works?
8. In your opinion, where does the role of bug bounty programs fit in the whole security testing process?
9. What is the organization of your blue/red teams? (i.e., what other types of teams/subteams do you have in your organization?)
10. What is the workflow that your blue team would follow once an incident of a malicious attack has occurred?
11. Do you recall any bad experiences that relate to red/blue/purple teaming?
12. If red/blue/purple teaming is done correctly, how do you make sure that the vulnerabilities reported are actually fixed? And what procedures do you follow to ensure that your security posture has actually improved as a result of the testing you've done?
13. Are people normally receptive/welcoming to the type of feedback they receive from red teamers? how about other teams? If no, how do you address this problem?
14. If you were to involve a red team, would you do it in-house or outsource the red teaming process to another company? Please explain the reasons behind your answer.
15. How do you perceive each team's contribution as a whole? If you were to structure these different teams in tiers, how would you do it?
16. In your opinion, how can an organization make the feedback loop between the different testing teams faster?
17. How do bug bounty programs integrate with whatever other testing teams an organization has in place (e.g., red/blue/purple and pentesting teams)?
18. Are there redundancies between what bug hunters do and what other testing teams do?
19. Do you think that some testing teams are biased towards discovering certain vulnerability types? If yes, how so?
20. Have you ever noticed that there is misconnect between different teams? If yes, please explain why such misconnect exists; i.e., was it due to miscommunication between the teams, or was it because there is a missing team in between.
21. Do have a bug bounty program? And what motivated you to create a bug bounty program in the first place?

A.8 Questions for bug bounty hunters

1. How do you view your role in software/application vulnerability discovery in general?
2. In your opinion, when should an organization consider creating a bug bounty program?
3. Why do companies care about bug bounty programs?
4. How do bug bounty programs integrate with whatever other testing teams an organization has in place (e.g., red/blue/purple and pentesting teams)?
5. Describe some ways in which VRPs/bug bounty programs can be improved.
6. What would push you away from a bug bounty program?

A.9 Questions for internal security engineers

1. What information do you know that is unique to your team that other teams find valuable?
2. What information do you feel like would be valuable to how well you can do your job that other teams have?
3. How does your team deal with security vulnerabilities that are found in production after the software has went through your testing pipeline?
4. How does your team deal with bugs found by other teams?
5. Where do you feel your team is weak in terms of finding security vulnerabilities?
6. What do you feel like is your team's strength in terms of improving the security of the organization as a whole?
7. How are internal testing teams structured?
8. Are your paired with an agile development team or a waterfall development team?
9. Are there any tools that would help you identify security vulnerabilities during the testing pipeline that you have heard are great or wish to have?

B Codebook

Tables 1, 2, 3, and 4 show the codebook that we developed and used to analyze our interview data.

Table 1: Codebook

Code	Explanation
blue-activities	Activities carried out by blue teamers.
blue-activities-IR	Any incident response work done by blue teams.
blue-background	A blue teamer's background.
blue-definition	Definition for a blue team.
blue-issues	Problems with blue teaming like when they lack a skill or do not know something.
blue-recon	Recon/early phases in blue teaming.
blue-resources	Discussions that relate to resources/information blue teams can get access to.
blue-skill	Skills participants think are needed for blue teaming.
blue-structure	Blue team structure.
blue-tools-own	Why a blue teamer would write own/custom tools.
blue-tools-why	Why blue teamers use tools.
blue-when	When to hire a blue team.
red-activity	Activities carried out by red teamers.
red-background	A red teamer's background.
red-definition	Definition for a red team.
red-issues	problems with red teaming like when they lack a skill or do not know something.
red-process	When a participant describes the testing process red teams follow.
red-recon	recon/early phases in red teaming.
red-structure	Red team structure.
vulnerability-severity-criteria	When a participant describes how they define vul severity levels, e.g., using CVSS or any other standard.
mature-vs-immature-bountyprograms	Discussions that relate to how bounty programs are different (big vs small companies, mature vs immature ones, etc)
red-tools-own	Why a red teamer would write own/custom tools.
red-tools-why	Why red teamers use tools.
red-skill	Skills participants think are needed for red teaming.
red-specializations	Discussions that relate to what red teaming companies specialize in, e.g., social engineering, physical security, etc.
red-scoping	Any considerations that relate to defining what's in-scope or out-of-scope for red teaming engagements.
red-when	When to hire a red team.
blue-red-conflict	Conflicts between red teams and blue teams in general.

Table 2: Codebook (cont.)

Code	Explanation
red-blue-conflict-solution	How to fix the red-blue conflicts.
red-relation-other-teams	How a red team relates to other teams, including whether it is operating internally or externally.
purple-activity	Activities carried out by purple teamers.
purple-background	Purple teamers' backgrounds.
purple-definition	Definition of purple teaming.
purple-issues	Issues experienced by purple teams.
purple-structure	Purple team structure.
purple-when	When to hire a purple team.
purple-skill	Skills participants think are needed for purple teaming.
purple-process	When a participant describes the testing process purple teams follow.
red-vs-purple	Differences between red and purple teams.
bugbounty-benefits	Pros of having a bounty program.
bugbounty-communication	Communications between bounty hunters and companies (e.g., when a company invites a hacker to fix an issue or discusses fixes with them).
bountyhunters-process	Testing process as described by bounty hunters.
bountyhunters-recon	Recon methodology as described by bounty hunters.
bountyhunters-tools	Tools bounty hunters use.
bugbounty-background	A bounty hunter's background.
bugbounty-full-time-job	Discussions that relate to doing bug bounties as a full or part time job.
bugbounty-incentives	Any considerations that relate to incentivizing bounty hunters to work for a program.
bugbounty-issues	Problems with bounty programs and considerations that might deter companies from having a program.
bugbounty-legal	Legal issues associated with bounty programs (e.g., safe harbors, CFAA-related discussions).
bugbounty-private	Any discussions that relate to private (invitation-only programs) or hybrid (public and private) approaches to bug bounties.
bugbounty-rogue	Discussions that relate to any misbehaviors by bounty hunters or why companies trust/not trust bounty hunters.
bugbounty-scoping	Any considerations that relate to defining what's in-scope or out-of-scope for bug bounty programs.
bugbounty-skill	Skills participants think are needed for bug bounties/ or discussions that relate to education.
bugbounty-triaging	Triaging bug bounty reports.

Table 3: Codebook (cont.)

Code	Explanation
bugbounty-trust-issue	Programs trusting bug bounty hunters or vice versa.
bugbounty-when	When to create or (not) create a bug bounty program, and the prep work required for that.
bughunters-program-selection	Considerations hunters make when deciding what program to work for.
pentest-process	When a participant describes the testing process pen testers follow.
pentest-skill	Skills participants think are needed for pen testing.
pentest-tools-own	Why a pen tester would write own/custom tools.
pentest-tools-why	Why pen testers use tools.
pentest-when	When to hire a pen testing firm.
pentesting-definition	Definition of pen testing.
pipeline	When participants describe/justify what teams to have at a high level (e.g., QA, red and blue), or the ordering/sequence of teams to plan to have in the long run.
red-vs-pentesting	When a participant compares red teaming to pen testing.
internal-activities	Testing activities by internal teams.
internal-process	When a participant describes the testing process internal testing teams follow.
internal-team-size	Size (number of engineers) in internal teams.
internal-team-structure	When a participant describes structure of internal teams or specializations required in internal teams.
internal-tools-why	Reasons for automation by internal teams, e.g., why do they write custom scripts?, and considerations that relate to their rationale behind using certain tools.
internal-team-when	When to create an internal testing team (e.g., QA).
internal-vs-external	When a participant justifies the rationale behind hiring an internal vs an external team, or any other discussions that relate to outsourcing.
external-scoping	Discussions that relate to scoping external tests, whether with pen testing or red teaming firms.
external-same	When a participant justifies working with the same external firm or switching from one firm to another every now and then (whether it is a pen testing or a red teaming firm).
external-internal communication	Communications between internal and external testers (external consulting firms).
external-criteria	What considerations are taken when selecting an external firm (whether it is a red teaming or a pen testing firm).
upper-management	What teams expect/wish to have from top management people.
vetting-testing-firms	Discussions that relate to how to select an external red teaming or pen testing firm.
third-party-code-vulnerabilities	Discussions that relate to vulnerabilities found in 3rd party code and how to deal with them.
bugbounty-vs-pentesting	Ddiscussions that relate to comparing bug bounties and pen testing activities.
bypassing-fixes	Discussions that relate to a tester's experience bypassing security fixes.
company-area-of-business	When a participant mentions the line of business of the company/organization he/she is working in.

Table 4: Codebook (cont.)

Code	Explanation
clients-misaligned-expectations	Discussions that relate to areas of conflict or misaligned expectations between pen testers/red teamers and their clients.
compliance-vs-security	Discussions that relate to taking compliance-based approaches to security testing.
learning-resources	Any discussions that relate to resources hackers or testers use to learn security testing.
legal	Legal issues for other teams, not bug bounties.
lifecycle	Discussions that relate to SDLC and how security testing is done as a lifecycle.
NDAs	Discussions that relate to non-disclosure agreements.
organization-maturity	How organization maturity affect what teams an organization has.
process-bias	Discussions that relate to idea like hunters are biased towards web vulnerabilities, pen testers are biased towards network vulnerabilities etc.
remediation/patch-management	Fixing/patching vulnerabilities or handling vulnerability reports.
reporting-issues	Considerations concerning vulnerability reports, such as what to include in such reports.
reproducing-vulnerabilities	Any approaches a participant finds effective for making sure that others can reproduce a vulnerability that has been reported.
security-problem-budget	Issues regarding allocating budgets for security.
security-problem-communication	Communication issues within security teams in an organization.
security-problem-communication-solution	Suggestions on how to solve communication problems between teams.
security-problem-culture	General issues in organizational culture that might hinder security.
security-problems-staffing	Issues regarding staffing in security teams (e.g., not being able to find people with required expertise).
security-positive-experiences	When a participant discusses practices that they found effective for improving information sharing, communication, vulnerability discovery, etc.
security-problem-other	Security problems other than communication, staffing and budget.
security-when-reactive	Organizations' reactive approaches to security.
security-problem-solution	Ways to fix security issues in organizations.
organization-maturity	How organization maturity affect what teams an organization has.
small-companies-security	Approaches small companies use/follow to make security testing cost effective.
soft-skills	All soft and people-oriented skills mentioned in the interviews.
teams-communication	Any info sharing or communications issues participants raise during the interviews.
tools-improvement	Discussions that relate to how to improve existing tools, and what tools CISOs would like to have to improve security processes.

C Participants' Demographics

Tables 5 and 6 include detailed demographic information about our participants and whether they held engineering roles, managerial roles or both. We mark a participant as a decider if he/she held a managerial position and a do-er if he/she held a security engineering position.

Table 5: Participants' Demographics

Participant	Job title	Other roles held in the past	Years of experience doing security testing	Organization size	Decider	Do-er
P1	CISO	Programmer, manager, CISO	Less than 1 year	Large	✓	
P2	Bug bounty hunter and part-time security engineer	External pen tester, internal red teamer, bug bounty hunter	1 to 5 years	Small		✓
P3	Security engineer	Bug bounty hunter, pen tester	1 to 5 years	Medium		✓
P4	Lead infrastructure security engineer	Internal pen tester, internal red teamer, blue teamer, bug bounty hunter	1 to 5 years	Medium		✓
P5	Bug bounty hunter	Bug bounty hunter, pen tester	1 to 5 years	Small		✓
P6	Team lead penetration tester	Internal pen tester, external pen tester, internal red teamer, external red teamer, bug bounty hunter	1 to 5 years	Small		✓
P7	Full time developer and runs a pen testing firm	Internal pen tester, external pen tester, bug bounty hunter, developer, security officer	1 to 5 years	Medium	✓	✓
P8	Full time red teamer	Internal pen tester, external pen tester, internal red teamer, external red teamer, bug bounty hunter, vulnerability researcher	6 to 10 years	Small		✓
P9	Security engineer	Pen tester,blue teamer, quality assurance engineer, purple teamer	6 to 10 years	Large		✓
P10	Security manager	Blue teamer, purple teamer, bug bounty hunter, security manager	1 to 5 years	Large	✓	✓
P11	Security engineer	Quality assurance engineer, bug bounty hunter	1 to 5 years	Small		✓
P12	CISO	CISO	More than 10 years	Small	✓	
P13	Security engineer	Internal pen tester, external pen tester, internal red teamer, external red teamer, blue teamer, bug bounty hunter	More than 10 years	Large		✓
P14	Bug bounty hunter	Bug bounty hunter	1 to 5 years	NA		✓
P15	Pen tester	External pen tester, bug bounty hunter	More than 10 years	Small		✓
P16	Pen tester	Internal pen tester, bug bounty hunter	1 to 5 years	Small		✓
P17	Security engineer	Internal pen tester, external pen tester, internal/external red teamer, blue teamer, CISO	More than 10 years	Large	✓	✓
P18	Bug bounty hunter	External pen tester, internal red teamer, bug bounty hunter	1 to 5 years	NA		✓
P19	Security engineer	Internal pen tester, bug bounty hunter	6 to 10 years	Large		✓
P20	Pen tester	Quality assurance engineer, external pen tester, external red teamer	1 to 5 years	Small		✓
P21	Bug bounty hunter	Bug bounty hunter, external pen tester	1 to 5 years	NA		✓
P22	Security engineer	Security engineer, blue teamer	1 to 5 years	Medium		✓
P23	Security engineer	Internal pen tester, external pen tester, internal red teamer, bug bounty hunter	1 to 5 years	Large		✓
P24	Application security analyst	Security engineer, red teamer, pen tester, bug bounty hunter	6 to 10 years	Small		✓

Table 6: Participants' Demographics (cont.)

Participant	Job title	Other roles held in the past	Years of experience doing security testing	Organization size	Decider	Do-er
P25	Security engineer	Security engineer, bug bounty hunter	Less than 1 year	Large		✓
P26	Red teamer	Internal pen tester, red teamer	6 to 10 years	Large		✓
P27	Bug bounty hunter	Internal pen tester, bug bounty hunter	1 to 5 years	Small		✓
P28	Bug bounty hunter	Bug bounty hunter	1 to 5 years	NA		✓
P29	CTO	CTO, external pen tester	1 to 5 years	Small	✓	✓
P30	Security manager	Internal pen tester, external pen tester, internal red teamer, external red teamer, blue teamer, quality assurance engineer, security manager	More than 10 years	Large	✓	✓
P31	Security engineer	Internal pen tester, internal red teamer, blue teamer, purple teamer	1 to 5 years	Large		✓
P32	Security consultant	Pen tester, red teamer, blue teamer, purple teamer	1 to 5 years	Large		✓
P33	Security engineer	Internal pen tester	More than 10 years	Small		✓
P34	Security manager	Internal pen tester, external pen tester, external red teamer, purple teamer, CISO	More than 10 years	Small	✓	✓
P35	Security engineer	Bug bounty hunter, security engineer	1 to 5 years	Large		✓
P36	Pen tester	Internal pen tester, external pen tester, external red teamer,bug bounty hunter	Less than 1 year	Small		✓
P37	CTO	External pen tester, external red teamer, purple teamer, bug bounty hunter	More than 10 years	Small	✓	✓
P38	Director of pen testing firm	Internal pen tester, external pen tester, bug bounty hunter	More than 10 years	Small	✓	✓
P39	Blue teamer	Blue teamer	More than 10 years	Large		✓
P40	Bug bounty hunter	External pen tester, internal red teamer, external red teamer, bug bounty hunter	More than 10 years	NA		✓
P41	Application security lead	Internal pen tester, external pen tester, internal red teamer, external red teamer, purple teamer, bug bounty hunter	More than 10 years	Large		✓
P42	Technical Cybersecurity Architect	Red teamer, blue teamer, purple teamer	More than 10 years	Large		✓
P43	Security manager	Pen tester, blue teamer	More than 10 years	Large	✓	✓
P44	Security manager	Internal pen tester, blue teamer, quality assurance engineer	More than 10 years	Small	✓	✓
P45	CTO	Internal pen tester, external pen tester, internal red teamer, external red teamer, purple teamer	6 to 10 years	Small	✓	✓
P46	Blue teamer	Internal pen. tester, external pen tester, internal red teamer, blue teamer, bug bounty hunter	More than 10 years	Medium		✓
P47	Pen tester	Internal pen tester, external pen tester, internal red teamer, external red teamer	6 to 10 years	Small		✓
P48	CISO	Internal pen tester, external pen tester, blue teamer, purple teamer, CISO	More than 10 years	Medium	✓	✓
P49	CISO	CISO	Less than 1 year	Large	✓	
P50	CISO	Internal pen tester, External pen tester, Internal red teamer, external red teamer, blue teamer, purple teamer	More than 10 years	Small	✓	✓
P51	Secrity manager	Security engineer, security manager	More than 10 years	Large	✓	✓
P52	Security manager	Security manager, blue teamer	6 to 10 years	Large	✓	✓
P53	Security manager	Security manager	More than 10 years	Small	✓	

User Mental Models of Cryptocurrency Systems - A Grounded Theory Approach

Alexandra Mai
SBA Research

Katharina Pfeffer
SBA Research

Matthias Gusenbauer
Tokyo Institute of Technology, SBA Research

Edgar Weippl
University of Vienna

Katharina Krombholz
CISPA Helmholtz Center for Information Security

Abstract

Frequent reports of monetary loss, fraud, and user-caused security incidents in the context of cryptocurrencies emphasize the need for human-centered research in this domain. We contribute the first qualitative user study ($N = 29$) on user mental models of cryptocurrency systems and the associated threat landscape. Using Grounded Theory, we reveal misconceptions affecting users' security and privacy.

Our results suggest that current cryptocurrency tools (e.g., wallets and exchanges) are not capable of counteracting threats caused by these misconceptions. Hence, users frequently fail to securely manage their private keys or assume to be anonymous when they are not. Based on our findings, we contribute actionable advice, grounded in the mental models of users, to improve the usability and secure usage of cryptocurrency systems.

1 Introduction

More than ten years after the first Bitcoin transaction was performed [13], cryptocurrencies have gained popularity among different types of users, ranging from technology enthusiasts to investors, gamblers, and people who are simply curious. Media reports often contain anecdotes of negative user experiences with security and privacy in cryptocurrency systems. Cryptocurrencies obviate the need for central control by maintaining a decentralized public ledger. While technical aspects of cryptocurrencies have been heavily studied (e.g., [6], [17], [34], [3]), human-centered studies are still rare.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

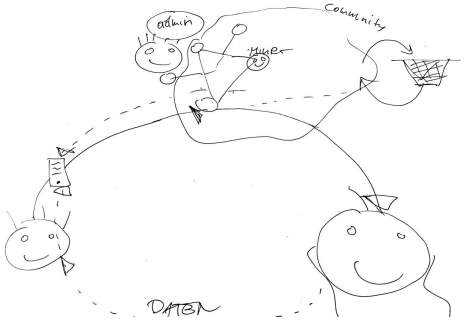


Figure 1: Drawing assignment of the transaction process (S8).

Gao et al. [16] used semi-structured interviews to explore specific aspects of the Bitcoin system out of context. Krombholz et al. [29] quantitatively examined user perceptions on Bitcoin security mainly based on closed-ended questions. However, so far no research investigated mental models which are based on users' *tacit knowledge*. Such knowledge consists of implicit and subjective assumptions that cannot easily be verbalized, but are heavily influencing human behavior [24].

We extend the state of the art by providing the first qualitative user study ($N = 29$) to learn about people's **mental models of cryptocurrencies**[1] **and associated security and privacy threats**. Therefore, we use drawing and card assignment tasks (see Figure 1). Our study methodology follows an inductive approach based on Grounded Theory (GT) [18, 31, 42].

Thereby, we answer the following research questions:

- **Q1** What mental models of cryptocurrency systems and their functional components do users have?
- **Q2** Which mental models interfere with a secure and privacy-preserving usage of cryptocurrency systems?
- **Q3** How can cryptocurrency tools prevent security and privacy threats caused by users' mental models?

[1] We focus on Bitcoin and Ethereum since they were the most prevalent cryptocurrency systems in terms of market capitalization [9] at the time of our study.

Our work aims at explaining (some of) the reasons for user-caused security incidents. This is necessary in order to re-design tools and create effective strategies for behavior change. We argue that cryptocurrency tools (e.g., wallets, online exchanges) should be designed in a way to avoid security or privacy risks even when used by people with incorrect or incomplete mental models. This is in line with Wash et al. [43] claiming that instead of attempting to force users into more 'correct' mental models, technology should be shaped to work well with existing mental models.

Through our study we identify the gaps between the actual protocol functionality and users' mental representations. Although not all the incorrect or incomplete mental models found imply security pitfalls, some partly explain why users of current cryptocurrency tools fail to securely manage their digital assets and have wrong assumptions about privacy and anonymity. Mental models with negative consequences include an erroneous understanding of cryptocurrency systems with regards to (i) cryptographic keys, (ii) anonymity, and (iii) fees.

2 Related Work

Cryptocurrency systems differ from other public key systems (e.g., PGP and secure messaging), as keys are used to sign transactions which are transparently published in the blockchain, instead of ensuring confidentiality through encryption. The threat model is entirely different as well: losing a private key leads to severe problems in the context of cryptocurrency systems as monetary assets are involved. A plethora of research has been carried out to study security and privacy aspects of the Bitcoin system [45]. Nevertheless, several user-centric challenges remain, providing a breeding ground for security and privacy threats.

Only few studies have examined the usability and user perceptions of cryptocurrency systems, mainly focusing on Bitcoin. Baur et al. [4] found that users attribute a high potential to cryptocurrencies, but perceive the usefulness of current cryptocurrencies as low. According to Khairuddin et al. [25], the major motivation for users to buy Bitcoin is its potential for financial revolution, increased user empowerment, and its use as an investment. Sas and Khairuddin [40] as well as Lustig and Nardi [32] explored trust issues of Bitcoin users. Elsden et al. [11] proposed a typology of emerging blockchain applications making it easier for users to understand them. Gao et al. [16] conducted a qualitative study where they found that the major entry barrier for non-users is a perceived necessity for profound technical knowledge.

The first large-scale quantitative user study was presented by Krombholz et al. [29], revealing that many users neither understand nor use the security capabilities of coin management tools correctly. Although the authors also conducted a small number of qualitative interviews, those were only used to contextualize their quantitative findings, not to construct an inductive theory. Kazerani et al. [23] investigated the influence of (poor) usability of cryptocurrency management tools on the adoption of Bitcoin by lay people. Eskandari et al. [12] compared the usability of different cryptographic key management approaches.

However, these earlier studies either opted for a quantitative study design (e.g., [29]) or asked questions which the interviewees deemed too complex to answer given their background as non-users (e.g., [16]). To the best of our knowledge, our work is the first mental model study on cryptocurrencies that aims at discovering the tacit knowledge of the participants. Therewith, we answer open questions on why users commonly fail to manage private keys safely in the context of cryptocurrencies and which parts of current cryptocurrency tool interfaces put users with incorrect mental models at security or privacy risk. We give suggestions for future designs of cryptocurrency tools on how to ensure that user behavior does not compromise the users' security and privacy.

3 Methodology

The overall goal was to understand user perceptions and misconceptions of functional principles, and whether they prevent users from using cryptocurrencies in the most secure and privacy-preserving manner. We chose Bitcoin and Ethereum as examples of prevalent cryptocurrencies and excluded Ethereum's smart contract functionality to only focus on its native currency ether. Furthermore, payment channels are out of scope of our research. This allows us to make general assumptions about user perceptions with regard to the majority of cryptocurrencies that build on the same functional principles as Bitcoin and Ethereum (i.e., in relation to key generation and usage, transaction generation and confirmation, blockchain application, and mining operations). For the remainder of this paper, we will thus use the term *cryptocurrency* to refer to bitcoin, ether, and similar cryptocurrencies.

3.1 Grounded Theory

We follow an inductive approach and use Grounded Theory (GT) [18, 31, 42] to explore user perceptions based on qualitative data. GT is a set of systematic inductive methods to develop theories that are grounded in qualitative research data. A key characteristic is that it merges data collection and analysis in an iterative approach until (theoretical) saturation is reached [42]. Therefore, different phases of recruitment and coding are necessary (see below). By following a process during which we directly analyze the collected data, we generate descriptive theories that are as close to reality as possible. GT is traditionally used in social sciences and has gained popularity in human-computer interaction and usable security research [15, 20, 28].

3.2 Recruitment

Our goal was to recruit a diverse sample of current and potential future cryptocurrency users. We approached possible interviewees through Bitcoin mailing lists and social media as well as personal contacts, also to get in touch with organizations that work with blockchain technology.

We distributed a short description of our study and issued a questionnaire (Appendix A.1) for preselection. To prevent potential participants from reading up on the technical intricacies of blockchain technology, we did not disclose the concrete purpose of our study, only that it deals with cryptocurrencies. Then we selected a subset of participants fitting our recruitment criteria from the people who completed the questionnaire.

We chose the participants according to their self-reported level of knowledge about cryptocurrencies and information technology (ranging from lay users to experts) as well as to their usage of cryptocurrencies. We also chose to recruit participants with diverse exposure to and interaction with cryptocurrency. We recruited 7 people who were not actively using cryptocurrencies but who were working with cryptocurrencies in their professional life (e.g., organizing cryptocurrency meetups, conferences or projects with wallet/exchange operators). Further 10 participants considered cryptocurrencies mainly as an investment, 5 used them mainly for trading, and 7 actively used cryptocurrencies as a payment method.

While the self-reported data might not fully reflect the actual knowledge level of participants, we are confident that these measures are sufficiently accurate to reflect our inherently diverse target population and that a diverse sample was obtained.

3.3 Sampling

GT [42] requires to go back and forth between data collection and analysis in order to construct a theory which is derived from data and not chosen a priori (as it is the case in quantitative studies). Following GT, we conducted the selection of participants in two rounds (two weeks apart). First, we collected an initial sample of 18 cryptocurrency users (experts and non-experts) and then explored the obtained data through open coding.

Based on the concepts derived from our analysis, we extended our initial sample to people who are not actively using cryptocurrencies themselves, but work in institutions that use or deal with cryptocurrencies or blockchain technology (see Section 3.2). Since these people were confronted with cryptocurrency tools, at least at a superficial level, they have certain mental models but are not influenced by cryptocurrency tool interfaces. These mental models are particularly interesting as they represent perceptions of (potential future) first-time cryptocurrency users for whom cryptocurrency tools should be designed as well. By comparing non-users to users, we were able to explore how cryptocurrency tool interfaces might influence mental models (cryptocurrency tool bias) and also investigate biases of non-users (e.g., bank bias). For the second round of recruitment, we collected additional data, recruiting a sample of 11 participants based on the emerging theories.

Table 1: Participant demographics. Total N=29

Demographic	Participants (%)
Gender	
Male	19 (65.5%)
Female	10 (34.5%)
Age	
18 – 22	1 (3.4%)
23 – 27	12 (41.4%)
28 – 32	10 (34.5%)
33 – 37	4 (13.8%)
38 – 42	2 (6.9%)
Highest Completed Education	
High school	5 (17.2%)
Bachelor degree	10 (10.4%)
Master degree	14 (72.4%)

Hence, we had a final set of 29 participants (summarized in Table 1). In order to protect the privacy of our participants, we queried the age, beginning with 18 years, in intervals of five years.

3.4 Design and Procedure

As shown by Kearney and Kaplan [24], people commonly construct implicit knowledge maps to understand complex systems when the systems' functionality goes beyond their technical knowledge. They argue that such tacit knowledge influences people's decision-making and behavior in critical situations, although they are often not aware of it. We opted for a study design that encourages participants to expose their tacit knowledge and functional understanding by engaging them in drawing and card assignment exercises. During these exercises the participants had to assign cards with a function (e.g., "sign transaction") to the entities in their drawings.

Based on related work on human factors of Bitcoin [11, 12, 25, 35] and recent mental model studies in usable security [15, 16, 20, 37, 43, 46], we constructed an interview script for semi-structured interviews including a short pre-assessment questionnaire covering demographics and the participants' general cryptocurrency usage patterns, two drawing tasks, and a card assignment task. The complete study material can be found in the Appendix (Section A).

Our final dataset is based on 29 interviews which were conducted in person in two Austrian cities, namely Vienna and Graz. The interviews took place either in a room at our lab, the participants' workplace, or their home. The majority of the interviews were conducted by two researchers (one interviewer and one assistant taking notes). Two interviews were performed by only one interviewer due to scheduling conflicts.

With the informed consent of the participants, we recorded all interviews, fully transcribed them, and photographed all drawings and card assignments. The pictures (along with the transcribed verbal explanations) served as our baseline for coding.

3.4.1 Pilot Studies

We carried out three pilot studies to test the expressiveness and practicability of our interview script. At the end of each iteration, we requested feedback from the respective participant. We specifically asked the participants to explain their understanding of selected functional concepts and well-known buzzwords – e.g., blockchain, [de-]centralized system, miner – if the participants had not drawn or mentioned them during the interview. We modified the interview script only once (after the first interview). Therefore, we decided to include the remaining two interviews in our final sample.

3.4.2 Interview Procedure

Before the actual interview started, participants were briefed, they signed a consent form and received their compensation (20 Euro Amazon voucher). Each interview lasted roughly 30 minutes and consisted of semi-structured questions, two drawing tasks and one card assignment task. These tasks were based on a concrete scenario, namely transferring a certain amount of bitcoin or ether to a fictional friend called Alice.

In the course of the first drawing task, we asked participants to visualize and verbally express all components and actors involved in the transaction process, as well as their connections. We encouraged participants to think aloud while drawing to gather additional insights into their reasoning. Afterwards, we gave them 15 cards with short descriptions of selected functions (e.g., "generate private key", "generate transaction", "validate transaction", etc.; see Appendix A.2). Depending on which cryptocurrency the participants were more familiar with (self-assessment), the cards reflected the terminologies from either bitcoin or ether. We told the participants to assign the cards to the components and actors in their drawings. We did not provide full definitions but asked the participants to verbalize their own understandings of these technical terms and the associated context.

We added the card assignment task to assist the participants in refining and contextualizing their tacit knowledge. In order to eliminate the possibility of misunderstood terms and random guessing during the interview, the participants were encouraged to provide detailed definitions as far as possible and the interviewers asked follow-up questions if further clarification was needed.

The second drawing task was used to elicit understandings of attackers and threats, and how specific security and privacy risks are contextualized in transaction processes.

3.5 Ethical Considerations

Our organization, which is located in Austria, has no institutional review board but a series of guidelines to be followed when conducting user studies. One of the fundamental requirements is to preserve the participants' privacy and limit the collection of sensitive data as much as possible. Before the interview, all participants were asked to sign consent forms in which the goals of our study and data handling procedures were described. Those consent forms were stored securely and do not contain any links to the IDs we assigned to our participants. The study furthermore strictly followed the EU's General Data Protection Regulation (GDPR).

3.6 Coding

3.6.1 Open and Axial Coding

We followed a GT-based approach to interpret our data. After the first 18 interviews, two researchers independently coded the data (initial *open coding*) with the aim to group recurring statements and assertions relating to the same phenomena (preliminary categories). We created codes based on the drawings and think-aloud protocols. We refrained from assigning codes based on single denotations or terms (e.g., verify, confirm, encrypt). Instead, we coded entire statements and hence included the context in which a term was used.

In line with the full GT approach and while discussing the results, we decided to extend our participant pool by including people who do not actively use Bitcoin, but work in a field related to cryptocurrencies. We are confident that the perceptions and opinions of those people add a new perspective to our study outcome since they have no practical experience in using Bitcoin or Ethereum – and, as such, are not influenced by interfaces of exchange platforms or wallets – but possess some (theoretical) knowledge about blockchain technology. Moreover, this sample's knowledge structures are particularly relevant when thinking about the design of future technology for managing cryptocurrencies, since those people are either potential new users or decision makers in corporate environments.

Following the GT methodology, we performed a second round of open coding to refine the results of the first round. Three researchers independently coded the entire data-set in three rounds (Round 1: 10 interviews, Round 2: 10 interviews, Round 3: 9 interviews). In order to systematize the process, we applied *affinity mapping*, whereby we cut the interview transcripts into snippets and used sticky notes to label newly found categories that emerged from recurring or related statements. As we coded the interviews based on contextualized statements (instead of single terms), we generated a code book based on the participants' mental representations and reasoning. After each round of individual coding we discussed the relations between the newly found categories and agreed upon a set of higher-level categories (*axial coding*). For instance,

we decided to group the categories "public key generation external", "key and address independent", and "one public key for all" to the meta-category "address/key generation". The categories of the third round of coding served as a baseline for *selective coding*.

The final sample size was determined through reaching saturation [19], i.e., no new insights could be gathered from the interviews. As we achieved saturation in the newly emerging categories, we stopped interviewing after 29 participants.

3.6.2 Selective Coding

During this process, three independent researchers agreed upon a set of final core categories centered around the identified misconceptions which might compromise the users' security and/or privacy. The misconceptions are grouped into the following four different top-level categories which consist of multiple subcategories (final codebook see Figure 2[2]):

- ***Meta*** – This category includes statements which were meaningful for building our theory, but are not directly related to cryptocurrency systems and their functionality. It comprises general opinion changes during the interviews, prerequisites, statements about the control or power of the system, biases which influenced participants' descriptions, and misconceptions related to encryption and hashing.
- ***System*** – The system category includes statements describing the blockchain (*Blockchain Description*) as well as where and how it is stored (*Location*). Additionally, this category is split into *Structure*, *Behavior*, and *Function*. *Structure* includes descriptions about the connection between users and miners. The category *Behavior* refers to the behavior of the system (e.g., who receives fees). *Function* on the other hand categorizes the tasks of the keys and the addresses.
- ***Privacy*** – This category codes all mentioned attacks and possible prevention mechanisms on users' privacy.
- ***Security*** – This category includes all attacks and possible prevention mechanisms specific to the users' security.

3.6.3 Final Coding

With a final set of codes grouped into categories, two researchers independently went through all 29 interviews and assigned one or multiple codes, thus generating a comprehensive codebook. Thereby, the transcripts, drawings, and outcome of the card assignment task served as a baseline. We report an inter-rater reliability with a Krippendorff's Alpha value [27] of $\alpha = 0.89$, indicating a high level of agreement among the coders. We claim that this relatively high number is fostered by the technical classification and the granularity

[2]For category B all correct answers are marked with an asterisk *

A Meta	B.4.4 Verification	**C Privacy**
A.1 Bias	B.4.4.1 all peers*	C.1 Attacks
A.1.1 economy	B.4.4.2 blockchain	C.1.1 anonymous
A.1.2 wallet	B.4.4.3 peers: majority/n	C.1.2 identity disclosure @ 3rd party
A.1.3 bank	B.4.4.4 central	C.1.3 address mapping
A.1.4 exchange	B.4.4.5 user	C.1.4 doxxing
A.2 Preassumptions	B.4.5 Transactions: Generation	C.1.5 endpoints
A.3 Opinion Change	B.4.5.1 directly written in blockchain	C.1.6 attacker: state
A.4 Control/Power	B.4.5.2 by user/wallet*	C.1.7 attacker: system participant/s
A.5 Encryption Misconception	B.4.5.3 by blockchain	C.1.8 attacker: external
B System	B.4.5.4 by miners	C.2 Prevention
B.1 Blockchain Description	B.4.6 Transactions: Propagation	C.2.1 self-initiated: mining
B.1.1 system/software	B.4.6.1 client sends to all	C.2.2 self-initiated: info inquiry (3rd party)
B.1.2 algorithmus/actor	B.4.6.2 client sends to part*	C.2.3 self-initiated: address generation
B.1.3 deletable	B.4.6.3 central	C.2.4 self-initiated: identity hiding
B.1.4 datastructure: all transactions*	B.4.6.4 storage pool	C.2.5 self-initiated: anonymized shopping
B.1.5 datastructure: parts	B.4.6.5 direct	C.2.6 user does not care
B.2 Location	B.4.7 Transactions: Confirmation	**D Security**
B.2.1 storage: chunked	B.4.7.1 n blocks*	D.1 Attacks
B.2.2 storage: copied*	B.4.7.2 blockchain	D.1.1 no/secure
B.2.3 central	B.4.7.3 Alice	D.1.2 human failure
B.2.4 internet	B.4.7.4 central	D.1.3 man-in-the-middle
B.2.5 distributed: all*	B.4.7.5 n miners verify	D.1.4 hacking: endpoints
B.2.6 distributed: nodes with shares	B.4.7.6 through fees	D.1.5 hacking: central entity
B.3 Structure	B.4.8 Coin generation	D.1.6 hacking: exchange
B.3.1 User-system connection	B.4.8.1 miner*	D.1.7 DoS
B.3.1.1 automatic *	B.4.8.2 coins equal fees	D.1.8 mining majority
B.3.1.2 cloud	B.4.8.3 exchange	D.1.9 future technology/ theoretical
B.3.1.3 central	B.4.8.4 central	D.1.10 price: volatility
B.3.2 Miner: Connection internal	B.4.8.5 all/system/blockchain	D.1.11 price: manipulation
B.3.2.1 fully connected	B.4.9 Address/Key Generation	D.1.12 attacker: state
B.3.2.2 connected graph*	B.4.9.1 pub key generation: miner	D.1.13 attacker: external
B.3.2.3 pools*	B.4.9.2 pub key generationl: system	D.1.14 attacker: miner
B.3.2.4 not connected	B.4.9.3 one public key for all	D.1.15 technical failure
B.3.2.5 master/server	B.4.9.4 key/address: independent	D.1.16 social engineering
B.3.3 Miner: Connection external	B.4.9.5 send priv key: cloud	D.2 Prevention
B.3.3.1 miner equals blockchain	B.4.9.6 send priv key: between users	D.2.1 system initiated
B.3.3.2 separate system	B.4.9.7 client/wallet generates keys*	D.2.2 self initiated: software
B.3.3.3 user cannot mine	B.4.9.8 key/address: dependent*	D.2.3 self initiated: behavior
B.4 Behavior	B.4.10 PoW/Crypto Puzzle	D.2.4 self initiated: hardware
B.4.1 Fees: recipient	B.4.10.1 encrypted code solving	D.2.5 helplessness
B.4.1.1 cryptocurrency operator	B.4.10.2 find pre-computed value	D.2.5 helplessness
B.4.1.2 exchange	B.4.10.3 compute blockhash*	
B.4.1.3 user	B.5 Function	
B.4.1.4 all	B.5.1 Key	
B.4.1.5 miner*	B.5.1.1 sign: me*	
B.4.1.6 broker and miner	B.5.1.2 sign: Alice	
B.4.1.7 rich participants	B.5.1.3 sign: Alice and me	
B.4.2 Fees: amount	B.5.1.4 sign: Miner	
B.4.2.1 user selected*	B.5.1.5 sign: other system participants	
B.4.2.2 admin selected	B.5.1.6 signing equals transaction verification	
B.4.2.3 miner selected	B.5.1.7 approval	
B.4.2.4 fixed	B.5.1.8 access blockchain	
B.4.3 Block Generation	B.5.1.9 private key equals address	
B.4.3.1 user	B.5.1.10 private key is wallet/account password	
B.4.3.2 blockchain	B.5.2 Address	
B.4.3.3 central	B.5.2.1 nickname	
B.4.3.4 miner*	B.5.2.2 payment destination*	

Figure 2: Final codebook

of the codebook. Conflicts mostly appeared due to slightly different interpretations of the drawings, which sometimes conflicted with the think-aloud protocols. When we detected a conflict, we consulted the drawings and transcripts and discussed the results again. In these cases, we agreed that the verbal explanations should weigh more than the card assignments, since the latter were sometimes less expressive than the participants' verbal descriptions. All conflicts among the coders were resolved.

3.6.4 Theory and Mental Model Construction

The last step of our GT approach was to form theories including the overarching mental models which describe how our participants perceive cryptocurrency systems. First, two independent researchers generated two draft mental models:

one incomplete model and one inaccurate model for the structure, function, and behavior of components in cryptocurrency systems. We constructed the models based on our results, centered around those categories which resulted from our selective coding (codebook). Then, the two coders met in person to reach an agreement. We validated our constructed mental models through negative case analysis [7] by going through all interviews to check whether the participants' statements can be assigned to one of our draft mental models. If not, we sought to understand how they diverged from our draft and adopted it accordingly. In doing so, we iteratively refined our draft mental models until all statements could be assigned. A participants' mental model can contain aspects of the incorrect and incomplete mental model. In order to (i) construct our theory, (ii) examine whether the mental models interfere with secure and privacy-preserving usage of cryptocurrencies, and (iii) understand how resulting issues can be solved, we ran a focus group (Section 4.7) with four experts in the field of cryptocurrencies and blockchain technology. The two final models are presented in Section 4.2 and 4.3.

4 Mental Models of Cryptocurrency Systems

In this section, we first provide a simplified description of the Bitcoin and Ethereum system to provide the appraisal factors for the assessment of our data. Then, we present our participants' veritable mental models. These models represent incomplete and inaccurate descriptions of our participants in correspondence to the structure, functionality, and behavior of cryptocurrency systems. Direct participant quotes (translated to English) are provided for illustration. Since quantitative results (numbers) in qualitative research cannot be used to generalize findings, we will discuss all statements without providing numbers. Nonetheless, coding frequencies can be found in Appendix A.4.

4.1 Appraisal Factors

Before conducting the study, we constructed a ground truth model together with two cryptocurrency experts. These experts were also part of our focus group. We do not claim exhaustiveness of our expert mental models which can be incomplete and diverse as well. To increase the validity, we interviewed two experts and constructed one mental model incorporating both statements. Similar to the user interviews, we asked both experts to draw all components and actors involved in a transaction process and verbalize their thoughts. We then constructed a simplified representation of their mental models (see Figure 3). This model serves as a basis for the evaluation of user mental models and only focuses on important parts for user transactions. We found the participants' mental models to be consistently sparser than the expert mental model. The comparison of users' and expert mental models is purely

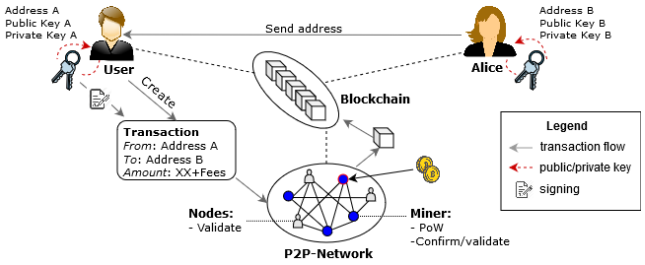


Figure 3: Ground truth of the transaction process of cryptocurrency systems based on blockchains.

illustrative and non-judgmental. We defined the assessment basis as follows:

Bitcoin and Ethereum are blockchain-based, peer-to-peer (P2P) networks which enable users to perform transactions with virtual (crypto-)currencies. The system consists of multiple participants (peers) that we group in four different roles: (i) sender, (ii) receiver, (iii) miner, and (iv) other users. Each participant can hold multiple roles.

The basic requirement to perform a transaction within a cryptocurrency system is that sender and receiver must have some kind of wallet, or are enrolled with an online exchange service. A wallet consists of public keys, private keys, and addresses. The private key is randomly generated and permits the user to spend cryptocurrency units. In the case of offline wallets, it is the user's responsibility to securely store and back up the private key. The address is a hashed version of the public key and acts as a public identifier of the asymmetric key pair.

Prior to performing a transaction, the receiving party communicates its address to the sending party. The sender creates the transaction which comprises the sending and receiving address as well as the transferred amount, including fees. The amount of the fees can be selected by the user and determines the processing speed of the transaction (transactions with higher fees are more likely to be included within the next block). Afterwards, the sender signs the transaction with the private key and broadcasts it to the P2P network. The verification – for both the transactions and the blocks – is performed by peers in the network. Thereby, not all peers necessarily perform full validation (e.g., SPV wallets or thin clients do not check whether transactions are valid, but they rather evaluate whether full nodes have validated them correctly).

A specific transaction t is considered to be confirmed when (i) a miner successfully constructed the Proof-of-Work (PoW) for a block b containing t, (ii) b ends up in the heaviest chain (i.e., the chain with the most cumulative PoWs), and (iii) a certain amount of blocks is succeeding b (as the blockchain gets longer, the confirmation can be considered to be more secure). The miner (or mining pool, i.e., a cluster of miners that work together) who solves the PoW first gets rewarded

with newly created currency (a specific amount depending on the implementation of the system) and the transaction fees. As soon as the transaction is confirmed, the amount is credited from the sender's to the receiver's wallet.

4.2 Incomplete Mental Model

Figure 4 depicts the best-case mental model grounded in our qualitative analysis. It includes technically correct yet sparse perceptions compared to the ground truth (see Figure 3). We did not encounter poor decision-making as a result of incomplete mental models, hence the missing details are not crucial for secure usage of the cryptocurrency system.

Several users correctly stated that cryptocurrency systems are decentralized with annotations reflecting an outline of a peer-to-peer (P2P) system, and a transaction flow matching our ground truth (illustrated through lighter grey continuous lines in Figure 4). The majority correctly stated that the user's wallet software generates the public/private key pair (illustrated by a dotted red line in Figure 4). Some of them also knew that in order to send coins to another party, the sender has to sign the transaction with the generated keys. They correctly mentioned that an address is the payment destination in our proposed scenario. Many participants correctly understood that miners receive the transaction fees. However, only a few participants knew how fees are actually calculated and could give a correct explanation of the mining process.

4.3 Inaccurate Mental Model

The mental model presented in Figure 5 incorporates the participants' misconceptions of cryptocurrency systems. However, not all illustrated components are reflected in all mental models of our participants. Misconceptions related to the transaction flow are illustrated by a grey, continuous line, and those related to the key generation are shown through dashed, red lines in Figure 5. We found that many misconceptions do not jeopardize users' security or privacy. In the following we discuss which misconceptions are crucial and which are not.

Some participants assumed a central management entity as part of a cryptocurrency system, such as a server or broker. Others thought that a direct end-to-end connection existed

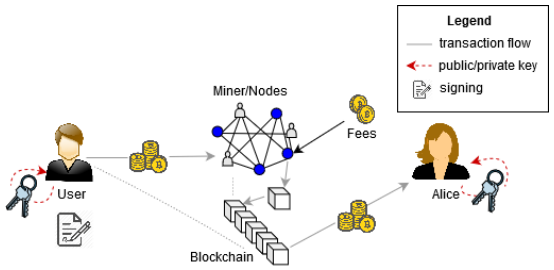


Figure 4: Incomplete mental model

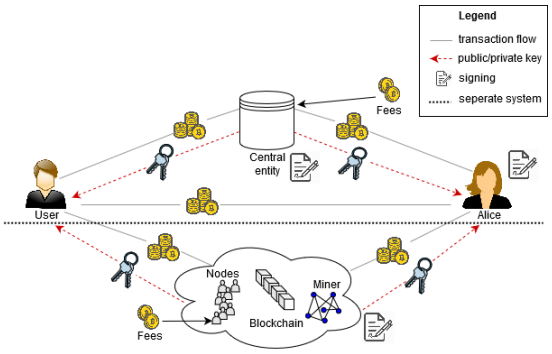


Figure 5: Inaccurate mental model

between sender and receiver, via which transactions are performed.

> *It is a de-centralized system because there is no pivotal element. Only the two accounts interact with each other directly without a third person interfering.* (S6)

One participant hypothesized that in addition to an end-to-end user connection, a further connection to a cloud exists through which users can get initial approval for transactions in order to afterwards send confirmed transactions directly to the receiver. Participants with incorrect mental models often described the blockchain, other nodes, and the miners either only through keywords without being able to explain them, or as a separate system or cloud. Therefore we depict them as a cloud and (partly) separated system (see bottom half in Figure 5). In the following sections, we discus these misconceptions and their impact on security and privacy in detail.

4.3.1 Cryptographic Keys

We identified many misconceptions related to the keys used in cryptocurrency systems. Although users' problems with cryptographic keys (and their management) have already been investigated for other application areas – for example secure messaging and PGP – the effects of mistakes from the users' perspective are different for cryptocurrency systems (e.g., direct monetary impact). In particular, we found that participants do not understand who generates the keys. Some participants claimed that the miners carry out key generation or expected the whole cryptocurrency system to generate keys.

> *Hmmm well, I don't generate my private keys myself, they are saved in my smartphone app. It is... generally the blockchain who generates it [the key] for me, or the network, the blockchain. It is floating in the air somehow. I don't know. It comes from the internet.* (S19)

One participant thought that all parties in the Bitcoin system share one common key. This would break cryptocurrency systems because everybody would have access to everybody else's funds. Other participants presumed that in order to send coins between two parties, the users' private keys have to be sent to "the cloud".

> *I generate my private key and send it to the cloud. Then I get back [from the cloud] a public key... I must be able to rely on the channel to be secure, e.g., encrypted, when I send my private key to the cloud.* (S22)

Through contextual information from this interview, we can deduce that S22 was not referring to storing a key in the cloud (which would be a correct mental model), but to sending it to the cloud in order for the blockchain to get decrypted.

One participant assumed that they have to send the private key directly to the recipient. This would crucially harm the user's security as it would enable the receiver to have access to the sender's account. As most of our participants were not aware of the fact that the private key is generated on their side, they also did not understand that the private key should never be exposed to external entities (such as miners, central entities, or other system participants).

Our participants also lacked understanding of the signing process. Some stated that the receiver has to sign the transaction. Others thought that both, the sender and the receiver have to sign. A few participants inaccurately stated that the miners have to sign transactions. Several participants assumed that other end users in the system are signing transactions in order to validate them. One participant stated that other end users as well as the miners have to sign a transaction. A participant claimed that a user's keys were necessary in order to access the blockchain. As a result, users frequently did not understand why and how they should keep their private key safe, given that they did not understand what a private key can be used for.

We observed many incorrect card assignments and descriptions not matching our ground truth model in relation to cryptocurrency addresses. It was unclear to many participants what a cryptocurrency address actually is. One participant thought that the private key is a user's Bitcoin address. This misconception is especially severe as it might encourage a participant to share the private key with other participants. Many participants assumed that the generated keys and cryptocurrency address are entirely independent. Some participants assumed that the address is a form of nickname, similar to a pseudonym which you choose on a message board.

Such misinterpretations of key generation and usage can have a major security impact if cryptocurrency tools delegate the responsibility of key generation or management to the users without providing guidelines. If, due to misconceptions, users make their keys accessible to others, they become susceptible to theft.

4.3.2 Fees

Our participants expressed many incorrect assumptions about how fees are calculated and what their purpose is. A few participants explicitly stated a lack of knowledge in this regard. One participant thought that fees are defined by an administrator, two said that the miners select the amount. Others stated that the amount of fees is fixed.

> *Miners ask for transaction fees, I don't know if I can choose the amount...If I want to send money and I am in a hurry, for example in the case of smart contracts, then it is possible that the miner knows that I am in a hurry and the miner adds an exorbitant amount of transaction fees [to my transaction].* (S20)

As a result of such misconceptions, users might pay transaction fees that are too high in comparison to the amount that would have been needed to fulfill their requirements, if no guidelines are provided by cryptocurrency tools.

4.3.3 Anonymity Misconceptions

During the coding process, further themes related to anonymity in cryptocurrency systems emerged from our collected data which are not directly related to our generated mental models. A few participants assumed that transactions stored in the blockchain are deleted after some time.

> *After 8 blocks one blockchain is ready and it becomes one instance...Then, the old one is deleted.* (S8)

This entails a wrong assessment of privacy features offered by the blockchain. Participant S8 perceived the blockchain as oblivious and drew a garbage can where old transactions are disposed/recycled (this is only correct within the Lightning network [1], which S8 was not referring to). S8 stated that it is not possible to store a too big amount of data in the blockchain. Many participants incorrectly assumed that the cryptocurrency system applies some form of encryption by default. The participants imagined that either the blockchain, the transaction, or the transaction channel between the end points is encrypted.

> *The transaction itself must be encrypted to ensure a secure connection between server and client.* (S29)

One participant argued that the cryptographic puzzle or hashing is an en-/decryption operation necessary to get access to the money which was sent.

> *Alice receives the cryptographic puzzle, but I don't know what happens if she can't solve it...Because, I mean the bottom line is, I encrypt it [some kind of transaction code] and she receives it.* (S17)

Furthermore, some participants thought about encryption as a major factor used for security purposes. However, those participants commonly also stated that it is necessary to have some kind of additional knowledge in order to pursue this kind of cryptographic task.

> *I guess you can encrypt them [the transactions], however I do not know how.* (S20)

These misconceptions violate participants' privacy as they incorrectly assume that information in the blockchain is unreadable by the public. Moreover, in line with the findings by Gao et al. [16], it might discourage people from using cryptocurrencies when they are under the assumption that only participants with cryptographic knowledge are able to correctly apply privacy or security measures (i.e., encrypting the blockchain).

4.4 Mental Models of Security Threats and Prevention

Most of our participants were able to explain a broad spectrum of (potential) security risks. The majority of our participants mentioned threats related to compromised end points (e.g., mobile phones), which are indeed present as shown in a Kaspersky [22] report. However, this threat is not limited to cryptocurrency applications. Furthermore, our participants named mining majority attacks (i.e., an attacker controlling more than 50% of the mining power in the network). No mining majority attack has yet been performed on Bitcoin or Ethereum, although Bitcoin Gold experienced a 51% attack in May 2018, and a theoretical approach of an Eclipse attack on Ethereum has been described by Yuval et al. [33]. Therefore, there is a possibility that such an attack could happen in a larger cryptocurrency system, especially when ownership and mining are increasingly concentrated on a small group of people [36].

Many participants referred to attacks related to human failure, such as people losing their private keys or failing to store keys in a secure way. This is in line with results from Krombholz et al. [29] and newspaper articles [21] providing evidence that key loss is often caused by the users themselves. Some mentioned the threat of online exchanges being hacked, which has indeed been reported frequently [30]. Others mentioned price fluctuations or intentional price manipulation (e.g., through fake news) as risk factors. Furthermore, some participants correctly stated that Denial-of-Service (DoS) attacks on cryptocurrency systems [8] pose a potential security risk.

In contrast, some participants revealed an incorrect understanding of the threat landscape in cryptocurrency systems and described attacks which are not feasible in a decentralized system. A few participants stated that hacking of central entities, such as the miners, full nodes, or (parts of) the P2P network is feasible. Some described Man-in-the-Middle attacks as a possibility, where an attacker interferes or manipulates the transaction process and possibly alters information (e.g., the recipient's address).

Other participants reported not to be aware of any security risks and to consider cryptocurrency systems to be secure by design. Some of our participants assumed theoretical threats such as broken or weak cryptography that might expose users to a security risk.

Related to *prevention mechanisms against security threats*, more than half of the participants mentioned self-initiated behavior (such as storing private keys securely). Moreover, many referred to the usage of specific hardware (e.g., hardware wallets) and mentioned software (e.g., secure wallets) as a remedy against security breaches. In relation to that, participants described possible prevention mechanisms initiated by the cryptocurrency system, thinking that users cannot influence their execution. Many participants described feeling helpless as they do not think that (technically non-adept) users can actively apply any measures to circumvent such threats.

> *Maybe I can keep a low profile and I shouldn't sit in the tram with the app because of shoulder-surfing... As a non-professional I cannot really do more.* (S22)

4.5 Mental Models of Privacy Threats and Prevention

Some participants assumed that they are anonymous when using cryptocurrencies. However, the majority mentioned address mapping as a possible privacy threat, which is indeed possible [2, 26]. The second biggest privacy threat people mentioned was identity disclosure through third parties, since it is often mandatory to provide identification when purchasing or exchanging cryptocurrencies. Doxxing (writing private data into the blockchain) and a privacy-threatening attack of the end points (e.g., hacking) were also mentioned. Notably, potential future attacks with the help of quantum computers or artificial intelligence were referred to by several participants. Some thought that the state might be a possible attacker or named external persons with bad intentions as relevant attackers. In contrast, others thought that the system participants themselves might carry out attacks on their privacy.

With respect to *prevention mechanisms against arising privacy threats*, participants referred to the possibility to mine themselves in order to prevent identity disclosure when buying cryptocurrencies. A few participants explained that it is possible to buy cryptocurrencies from a specific third party which does not require identity disclosure. One participant assumed that the usage of two-factor authentication would ensure privacy:

> *To secure myself against the threat that IP addresses can be mapped [to bitcoin addresses], I use two-factor authentication.* (S7)

Some explicitly stated not to care about the prevention of privacy threats as they do not consider them important or do not assume that privacy issues exist in decentralized systems.

4.6 Tool Bias

Many cryptocurrency users focused their explanations and drawings of the transaction process on the graphical user interface which they are exposed to when performing transactions, either via a mobile wallet, a PC wallet, or an online exchange. We observed that wallet interfaces shaped the way participants perceived the blockchain location (centralized vs. decentralized), its functionality (persistent, transparent), and the users' role within the cryptocurrency system.

Figure 6 shows a drawing (example 1) which is influenced by the interface displayed to users when carrying out transactions via mobile phone. In particular, we found that our participants were frequently influenced by a feature of the interfaces currently used by many online exchanges and wallets (Figure 6 example 2). Thereby, the current number of confirmations is displayed to show how many blocks are already successfully mined and incorporated in the heaviest chain of the blockchain. After a specific number of succeeding blocks the current transaction is marked as "accepted". However, we can deduce from our study that users commonly misinterpret these confirmations as a specific number of miners or peers who signed, approved, or validated their transaction. Even among experts, a specific fixed number of confirmations is assumed, although the security actually depends on the weight of the longest chain (see Sompolinsky and Zohar [41]).

In contrast to the cryptocurrency tool bias for cryptocurrency users, we discovered a **bank bias** for non-users. They often stated that the blockchain is centrally managed or that transactions are conducted directly between users.

4.7 Expert Focus Group

In order to construct our theory, we discussed the security and privacy impact of our participants' mental models in an expert focus group which consisted of four members from a different research group at our institution who are primarily researching blockchain technology. One researcher led the discussion and two researchers took notes and asked follow-up questions. First, we presented our incorrect model to all participants and provided printouts. Then, we discussed the incorrect model in three rounds based on the categories resulting from the selective coding (keys, fees, anonymity misconceptions). In each round we first presented the identified misconceptions and then asked our participants whether they think that these categories interfere with security and privacy. If the answer was yes, we asked for the experts' opinions on how these security problems could be prevented. Our discussion and improvement suggestions for cryptocurrency tools are based on the outcome of this focus group.

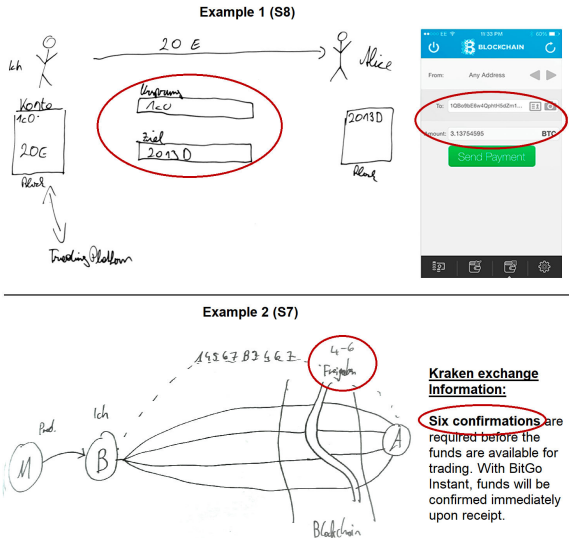


Figure 6: Illustrating cryptocurrency tool bias

We decided on a final set of categories which are important for our theory generation since they have a direct impact on users' security and privacy. These categories are (i) **keys**, (ii) misconceptions regarding **anonymity**, and (iii) **fees**. The resulting mental models are centered around these aspects of our participants' mental representations. The anonymity misconceptions only emerged from the participants' descriptions of the transaction process and were not reflected in their drawings.

Regarding the questions of how cryptocurrency tools could prevent security problems caused by incorrect mental models, the focus group brought up the challenge of designing tools which are adapted to the diverging mental models we found. There is a thin line between an easy-to-use system and a system that gives (expert) users the feeling of being too simple to be secure, and also provides too little information to evaluate the system. Therefore, the focus group proposed that the user interfaces of cryptocurrency tools should have options to switch between different levels of complexity, providing the user with the chance to interact with the system and obtain detailed information about it only if desired. This approach has been (partly) implemented by Coinomi [10] (see Appendix A.3) and should be a standard feature for all (future) wallets.

5 Discussion

Our results explain the roots of several misconceptions with impact on security and privacy found in related work [12, 16, 29] and can be directly linked to concrete improvement suggestions for cryptocurrency tools (e.g., wallets or exchanges).

We claim that modifications of the interface of cryptocurrency management tools can prevent security and privacy threats caused by incorrect mental models. We base this claim

on our observation that there is a **cryptocurrency tool bias** of cryptocurrency users (see Section 4.6). Our results indicate that users' mental models are influenced by the interfaces of tools and technologies they use, which will be subject to further research.

5.1 Challenges and Improvement Suggestions

We found a wide range of mental models, from very detailed to sparse and from correct to incorrect. Hence, we suggest – in line with the outcome from our focus group – to design cryptocurrency tools adapted to diverging mental models and different user groups (e.g., experts and non-experts). Therefore, we suggest that cryptocurrency tool providers ought to offer **different levels of complexity**.

In the following, based on the results from our study and the experts focus group, we discuss how current cryptocurrency tools should be adapted to allow people to use them in a secure and privacy-preserving manner, irrespective of their (incorrect) mental models.

5.1.1 Anonymity

We noticed that about a quarter of our participants used the term "encryption" when describing a transaction process in cryptocurrency systems. Many participants stated that the blockchain is encrypted. We hypothesize that these users mixed up authentication/signing (which indeed takes place during a transaction process) and encryption. Most of these participants assumed that encryption is a safety measure against security- or privacy breaches. Moreover, many participants presumed that transactions cannot be tracked due to the encryption of the blockchain. We claim that such misconceptions jeopardize people's privacy as some participants were incorrectly assuming that their information is hidden from the public or that all information is deleted after some time. Our results suggest that people with these misconceptions refrain from taking measures to safeguard their privacy while believing that they are anonymous.

Furthermore, we revealed misconceptions about the persistence of the blockchain. We infer from discussions with industrial partner institutions that blockchain technologies are commonly applied in areas where it does not make sense, such as for ephemeral data. The mental models we found in the course of this study explain such a contradiction.

Recommendation: *Interfaces of cryptocurrency tools should illustrate the openness, persistence, and transparency of the blockchain. For example, a block explorer could be integrated, visualizing in which block a transaction is integrated and how many succeeding confirmed blocks currently exist. Some wallets (see Appendix A.3) provide access to textual block explorers as an additional feature; however, there is no graphical visualization integrated into the wallets. Furthermore, a pop-up could be shown before pursuing a transaction, stating that this transaction will be broadcasted in clear text to the cryptocurrency network and no information can be altered later on.*

5.1.2 Cryptographic Keys

Previous research on public key cryptography for e-mail encryption has shown that users have difficulties managing and understanding asymmetric keys [38, 44]. Our study supports this finding as less than half of our participants were able to correctly describe how keys are generated and used. Until the time of writing, no holistic solution has been proposed to solve these issues. Bitcoin and Ethereum use keys differently than for example PGP (i.e., it is only used to sign data instead of also encrypting it) and come with an unexplored and diverse user group. Nevertheless, no research has been conducted so far to examine how people understand the function of keys in the context of cryptocurrency systems.

The misconceptions about cryptographic keys, as found during our study, directly influence the way users manage their keys, thus putting them at risk for monetary loss and fraud.We observed that many users did not draw a connection between their private key and the ability to carry out transactions from their account. Moreover, we discovered misconceptions in relation to the key generation. We suppose that these incorrect perceptions interfere with a secure key management if users are not aware of the fact that private keys give access to their funds and should be known only to their owner, hence being kept safe locally.

In line with research on usable key management in other domains [38, 39], we suggest to automate tools as far as possible so that users do not have to deal with key generation or key back-ups, while still providing as much transparency and information as needed to not expose users to security or privacy risks (for a feature overview of key storage and back-up systems from popular wallets, see Appendix A.3).

For cryptocurrency systems this means that users must at least understand that their seed phrase (or private key) (i) should not be shared with anybody else, and (ii) can currently not be recovered in case of loss, leading to the loss of all funds. These facts should be emphasized to the user during wallet initialization and whenever the wallet is used, as discussed in the above sections.

Recommendation: *In order to avoid that users lose their seed phrases, wallets should enforce seed phrase back-ups by asking the users to input a certain number of words from their phrase after making a copy (e.g., writing it down on paper, taking a picture, copying it on a USB device). Furthermore, wallets should ask users to enter their seed phrases in specific time intervals to ensure that they maintain access. Most current wallets do not implement these features (see Appendix A.3). Alternatively, we suggest using automatic key recovery (e.g., similar to trusted friends [14]).*

5.1.3 Fees

Currently, many cryptocurrency tools only offer one fixed amount for fees. Our results show that due to this practice, the majority of participants do not know that users can actively select how much they want to pay as mining fees during the creation of a transaction. Therefore, users are not aware that it is in their power to select how quickly their transaction will be included in the blockchain. As a result, users might pay transaction fees that are too high in comparison to the actual amount needed for their requirements.

Recommendation: *User interfaces of cryptocurrency tools should remind users that by choosing the amount of transaction fees they can influence how quickly their transaction will be included in the next block. The amount of fees should be precomputed based on heuristics (leading to different amounts for each user and transaction) and labeled with understandable terms (e.g., "slow—low fees", "default" and "fast—high fees"). A comparable approach is provided by the Blockchain [5] and Coinomi [10] wallet (see Appendix A.3).*

5.1.4 Security and Privacy Threats and Prevention

We discovered that while our participants showed a basic understanding of the threat landscape in cryptocurrency systems, their knowledge about possible prevention mechanisms was poor and led to a feeling of helplessness among half of them. These participants either believed that users cannot take any measures, but need to rely on the system, or they assumed that prevention mechanisms (e.g., wallet encryption) can only be pursued by technologically knowledgeable users. This coincides with the results found by Krombholz et al. [29] which showed that many users do not apply security measures offered by state-of-the-art cryptocurrency tools.

Recommendation: *We suggest that cryptocurrency tools should perform encryption by default and inform the users about this safety measure (see Appendix A.3 for the status of popular wallets). Moreover, they should add cues and visualizations to explain to the users which security measures (e.g., encryption) are implemented so that users can make informed trust decisions.*

6 Limitations & Future Work

Participant recruitment via mailing lists, social media, and personal contacts provided us with a diverse sample regarding age and profession. However, our sample still has its limitations as it is biased towards a higher educated social stratum; also, non-users without any connection to cryptocurrencies were excluded. Furthermore, the recruiting area was limited to two cities in Austria. Therefore we cannot compare or evaluate cultural differences to other countries/continents, and the European legal landscape with regard to security and privacy (GDPR) also most likely influenced the participants.

The interviews were conducted in German, which is why language-specific expressions in direct participant quotes may have been lost in translation. However, all direct translations were double-checked by a translator, which is why we are confident that such issues have been kept to a minimum.

We followed an inductive approach for our qualitative study to gather insights into user perceptions of cryptocurrency systems. However, our methodology also has its limitations as the data is self-reported and, in comparison to quantitative studies, the sample size is fairly small. Still, we feel confident that our sample is sufficiently large to observe general tendencies.

This study provides the basis for future work to quantify our findings. We plan to examine the connection between mental models, experiences with cryptocurrency management tools, and security-critical errors. Moreover, usable cryptocurrency management tools can be designed and evaluated based on our findings.

7 Conclusions

We explored user perceptions and misconceptions of cryptocurrency users ($N = 29$) enriched with drawing and card assignment tasks. Although our study focused on Bitcoin and Ethereum, our findings can be further useful for improving the security and privacy of a large body of (existing or future) altcoins which also build on the blockchain technology.

We discovered that flaws and inconsistencies in user mental models of cryptocurrency systems expose users to security and privacy risks when using current cryptocurrency tools. These risks include money loss, fraud, or deanonymization. Most importantly, we revealed major misconceptions related to the functionality and management of cryptographic keys which are not compensated by the cryptocurrency tools. Our findings explain why cryptocurrency users fail to manage their private keys securely and, as a result, frequently fall victim to money loss and fraud. Furthermore, we revealed that users think that the blockchain is encrypted or oblivious, which prevents them from taking measures to safeguard their privacy. Another interesting result was that many participants were not aware of the fact that the amount of mining fees can be actively selected to influence the transaction speed.

We proposed several concrete enhancements to state-of-the-art cryptocurrency tools (e.g., wallets or exchanges) with the purpose of protecting users with misconceptions from security and privacy threats. Among others, we suggest to automate key generation, -management, and -back-up as much as possible. With our work, we lay the foundation for improving the usability of state-of-the-art cryptocurrency management tools to prevent security and privacy breaches.

8 Acknowledgments

We thank the anonymous reviewers, our shepherd, and Michael Schilling for their valuable feedback on our work. SBA Research (SBA-K1) is a COMET Centre within the framework of COMET – Competence Centers for Excellent Technologies Programme and funded by BMK, BMDW, and the federal state of Vienna. The COMET Programme is managed by FFG. The financial support by the Christian Doppler Research Association, the Austrian Federal Ministry for Digital and Economic Affairs and the National Foundation for Research, Technology and Development is gratefully acknowledged.

References

[1] Lightning network. https://lightning.network/, 2020. Accessed: 2020-01-31.

[2] Elli Androulaki, Ghassan O. Karame, Marc Roeschlin, Tobias Scherer, and Srdjan Capkun. Evaluating user privacy in bitcoin. *International Conference on Financial Cryptography and Data Security (FC'13)*, 2013.

[3] Maria Apostolaki, Aviv Zohar, and Laurent Vanbever. Hijacking bitcoin: Routing attacks on cryptocurrencies. *IEEE Symposium on Security and Privacy (S&P'17)*, 2017.

[4] Aaron W. Baur, Julian Bühler, Markus Bick, and Charlotte S. Bonorden. Cryptocurrencies as a Disruption? Empirical Findings on User Adoption and Future Potential of Bitcoin and Co. In Marijn Janssen, Matti Mäntymäki, Jan Hidders, Bram Klievink, Winfried Lamersdorf, Bastiaan van Loenen, and Anneke Zuiderwijk, editors, *Open and Big Data Management and Innovation*, pages 63–80, Cham, 2015. Springer International Publishing.

[5] Blockchain. Blockchain.com. https://www.blockchain.com/wallet. Accessed: 2020-05-26.

[6] Joseph Bonneau, Andrew Miller, Jeremy Clark, Arvind Narayanan, Joshua A Kroll, and Edward W Felten. Sok: Research perspectives and challenges for bitcoin and cryptocurrencies. *IEEE Symposium on Security and Privacy (S&P'15)*, 2015.

[7] Anne E Brodsky. Negative case analysis. *The SAGE encyclopedia of qualitative research methods*, page 553, 2008.

[8] David Canellis. Crippling DoS vulnerability put the entire Bitcoin market at risk. https://thenextweb.com/hardfork/2018/09/20/bitcoin-core-vulnerability-blockchain-ddos/, 2018. Accessed: 2020-01-31.

[9] Coinmarketcap. Coinmarket. Cryptocurrency Market Capitalizations. https://coinmarketcap.com/coins/, Accessed: 2019-05-06.

[10] Coinomi. Coinomi. https://www.coinomi.com/en/. Accessed: 2020-05-26.

[11] Chris Elsden, Arthi Manohar, Jo Briggs, Mike Harding, Chris Speed, and John Vines. Making Sense of Blockchain Applications: A Typology for HCI. *SIGCHI Conference on Human Factors in Computing Systems (CHI'18)*, 2018.

[12] Shayan Eskandari, Jeremy Clark, David Barrera, and Elizabeth Stobert. A first look at the usability of bitcoin key management. *arXiv preprint arXiv:1802.04351*, 2018.

[13] Blockchain Explorer. Block #0. https://www.blockchain.com/btc/block-index/14849, 2019. Accessed: 2019-09-17.

[14] Facebook. How can I choose friends to help me log in if I ever get locked out of my account? https://www.facebook.com/help/119897751441086/, 2019. Accessed: 2019-08-27.

[15] Kevin Gallagher, Sameer Patil, and Nasir Memon. New Me: Understanding Expert and Non-Expert Perceptions and Usage of the Tor Anonymity Network. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS'17)*, pages 385–398. USENIX Association, 2017.

[16] Xianyi Gao, Gradeigh D. Clark, and Janne Lindqvist. Of Two Minds, Multiple Addresses, and One Ledger: Characterizing Opinions, Knowledge, and Perceptions of Bitcoin Across Users and Non-Users. *SIGCHI Conference on Human Factors in Computing Systems (CHI'16)*, 2016.

[17] Juan Garay, Aggelos Kiayias, and Nikos Leonardos. The bitcoin backbone protocol: Analysis and applications. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 281–310. Springer, 2015.

[18] Barney G Glaser and Anselm L Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Transaction publishers, 1967.

[19] Greg Guest, Arwen Bunce, and Laura Johnson. How many Interviews are enough? An Experiment with Data Saturation and Variability. *Field methods*, 18(1):59–82, 2006.

[20] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. My Data just goes Everywhere: User Mental Models of the Internet and Implications for privacy and security. In *Symposium on Usable Privacy and Security (SOUPS'15)*, pages 39–52. USENIX Association Berkeley, CA, 2015.

[21] Michael Kaplan. I accidentally threw away $60M worth of Bitcoin . https://nypost.com/2018/05/26/i-accidentally-threw-away-60m-worth-of-bitcoin/, 2019. Accessed: 2019-05-02.

[22] Kaspersky. The number of mobile malware attacks doubles in 2018. https://www.kaspersky.com/about/press-releases/2019_the-number-of-mobile-malware-attacks-doubles-in-2018-as-cybercriminals-sharpen-their-distribution-strategies, Accessed: 2019-05-06.

[23] Ali Kazerani, Domenic Rosati, and Brian Lesser. Determining the Usability of Bitcoin for Beginners Using Change Tip and Coinbase. pages 1–5. ACM Press, 2017.

[24] Anne R. Kearney and Stephen Kaplan. Toward a Methodology for the Measurement of Knowledge Structures of Ordinary People: The Conceptual Content Cognitive Map (3CM). *Environment and Behavior*, 29(5):579–617, 1997.

[25] Irni Eliana Khairuddin, Corina Sas, Sarah Clinch, and Nigel Davies. Exploring Motivations for Bitcoin Technology Usage, 2016.

[26] Philip Koshy, Diana Koshy, and Patrick McDaniel. An analysis of anonymity in bitcoin using p2p network traffic. *International Conference on Financial Cryptography and Data Security (FC'14)*, 2014.

[27] Klaus Krippendorff. Content Analysis: An Introduction to It's Methodology. pages 241–243. SAGE Publications, 2004.

[28] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel von Zezschwitz. " if https were secure, i wouldn't need 2fa"-end user and administrator mental models of https. *IEEE Symposium on Security and Privacy (S&P'19)*, 2019.

[29] Katharina Krombholz, Aljosha Judmayer, Matthias Gusenbauer, and Edgar Weippl. The Other Side of the Coin: User Experiences with Bitcoin Security and Privacy. *International Conference on Financial Cryptography and Data Security (FC'16)*, 2016.

[30] Eric Larcheveque. 2018: A record-breaking year for crypto exchange hacks. https://www.coindesk.com/2018-a-record-breaking-year-for-crypto-exchange-hacks, Accessed: 2019-05-06.

[31] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research Methods in Human-Computer Interaction*. Morgan Kaufmann, 2017.

[32] C. Lustig and B. Nardi. Algorithmic Authority: The Case of Bitcoin. In *2015 48th Hawaii International Conference on System Sciences*, pages 743–752, Jan 2015.

[33] Yuval Marcus, Ethan Heilman, and Sharon Goldberg. Low-resource eclipse attacks on ethereum's peer-to-peer network. *IACR Cryptology ePrint Archive*, 2018(236), 2018.

[34] Rafael Pass, Lior Seeman, and Abhi Shelat. Analysis of the blockchain protocol in asynchronous networks. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 643–673. Springer, 2017.

[35] Mark Perry and Jennifer Ferreira. Moneywork: Practices of Use and Social Interaction around Digital and Analog Money. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(6):41, 2018.

[36] Burgess Powell. Not Only Is A 51% Attack On Blockchain Possible, But It's Coming. https://blocklr.com/news/51-attack-blockchain-more-likely-than-you-think/, 2018. Accessed: 2019-04-24.

[37] Karen Renaud, Melanie Volkamer, and Arne Renkema-Padmos. Why doesn't Jane Protect her Privacy? In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 244–262. Springer, 2014.

[38] Scott Ruoti, Jeff Andersen, Scott Heidbrink, Mark O'Neill, Elham Vaziripour, Justin Wu, Daniel Zappala, and Kent Seamons. "we're on the same page": A usability study of secure email using pairs of novice users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4298–4308, New York, NY, USA, 2016. Association for Computing Machinery.

[39] Scott Ruoti, Jeff Andersen, Tyler Monson, Daniel Zappala, and Kent Seamons. A comparative usability study of key management in secure email. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS'18)*, pages 375–394, 2018.

[40] Corina Sas and Irni Eliana Khairuddin. Design for Trust: An Exploration of the Challenges and Opportunities of Bitcoin Users. 2017.

[41] Yonatan Sompolinsky and Aviv Zohar. Bitcoin's Security Model Revisited. *arXiv preprint arXiv:1605.09193*, 2016.

[42] Anselm Strauss, Juliet Corbin, et al. *Basics of Qualitative Research*, volume 15. Newbury Park, CA: Sage, 1990.

[43] Rick Wash and Emilee Rader. Influencing Mental Models of security: A Research Agenda. In *Proceedings of the 2011 workshop on New security paradigms workshop*, pages 57–66. ACM, 2011.

[44] Alma Whitten and J Doug Tygar. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In *USENIX Security Symposium*, volume 348, 1999.

[45] Jesse Yli-Huumo, Deokyoon Ko, Sujin Choi, Sooyong Park, and Kari Smolander. Where Is Current Research on Blockchain Technology? - A Systematic Review. *PLOS ONE*, 11(10), October 2016.

[46] Eric Zeng, Shrirang Mare, and Franziska Roesner. End User Security & Privacy Concerns with Smart Homes. In *Symposium on Usable Privacy and Security (SOUPS'17)*, 2017.

A Mental Model User Study

A.1 Demographics gathered via a pre-study questionnaire

- Age/ Gender
- Profession/ Highest completed level of education/ Recent professional status
- I have a good understanding of Computers and the Internet: Likert Scale from 5 (agree) - 1 (disagree)
- I often ask other people for help when I am having problems with my computer: Likert Scale from 5 (agree) - 1 (disagree)
- I am often asked for help when other people have problems with their computer. Likert Scale from 5 (agree) - 1 (disagree)
- Which cryptocurrencies have you heard of?
- Was the subject of cryptography and/or cryptocurrencies part of your education or your profession?
- If yes, briefly outline the topics you heard of.
- Do you use Bitcoin/Ethereum?
- For which matters do you mainly use Bitcoin/Ethereum?

A.2 Interview Protocol

General

- Which kind of education do you have and what is your current profession?
- When and how did you become aware of cryptocurrencies?
- How have you been dealing with cryptocurrencies so far?
- Why do you use Bitcoin/Ethereum? (just asked if the participant owns a cryptocurrency)
- What is in your opinion the cryptographic part of cryptocurrencies?

Mental Models

- **[Drawing Task 1]** Please draw a picture of how you think the transaction process works between you and a second person called Alice. Imagine you transfer BTC/ETH 20 to Alice. Remember to include all relevant persons and components into your drawing.
- **[Card Assignment Task]** We prepared some cards which describe various functionalities of a cryptocurrency system. Please assign these cards to the components you drew in Phase 1. If you feel you missed a component before, please draw them with green colour. The cards we provided during this task:
  - Generate address
  - Generate public key
  - Generate private key
  - Transaction confirmed
  - Generate transaction
  - Sign transaction
  - Broadcast transaction
  - Verify transaction
  - Generate block
  - Validate block
  - Perform Proof of Work
  - Solve cryptographic puzzle
  - Receive transaction fees
  - Generate coins
  - Only Bitcoin: Receive unspent transaction output (UTXO)
  - Only Ethereum: Receive balance

Attacker Models

- There are two words which are lately frequently used in the media in relation to cryptocurrencies, namely "security" and "privacy". What do these two words mean to you and what are the differences between them?

- **[Drawing Task 2]** Please have a look on the model you created during Phase 2. Take a red marker for drawing security risks and a blue marker for drawing privacy risks. While drawing, keep the following two questions in mind:

  - Where do you think the potential threats occur?
  - Who is causing those threats?

  After the participant has finished the drawing, ask: "What countermeasures do you know to prevent those risks?"

A.3 Wallet Feature Overview

Table 2: Feature overview of 4 popular software wallets at the time of our study

	Blockchain.com	**Coinbase.com**	**Coinomi**	**Exodus**
Founded	2011	2012	2013	2015
Supported Cryptocurrencies	BTC, ETH, BCH, XLM, USD-D	BTC, ETH, BCH, ETC, LTC, ERC-20 tokens	BTC, ETH, BCH, ETC, LTC, etc.	BTC, ETH, BCH, ETC, LTC, etc.
Type	wallet/exchange	wallet/exchange	wallet/exchange	wallet/exchange
Private key storage	local	local	local	local
Back-ups	user initiated (seed phrase)	user initiated (seed phrase, gdrive with PIN)	user initiated (seed phrase)	user initiated (seed phrase)
Force seed phrase back-up	yes	no	no	no
Transaction Fees	options (pre-calculated/custom)	no options	options (low/normal/high priority)	no options
Wallet encryption	password (forced)	fingerprint/ PIN (forced)	password (standard)/ biometric/ none	none (standard)/ PIN / fingerprint
Periodic seed phrase querying	no	no	no	no
Block explorer included	yes (textual)	yes (textual)	yes (textual)	no
Different complexity levels	no	no	yes (creation: "fast", "advanced")	no

A.4 Coding Frequencies

Table 3: Coding Frequencies

A	Total	K	LA	BA
A.1				
A.1.1	1	0	0	1
A.1.2	6	3	3	0
A.1.3	6	2	2	2
A.1.4	1	1	0	0
A.2	7	3	3	1
A.3	10	3	5	2
A.4	2	0	0	2
A.5	7	1	2	4

B	Total	K	LA	BA
B.1				
B.1.1	10	4	2	4
B.1.2	4	1	2	1
B.1.3	2	0	1	1
B.1.4	14	5	7	2
B.1.5	2	0	2	0
nm	4	1	1	2
LOK	0	0	0	0
B.2				
B.2.1	4	2	2	0
B.2.2	6	3	2	1
B.2.3	3	0	3	0
B.2.4	2	1	0	1
B.2.5	20	8	10	2
B.2.6	1	0	0	1
nm	6	1	1	4
LOK	1	0	0	1
B.3				
B.3.1				
B.3.1.1	6	0	5	1
B.3.1.2	1	1	0	0
B.3.1.3	0	0	0	0
B.3.1.4	5	2	2	1
nm	15	6	5	4
LOK	5	1	2	2
B.3.2				
B.3.2.1	1	0	0	1
B.3.2.2	9	3	5	1
B.3.2.3	7	2	1	4
B.3.2.4	3	1	1	1
B.3.2.5	1	0	0	1
nm	11	5	4	2
LOK	0	0	0	0
B.3.3				
B.3.3.1	2	2	0	0
B.3.3.2	5	2	2	1
B.3.3.3	1	0	0	1
nm	0	0	0	0
LOK	0	0	0	0
B.4				
B.4.1				
B.4.1.1	1	0	1	0
B.4.1.2	1	1	0	0
B.4.1.3	1	0	1	0
B.4.1.4	3	1	1	1
B.4.1.5	19	8	8	3
B.4.1.6	2	1	1	0
B.4.1.7	1	0	0	1
nm	1	0	0	1
LOK	3	0	1	2
B.4.2				
B.4.2.1	4	0	4	0
B.4.2.2	1	0	1	0
B.4.2.3	2	1	0	1
B.4.2.4	2	1	0	1
nm	17	8	6	3
LOK	3	0	1	2
B.4.3				
B.4.3.1	4	0	0	4
B.4.3.2	3	1	1	1
B.4.3.3	1	0	1	0
B.4.3.4	16	7	7	2
nm	3	1	1	1
LOK	3	0	2	1
B.4.4				
B.4.4.1	8	3	4	1
B.4.4.2	4	2	1	1
B.4.4.3	11	3	2	6
B.4.4.4	2	0	2	0
B.4.4.5	4	1	2	1
nm	1	0	1	0
LOK	1	1	0	0

	Total	K	LA	BA
B.4.5				
B.4.5.1	2	1	0	1
B.4.5.2	25	9	9	7
B.4.5.3	1	0	1	0
B.4.5.4	2	0	1	1
nm	1	0	1	0
LOK	0	0	0	0
B.4.6				
B.4.6.1	1	0	0	1
B.4.6.2	6	3	3	0
B.4.6.3	3	1	1	1
B.4.6.4	3	0	2	1
B.4.6.5	2	0	1	1
nm	12	5	3	4
LOK	4	1	3	0
B.4.7				
B.4.7.1	4	1	2	1
B.4.7.2	3	2	1	0
B.4.7.3	3	1	2	0
B.4.7.4	1	0	1	0
B.4.7.5	6	3	1	2
B.4.7.6	0	0	0	
nm	12	3	5	4
LOK	0	0	0	0
B.4.8				
B.4.8.1	23	10	11	2
B.4.8.2	1	1	0	0
B.4.8.3	1	0	1	0
B.4.8.4	3	0	1	2
nm	1	0	0	1
LOK	2	0	0	2
B.4.9				
B.4.9.1	2	1	0	1
B.4.9.2	4	1	1	2
B.4.9.3	1	0	0	1
B.4.9.4	11	2	6	3
B.4.9.5	2	0	0	2
B.4.9.6	1	0	1	0
B.4.9.7	18	8	8	2
B.4.9.8	6	5	1	0
nm	0	0	0	0
LOK	4	0	2	2
B.4.10				
B.4.10.1	2	0	1	1
B.4.10.2	2	0	0	2
B.4.10.3	9	4	5	0
nm	11	4	2	5
LOK	5	1	4	0
B.5				
B.5.1				
B.5.1.1	13	4	6	3
B.5.1.2	1	0	1	0
B.5.1.3	4	1	0	3
B.5.1.4	6	3	2	2
B.5.1.5	6	2	1	3
B.5.1.6	2	0	0	2
B.5.1.7	2	2	0	0
B.5.1.8	1	0	0	1
B.5.1.9	1	0	1	0
B.5.1.10	2	2	0	0
nm	0	0	0	0
LOK	3	0	2	1
B.5.2				
B.5.2.1	2	0	0	2
B.5.2.2	8	4	4	0
nm	16	6	6	4
LOK	3	0	2	1

C	Total	K	LA	BA
C.1				
C.1.1	3	1	2	0
C.1.2	14	2	8	4
C.1.3	18	7	7	4
C.1.4	2	1	1	0
C.1.5	1	0	0	1
C.1.6	5	1	3	1
C.1.7	2	0	1	1
C.1.8	5	2	1	2
C.2				
C.2.1	0	0	0	0
C.2.2	4	1	2	1
C.2.3	5	0	4	1
C.2.4	3	1	1	1
C.2.5	5	1	1	3
C.2.6	3	2	1	0
C.2.7	4	0	3	1
nm	13	5	6	2

D	Total	K	LA	BA
D.1				
D.1.1	2	0	2	0
D.1.2	11	2	6	3
D.1.3	7	0	5	2
D.1.4	15	4	8	3
D.1.5	5	1	3	1
D.1.6	9	3	4	2
D.1.7	4	2	1	1
D.1.8	14	4	6	4
D.1.9	7	2	4	1
D.1.10	2	1	1	0
D.1.11	2	0	0	2
D.1.12	3	2	1	0
D.1.13	10	2	7	1
D.1.14	7	3	2	2
D.1.15	3	1	1	1
D.1.16	4	0	1	3
D.2				
D.2.1	5	1	2	2
D.2.2	7	2	4	1
D.2.3	15	2	7	6
D.2.4	10	3	5	2
D.2.5	13	4	4	5
nm	2	0	2	0

The table displays our resulting numbers of the interviews, categorized according out three participation groups (i) knowledgeable user (K) (ii) lay active user (LA) and (iii) blockchain activity (BA).

Cloudy with a Chance of Misconceptions: Exploring Users' Perceptions and Expectations of Security and Privacy in Cloud Office Suites

Dominik Wermke
Leibniz University Hannover

Nicolas Huaman
Leibniz University Hannover

Christian Stransky
Leibniz University Hannover

Niklas Busch
Leibniz University Hannover

Yasemin Acar
Leibniz University Hannover

Sascha Fahl
Leibniz University Hannover

Abstract

Cloud Office suites such as Google Docs or Microsoft Office 365 are widely used and introduce security and privacy risks to documents and sensitive user information. Users may not know how, where and by whom their documents are accessible and stored, and it is currently unclear how they understand and mitigate risks. We conduct surveys with 200 cloud office users from the U.S. and Germany to investigate their experiences and behaviours with cloud office suites. We explore their security and privacy perceptions and expectations, as well as their intuitions for how cloud office suites should ideally handle security and privacy. We find that our participants seem to be aware of basic general security implications, storage models, and access by others, although some of their threat models seem underdeveloped, often due to lacking technical knowledge. Our participants have strong opinions on how comfortable they are with the access of certain parties, but are somewhat unsure about who actually has access to their documents. Based on our findings, we distill recommendations for different groups associated with cloud office suites, which can help inform future standards, regulations, implementations, and configuration options.

1 Introduction

During the 1970s, office software began to emerge in the world of personal computing. Early word processors such as Electric Pencil for the MITS Altair in 1976, WordStar for the CP/M in 1978, and later dedicated spreadsheet applications such as VisCalc were considered "*killer applications*" for their respective systems. These dedicated office tools helped the adoption of personal computers over more dedicated or mechanical systems for word processing. In recent years, another major shift is happening in the world of office applications. With Microsoft Office 365, Google Drive, and projects like LibreOffice Online, most major office suites have moved to provide some sort of cloud platform that allows for collaboration between multiple editors, automatic real-time storage on cloud or internal network servers, and easy access through the browser without requiring the installation of software.

The major selling points for these cloud office platforms might as well be their most concerning (security & privacy) weaknesses: easy sharing of documents, cloud storage of data, and the high similarity in design and UI to previously prevalent offline office software hide a large array of potential privacy and security trapdoors from the average office user.

With the shift from offline to cloud, many cloud office providers also moved from a pay-once model to a subscription-based model with a trial period or even a model with completely free usage. This shift accompanied a questionable change in business model drive towards data collection and profiling: the processing and storing of documents in the cloud provides the possibility of large-scale privacy intrusion by the providers for both end users and businesses that utilize the cloud. Due to the similarity in design to offline office software, end users are unlikely to fully comprehend this major impact on their privacy. This impact on privacy is further amplified by governments and administrations updating their infrastructure to cloud-based solutions, potentially processing and uploading the data of citizens in the cloud without their explicit consent. In a recent example, the Department of Defense awarded a $7.6 billion contract to General Dynamics to provide the Pentagon with the cloud-based Microsoft Office 365 [27].

Another major selling point of cloud office applications is the ease of access, often from almost anywhere on earth with an internet connection, without requiring any additional installation of software. While the actual location of the underlying servers is rarely mentioned in cloud advertisements, it has

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

notable implications for privacy and security. In July 2019, the central German State of Hesse declared that schools may not legally use Microsoft Office 365 or similar cloud office platforms due to collected telemetry and the potential access to stored data on U.S. servers by U.S. officials [34, 36]:

> "*What is true for Microsoft is also true for the Google and Apple cloud solutions. The cloud solutions of these providers have so far not been transparent and comprehensibly set out. Therefore, it is also true that for schools the privacy-compliant use is currently not possible.*" - Hessian commissioner of Data Protection and Freedom of Information [34].

In August 2019, Microsoft announced that it will be able to provide cloud services from data centers in Germany in late 2019 "*to meet evolving customer needs*" and to being "*committed to making sure that the Microsoft Cloud complies with [the European General Data Protection Regulation] GDPR*" [9]. As of February 2020, Microsoft offers Office 365 and Dynamics 365 from new German data center regions [21].

In this work, we investigate privacy and security misconceptions by end users of cloud office applications in a user study including participants from both the U.S. and Germany. For this, we conducted two online surveys with 200 crowd workers from Amazon's Mechanical Turk and ClickWorker. With a combination of qualitative and quantitative methods, we modeled the two surveys to explore the following research questions:

RQ1: "*How and why do our participants interact with cloud office applications?*" Cloud office suites are compelling to use with features such as collaboration between multiple editors, automatic real-time storage, and easy online access without installation. We are interested why and how our participants interact with cloud office applications both in a home and organizational setting.

RQ2: "*What are end users' awareness, perceptions, and attitudes about privacy in cloud office applications?*" The switch from offline to a cloud environment in both home and organizational settings introduced abrupt changes for privacy and security assumptions regarding office suites. We examine our participants' security and privacy perceptions and expectations, as well as their intuitions for how cloud office suites should ideally handle security and privacy.

RQ3: "*What are participants' understandings and related mental models regarding protection and security of their cloud documents?*" The actual server location, access by providers or governments, and handling of deletions has an enormous impact on the privacy of cloud office applications. We survey the extent of our participants' understanding and their basic mental models regarding cloud office documents.

The remainder of this paper is structured as follows: after this introduction (Section 1) we provide a background to cloud office suites in Section 2. We describe the setup and structure of our two surveys in Section 3 and report our results in Section 4. We discuss related work in Section 5. Finally, we discuss findings and give recommendations in Section 6 and conclude this work in Section 7.

2 Cloud Office Suites

For this work, we define Cloud office suites as cloud-based office applications that allow view, edit and comment on documents, spreadsheets and presentations in the browser.

Table 1 provides an overview of the most popular cloud office suites and their features relevant for this work. Prominent providers of cloud office suites are Google (Google Drive) [12], Microsoft (Office 365) [25], Apple (iWork for iCloud) [15], The Document Foundation (Libre Office Online) [11], and Ascensio System SIA (OnlyOffice) [37]. In contrast to traditional office suites such as Microsoft Office, cloud office suites provide browser-based user interfaces. Users are no longer limited to work on desktop computers using native office applications, but can access their files using any device that provides a modern browser. Hence, modern cloud office suites support mobile devices such as smartphones and tablets and allow easy access to their cloud applications wherever users have access to the internet.

In contrast to traditional office suites, cloud office suites allow users to easily share documents with multiple collaborators and edit the same document simultaneously. Cloud office documents can be shared using e-mail addresses or direct links to a document. For better user experience, cloud office suites allow their users to recover deleted documents. In addition to online access to their documents, Google Drive provides an offline mode that stores documents in the local browser storage and makes them available for offline editing. Offline documents are pushed to the cloud as soon as users have Internet access.

The three major providers Microsoft, Google, and Apple only provide cloud-hosted solutions while Ascensio System also provides a self-hosted community edition which allows keeping the data under users' control. Every hosted cloud office solution provides storage capabilities in the cloud. The amount of storage included depends on the license purchased and can be upgraded at any time. LibreOffice Online by The Document Foundation supports no storage by itself and is dependent on the underlying software like OwnCloud or NextCloud to provide the storage and authentication.

While all cloud office suites provide rudimentary access control for sharing, only Google Drive and OnlyOffice provide an option to share documents with read-only access that still allows to comment on documents.

	Storage	Offline Mode	Versions available				Mobile Version		Sharing				Document Recovery
			Self Hosted	Free	Paid	Trial	Android	iOS	E-Mail	Link	Read Only	Read & Comment	
Office 365	●	○	○	◐[1]	●	●	●	●	●	●	●	○	●
Google Drive	●	●	○	●	●	●	●	●	●	●	●	●	●
iWork for iCloud	●	○	○	●	●	-	○	●	●	●	●	○	●
LibreOffice Online	○	○	●	●	◐[2]	-	◐[3]	○	◐[4]	◐[4]	◐[4]	○	◐[4]
OnlyOffice	●	○	●	●	●	●	●	●	●	●	●	●	●

● Feature available ○ Feature unavailable ◐ Feature partially available
[1] Students and teachers receive a free online only version. [2] Support is only provided by third party companies and not directly by The Document Foundation. [3] Only a viewer is available. [4] Depends on underlying software.

Table 1: Overview of the most common cloud office suites and their related features.

3 Methodology

In this section we provide details on the procedure and structure of the two surveys we conducted with crowd workers from Amazon's Mechanical Turk ($n = 105$) and ClickWorker ($n = 95$). We also detail the coding process for our qualitative questions as well as the statistical analysis approach for our quantitative data. Finally, we report on our data collections and ethical considerations, and discuss the limitations of our work.

Note that while our two surveys may include participants living in the U.S. or in Germany, Austria, or Switzerland respectively, we refer to them as "U.S." and "German(y)" for a more succinct reporting.

3.1 Study Procedure

Both the German-speaking participants from ClickWorker and the English-speaking participants from Mechanical Turk were administered an almost identical survey, with the German survey being a direct translation from the English version by multiple native German speakers.

Questionnaire Development. The questionnaire development was guided by our established research questions. We included pre-tested and evaluated survey questions from previous work where appropriate to allow for a greater comparability between studies. In addition, we performed 5 in-depth, free-form interviews with both experts and non-experts to establish additional areas of interest for our survey.

Pre-Testing. Before we conducted the surveys, we pre-tested our questionnaires following the principle of cognitive interviews [31]. This allowed us to glean insights into how survey respondents might interpret and answer questions. We asked participants to share their thoughts as they answered each survey question and used our findings to iteratively revise and rewrite the survey questions to minimize bias and maximize validity. This first pre-test was conducted internally in both German and English with members of the groups, students of our university, and friends. In addition, we refined the surveys in multiple pilots with participants on Mechanical Turk ($n = 9$) and ClickWorker ($n = 20$) until a satisfactory convergence was reached.

Recruitment and Inclusion Criteria. We recruited participants for our study from Amazon's Mechanical Turk and ClickWorker during September 2019. We did not mention security or privacy in the initial recruitment ad to avoid certain recruitment biases. We generally required participants to be age 18 or older and to have used cloud office software before. For Amazon's Mechanical Turk, we additionally required participants to be comfortable with participating in the study in English and to live within the United States. To ensure sufficient data quality, we also required them to have completed a minimum of 1,000 hits and to have a task approval rate of at least 95% [30]. For ClickWorker, we additionally required participants to be comfortable with participating in the survey in German and to live within Germany, Austria, or Switzerland.

A total of 229 people responded to our surveys. Of those, 22 did not finish and 7 were excluded due to low-quality answers or due to failing at least one of our quality checks, resulting in 200 final participants whose responses we consider.

3.2 Survey Structure

We outline the survey structure in Figure 1 and below. Both our surveys consisted of a total of 9 sections, ranging from general cloud office questions to personal beliefs about the responsibilities of cloud office providers. The two survey versions differed slightly due to localized answer options (e.g., localized names for government agencies) and changes to concepts that do not exist or have a different privacy implication in German-speaking countries (e.g., social security number).

1. Use of Office Tools: Our surveys open with questions in which we explore the general usage patterns of offline and cloud office applications by our participants in both private and organizational contexts. We report general demograph-

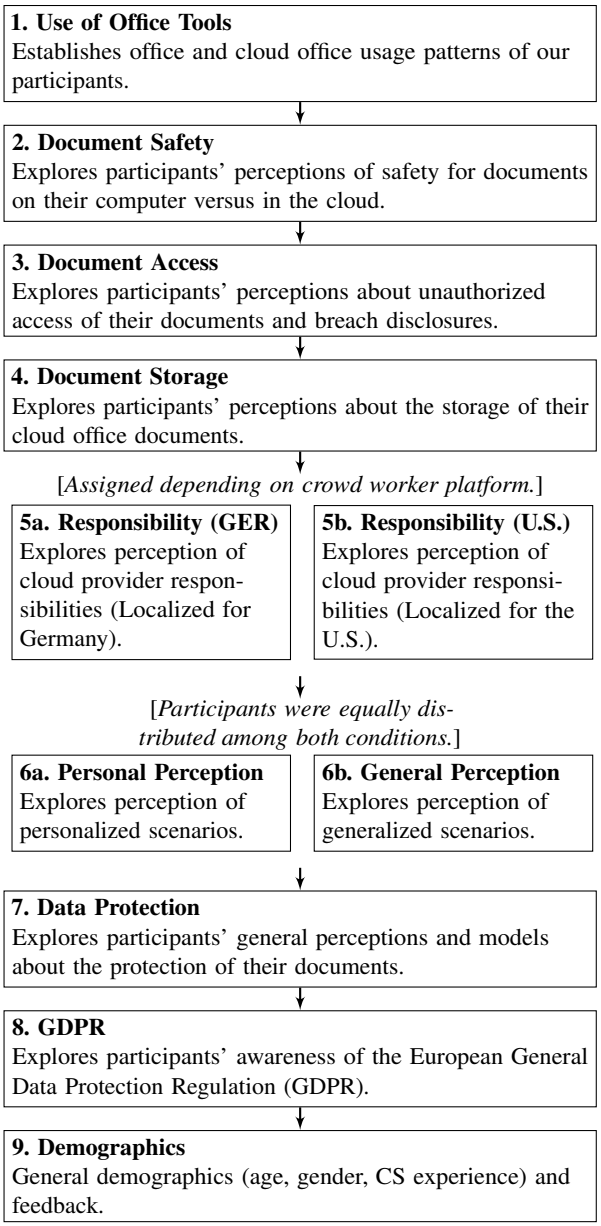


Figure 1: Illustration of the survey flow for both German and English surveys. Splits in the flow include a localized version of the "Responsibility" block for Germany and the U.S. and a split for generalized scenarios vs. personalized which were randomly assigned to participants.

ics and office-specific demographics of our participants in Section 4.1 and Table 3.

2. Document Safety: The "Document Safety" section explores how participants perceive the security of their documents in the cloud vs. locally on their computer and why. We report these results in Section 4.2.

3. Document Access: The "Document Access" section investigates participants' mental concepts and perceived risks related to the access of their documents. Questions related to which parties they think have access to their documents, who already might have accessed their documents without their authorization, and if the risk of unauthorized access by different parties is higher in the cloud or on their computer. Further, the section asks participants about who they think *would* inform them in case of an unauthorized access to their data and who they think *should* inform them and how. We report the results related to the access of cloud office documents in Section 4.3.

4. Document Storage: This section explores our participants' perception about the storage of their cloud office documents. We asked our participants about the number of copies they think exist of their documents and with whom they think copies remain after deleting their own versions. In addition, we asked who they think can delete their documents. We report the results for these questions in Section 4.4.

5a/b. Responsibility: In this section, we investigate our participants' perceptions about responsibilities of cloud office providers regarding access and protection of documents. The "Responsibility" section differs slightly between the German and English survey to allow for the localization of certain answer options such as law-enforcement agencies and government names. We report the results in Section 4.5.

6a/b. Perception: The "Perception" section contains questions to three different scenarios related to the processing of sensitive data in cloud document applications, either in a more personal or more generalized condition.

1. Data of children. The first scenario described the use of a cloud office application in an educational setting. We asked our participants to assess how much they felt at ease with using cloud office applications for handling data of children in schools, e.g., for storing grades or writing tasks.
2. Health data. The second scenario had a focus on health information. A general practitioner used a cloud office application to handle sensitive patient information including a patient's name, age, weight, diagnosis, and treatment plan. Again, we asked our participants to rate their level of comfort with the scenario.
3. Financial data. In the third scenario we illustrated a use case involving financial data. A financial advisor used a cloud office application to process client data. The processed documents include private information such as the client's name, social security number, and detailed financial information.

Participants of the study were equally distributed between both conditions and the order of scenarios was randomized for each participant. Results for the different scenarios are reported in Section 4.6.

7. Data Protection: The "Data Protection" section explores participants' mental models about the protection of their documents in the cloud. We asked our participants which data

they think is collected when they process documents in cloud office applications and how they think their data is protected. We report these results in Section 4.7.

8. General Data Protection Regulation: In the "GDPR" section we explored our participants' general knowledge about the European General Data Protection Regulation (GDPR) and what they know about the protections it offers. These questions link back to the "Responsibility" block, which asked participants about cloud office provider responsibilities directly implied by the GDPR. We report the general results for this block together with other demographics in Section 4.1 and combined it for our analysis of the responsibility section in Section 4.5.

9. Demographics: We administered demographic questions at the end of the questionnaire to prevent stereotype bias [22, 35]. Our demographic questions included age, gender (with free text), and previous experiences in CS education and CS jobs. Additionally we asked respondents for general feedback for the survey questionnaire. We report general demographics and office-specific demographics of our participants in Section 4.1 and Table 3.

3.3 Coding and Analysis

Our collected data includes both qualitative and quantitative data points.

Qualitative Coding. We analyzed all open-ended questions in an iterative open-coding process [7, 38]. Two researchers established an initial codebook [5], coded all open-ended questions together, and resolved emerging coding conflicts immediately in a consensus discussion or by introducing new codes. If new codes were introduced, all previous answers were revisited and re-coded if necessary. Due to the immediate resolving, reporting an intercoder agreement and reliability is uncommon for this approach [20]. The codebook remained stable once both researchers were satisfied that all important themes and concepts in the responses could be captured with the codes. Both surveys were coded with the same codebook and codes for the German survey were assigned by two native speakers.

Quantitative Analysis. We use the non-parametric Kruskal-Wallis H test (KW; non-parametric equivalent to the one-way ANOVA) to compare multiple independent groups. For multiple tests on paired groups, we use the Mann-Whitney U test (MWU) and control the results for multiple testing. We assume an alpha level of $\alpha = .05$ for significance in hypothesis tests. Where appropriate, we controlled our hypothesis tests for the multiple comparison problem with the conservative Bonferroni correction and report the "adjusted"/"adj." values. For certain tests, we map five-point Likert scale answers to numbers (-2, -1, 0, 1, 2).

We present the outcomes of our regressions in tables where each row contains a factor and the corresponding change of the analyzed outcome in relation to the baseline of the given factor. Linear regression models measure change from baseline factors with a coefficient (Coef.) of zero for the value of the outcome. For each factor of a model, we also list a 95% confidence interval (C.I.) and a p-value indicating statistical significance. We highlight p-values below a cut-off of .05 with a star ($*$).

As our regression analyses are intended to be exploratory, we consider a set of candidate models and select the final model based on the lowest Akaike Information Criterion (AIC) [4]. We consider candidate models consisting of the required factors "Country", "Condition", and "Scenario", as well as every possible combination of the optional variables. Required factors, optional factors, and corresponding baseline values are described in Table 2. In cases when we consider results on a per-scenario rather than a per-participant basis, we use a mixed linear model that adds a random intercept to account for multiple scenarios from the same participant.

3.4 Data Collection and Ethics

Our institutions did not require a formal IRB process for the studies conducted in this work. Nonetheless, we modeled our research plan and study procedures after an IRB-approved study, adhered to the strict German and U.S. data and privacy protection laws and the General Data Protection Regulation in the E.U., and structured our study following the ethical principals of the Menlo report for research involving information and communications technologies [10]. All participants approved to a consent form that informed them about the study purpose, the data we collected and stored, and included an e-mail address and phone number to contact the principal investigators in case of questions or concerns.

Recently, researchers faced issues with low data quality on Amazon MTurk [18]. Therefore, we included a number of filters to identify low-quality answers. During data cleaning and analysis, we identified 7 participants who did not pass our quality measures and excluded these invalid participants from further analysis.

We calibrated participants' compensations based on an average piloting time of 10 minutes and payed participants on Amazon's Mechanical Turk $1.70 and on ClickWorker €1.70 for an hourly wage of $10.20 and €10.20, respectively.

3.5 Limitations

As any study with online surveys, our work includes a number of limitations typical for this type of study and should be interpreted in context. In general, self-report studies may suffer from several biases, including over- and under-reporting, sample bias, and social-desirability bias. However, while we utilize self-report data, our central claims are not about the accuracy of respondents' answers to a given question, but

Factor	Description	Baseline
Required		
Country	Germany or U.S., participants assigned based on crowd working platform.	U.S.
Condition	General or Personal. Scenario condition, participants evenly distributed between both conditions.	General
Scenario	Child, Health, or Financial. Type of scenario, all 3 shown in randomized order.	Child
Participant	Random effect accounting for repeated measures (due to the 3 scenarios per participant).	n/a
Optional		
Office at work	True or False, uses office software at work, self-reported.	False
CS Education	True or False, has a CS education, self-reported.	False
CS Job	True or False, has a CS job, self-reported.	False
Age	Age in years, self-reported.	n/a

Table 2: Factors used in candidate regression models. Model candidates always included the required factors and covered all possible combinations of optional factors. Final models were selected based on lowest AIC. Categorical factors were individually compared to the baseline.

rather about the concepts and misconceptions conveyed by their answers.

Conducting user studies on crowd working platforms like Amazon's Mechanical Turk and ClickWorker is a commonly used and generally accepted procedure for human-computer interaction and usable security and privacy research [33]. While the quality of answers can suffer in a crowd worker context, we tried to ensure a high data quality by following best practices by limiting access to our surveys to high-reputation cloud workers [30] and by manually filtering low quality answers.

This study focuses on the responses of German and U.S. Internet users, and thus, we can offer no insight into the generalizability of results for international participants. We aimed to improve the internal validity of our study by providing localized answer options.

We explicitly ignored the implications of meta data collection and third party data of cloud office providers to allow participants to focus on their mental model of cloud document processing and access.

4 Results

In the following section we report and discuss results for all 200 valid participants of both the U.S. and German survey. Generally, participants were aware of certain security and privacy implications of writing their documents in cloud office applications, but were unaware or had severe misconceptions about others. Our reporting of results mostly follows the actual order of survey sections described in Section 3.2. After each subsection, we summarize our key findings.

4.1 Use of Office Tools

We report the general demographics of both surveys in Table 3. Overall, 127 participants responded to our survey on Amazon's Mechanical Turk (U.S.) and 102 on ClickWorker (German). Of those, 105 and 95 respectively completed the survey and were considered valid for a combined total of 200 participants for whom we report results.

Our participants identified predominantly as male (64.5%) with a median age of 33.0 years (mean $= 35.7$, $\sigma = 10.7$). Across both surveys, 28.0% of our participants classified themselves as having a CS education and 22.5% as having worked in a CS-related job. CS experiences are similar for both the U.S. and the German survey, with the exception of CS education (38.1% vs. 16.8%). We assume this discrepancy might be related to general differences in education systems, as the German school curriculum focuses less on IT education compared to the U.S. The majority of both our U.S. and German participants have a job that involves using office applications regularly with 80.0% and 77.8%, respectively.

The majority (97.1%) of our U.S. participants have used Google Drive (with its related cloud office tools such as Google Docs or Google Sheets) before, followed by Microsoft Office (Offline) (86.7%) and Microsoft Office 365 (Cloud) (70.5%). The majority of our German participants (87.4%) is more familiar with Microsoft Office (Offline), followed by Google Drive (80.0%) and Microsoft Office 365 (Cloud) (64.2%). We assume this difference is likely due to the extensive, almost exclusive usage of Microsoft Office products in German businesses and government[1]. These differences even out for office tools used in the last months where Google Drive prevails among both the U.S. and German participants (82.7%, 70.5%), followed by Microsoft Office (Offline) (50.5%, 65.3%) and Microsoft Office 365 (Cloud) (50.5%, 55.8%). The majority of our U.S. participants use office tools to process Spreadsheets (89.5%), Text (76.2%), and Emails (68.6%). As document types, the German participants process Text (90.5%), followed by Spreadsheets (82.1%) and Presentations (65.3%).

[1] E.g. the city of Munich decided to migrate to Windows 10 after it's 2003 decision to adopt Linux, partially due to incompatibility and communication problems with other organizations [13].

	U.S.	German	Combined
Participants			
Started	127	102	229
Finished	110	97	207
Valid ($n =$)	105	95	200
Gender			
Male	66.7%	62.1%	64.5%
Female	33.3%	37.9%	35.5%
Other (Free text)	0.0%	0.0%	0.0%
Age in years			
Mean	35.3	36.1	35.7
Std. dev. (σ)	9.9	11.5	10.7
Median	33.0	33.0	33.0
Computer Science			
CS Education	38.1%	16.8%	28.0%
CS Job	24.8%	20.0%	22.5%
Professional Usage			
Office software at work	80.0%	77.9%	79.0%
Office Usage*			
Google Drive	97.1%	80.0%	89.0%
Microsoft Office (Offline)	86.7%	87.4%	87.0%
Microsoft Office 365 (Cloud)	70.5%	64.2%	67.5%
LibreOffice Offline	18.1%	25.3%	21.5%
Apple's iWork Web (Offline)	9.5%	20.0%	14.5%
Apple's iWork Web (Cloud)	6.7%	17.9%	12.0%
LibreOffice Online	4.8%	9.5%	7.0%
Other	3.8%	5.3%	3.0%
OnlyOffice	1.0%	1.1%	2.5%
Document Usage*			
Spreadsheets	89.5%	82.1%	86.0%
Text	76.2%	90.5%	83.0%
Emails	68.6%	55.8%	62.5%
Presentations	49.5%	65.3%	57.0%
Calendar and Appointments	57.1%	50.5%	54.0%
Other	1.0%	2.1%	1.5%
Document Storage*			
Locally on my computer	73.3%	82.1%	77.5%
Google Drive	88.6%	52.6%	71.5%
Dropbox	33.3%	35.8%	34.5%
OneDrive	30.5%	29.5%	30.0%
iCloud	18.1%	24.2%	21.0%
Network Share	21.0%	16.8%	19.0%

* Multiple answers allowed, may not sum to 100%

Table 3: Demographics for all valid participants from the U.S. survey (Amazon's Mechanical Turk), German survey (ClickWorker), and combined.

Most of our U.S. participants prefer to store their documents in Google Drive (88.6%), followed by locally (73.3%), and Dropbox (33.3%). While the majority of German participants prefers local storage (82.1%), followed by Google Drive (52.6%), and Dropbox (35.8%). This mirrors the distribution of most used office tools for U.S. participants (Google Drive Office $\rightarrow$ Google Drive Storage) and German participants (Offline Microsoft Office $\rightarrow$ Local Storage).

Participants of both the U.S. and German survey agree on the top reasons why they (would) use cloud office applications over local office applications: easy remote access of documents (76.2%, 70.5%), ease of collaboration (58.1%, 59.0%), and free or cheap access (52.4%, 43.2%).

> Summary: Demographics. Somewhat unsurprisingly, participants prefer to store their documents on the platform they edit them with (e.g., locally for offline office). All of our participants agree on the benefits of cloud office applications: free access and easy collaboration for remote documents.

4.2 Document Security

In this question section we asked our participants to think about where they believe their documents are more secure from any unauthorized access, on their personal computer or in the cloud. Most participants reported that they feel their personal computer is more secure for storing their documents than the cloud (54.5% vs. 19.5%).

In addition to the quantitative questions, we asked our participants to explain their assessment. Most (94) of the participants who said they felt their documents would be more secure against unauthorized access on their personal computers mentioned that an attacker would require physical access to their machines to acquire access to documents, e.g., P30 said "*You would have to physically breach my computer to get to the documents, the drive is encrypted, no one can access it.*". Similarly P47 explained

> "*I think that documents are more secure on my computer because I'm the only one that can access them; and if there were any threats on my PC, I would use programs to get rid of them.*" - P47.

Some participants (21) who said files were more secure on their computer thought that it was easier to attack a cloud system than their personal computer, e.g., P27 mentioned that "*Because I know what security I have on my pc, but don't know about Google. Of course, I assume they've got top of the line security, but I don't actually know.*" (P27).

Participants who thought documents in the cloud were more secure (39; 19.5%) mostly mentioned two reasons. First, they believe that cloud office suite providers have more security expertise than they personally do. For example, P79 said

> "*The cloud is managed by big corporations. They probably take security more serious than individuals. They always have to worry about hackers so there [sic] security is likely very powerful.*" - P79.

Second, some participants assessed cloud office suites to be more "secure" than their personal computers because they have backups and losing data is less likely, e.g.,

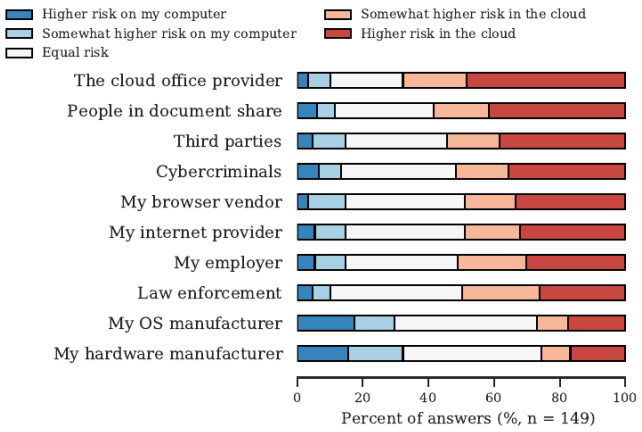


Figure 2: Likert scale for participants' associated risk of unauthorized access between their local computer and their cloud office documents for different parties.

> "*Local computers can be hacked and can crash. It happens. Too often, backups are not made regularly, so data can be lost in either case. With automatic backup to the cloud, documents are more secure in case of local computer issues.*" - P74.

Other reasons for believing in a secure cloud often seem to be based on insufficient technical knowledge, e.g., "*because I think it is not possible to hack the cloud.*" (P219). Few participants (3; 1.5%) mentioned the use of two-factor authentication and the application of encryption by cloud office suite providers as important security factors.

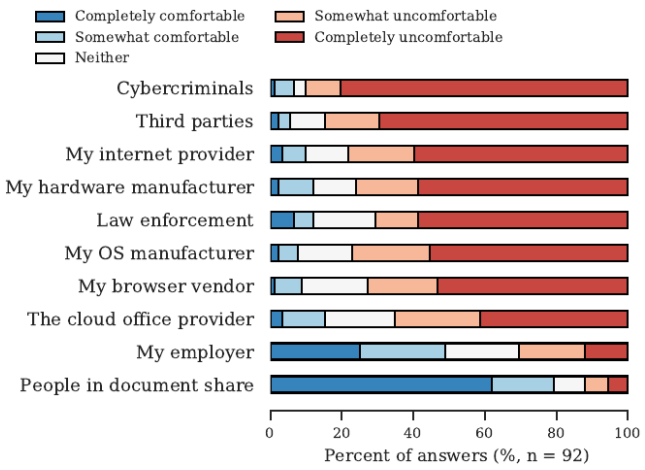
Summary: Document Security. Our participants seem to be aware of some general security implications of processing their documents in the cloud. They seem to prefer their local system in terms of security against unauthorized access, although some of their threat models appear to be underdeveloped.

4.3 Document Access

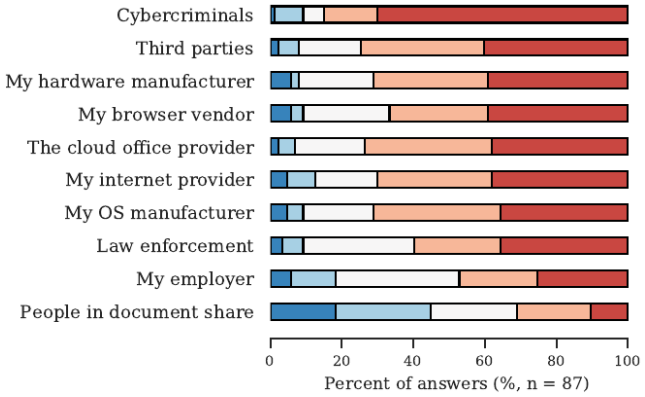
In this question section we explore our participants' perception, misconceptions, and mental models regarding the (unauthorized) access of specific parties to their potentially sensitive documents processed in cloud office applications.

We found a significant difference in the risk of different parties accessing the participants documents (KWH; $H = 102.33$; $p < 0.01$). This might indicate that participants seem to be aware of the changed attack surface for cloud office documents and associate a higher risk of unauthorized access by cybercriminals and third parties such as advertisers and plugin developers in the cloud (cf. Figure 2).

These answers coincide with parties of which participants thought that they already accessed their documents, although some participants have the misconception that their browser



(a) Survey participants from MTurk (U.S.).



(b) Survey participants from CrowdWorker (German).

Figure 3: How comfortable our participants are with different parties accessing their cloud documents. "*I don't know*" answers were omitted.

vendor and operating system provider also have accessed their cloud documents. Figure 3 shows the comfort level of our participants related to the access of different parties to their cloud office documents.

We also asked participants who *would* inform them if their cloud office documents are accessed by an unauthorized party and who *should* inform them. Participants' answers point at a responsible party here: While the German and U.S. participants are split on the cloud office provider (73; 69.5%) and nobody (39; 41.1%) as most common answer on who *would* inform them respectively, both groups agree that it is the cloud office provider that *should* inform them (153; 76.5%).

A large number of participants explicitly told us that they like to be informed about unauthorized access of their cloud office documents by email (119; 59.5%). In addition, some participants provided us with their wishes about the information they want to receive in case of such a data breach, e.g., P69 insisted that

"*[I] [n]eed to know basically everything that the person saw. When they saw it, what they saw, where they're from. I don't care who gives the analysis, just that its an accurate analysis and that they let me know.*" - P69.

Summary: Document Access. Overall, our participants seem to have a clear idea on by whom and how they should be informed about unauthorized access of their cloud documents: the cloud office provider via (secure) email. Our participants seem to have strong opinions on how comfortable they are with the access of certain parties, but are somewhat unsure about who actually has access to their documents.

4.4 Document Storage

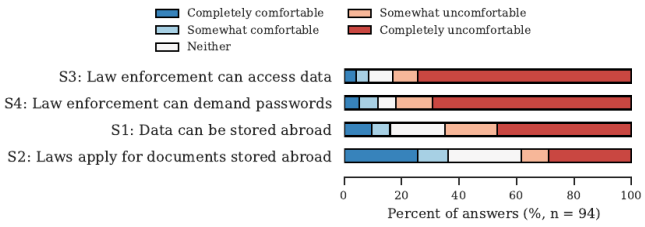
The majority of German participants believe that multiple copies of their cloud office documents exist (49; 51.6%), while most U.S. participants admit that they do not know (51; 48.6%). Of those that assume multiple copies exist (83; 41.5%), the majority thinks that only their copies are deleted if they delete a document (30 of 83; 36.1%), or they are unsure (21 of 83; 25.3%). Unsurprisingly the majority of our participants assume that their cloud provider can delete their documents (138; 69.0%), followed by people they shared the documents with for U.S. participants (43 of 105; 41.0%) and cybercriminals for German participants (46 of 95; 48.4%).

Some of our participants assume a rather basic mental model of why copies of their cloud documents might exist, e.g., P123 believes "*[. . .] that these copies exist just in case that [sic] the original documents get lost.*" (P123). Other participants had a less utilitarian view on the existence of potential copies, e.g., P79 had some rather dystopian thoughts about why copies of their documents are created: "*[T]o use against me when the time is right.*" (P79). For why not all of the copies are deleted, some participants had some very convincing arguments: "*[They are] used to train artificial intelligence or to make a profile of me for the future.*" (P79), "*possibly to sell to 3rd-party vendors for advertising*" (P96), and "*so they can be used for law enforcement.*" (P97).

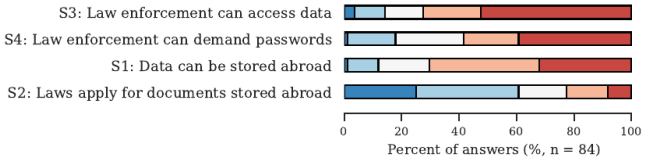
Summary: Document Storage. Overall, our participants seem to be rather unsure about the actual number of copies, access rights, and deletion procedures of their cloud documents. They appear pessimistic regarding the reasons of why additional copies are kept.

4.5 Document Responsibility

In this section we asked participants about which party they think is responsible for the protection of their documents. The majority of U.S. participants sees the cloud provider as responsible (83; 79.0%), while the majority of the German participants sees themselves as responsible (69; 72.6%),



(a) Survey participants from MTurk (U.S.).



(b) Survey participants from CrowdWorker (German).

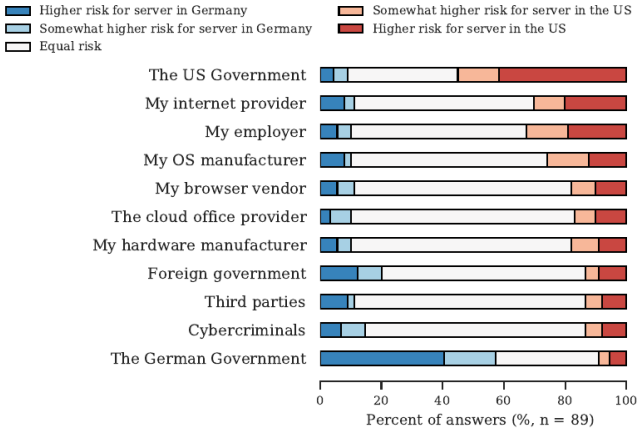
Figure 4: Participants' comfort with potential privacy violations by their government.

We also compared U.S. and German participants in their agreement regarding four scenarios exploring the responsibilities of cloud office providers:

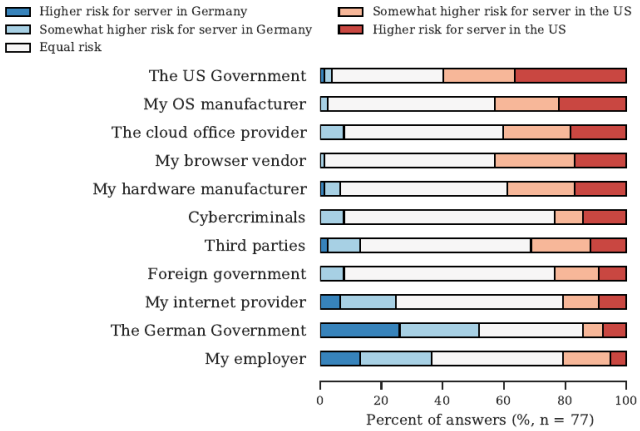
S1: "*Cloud office providers should offer adequate protection for cloud office documents.*"
(MWU; $U = 4445$; adj. p-value $= 1$)
S2: "*I should have the right to demand a full overview of my data collected by cloud office providers.*"
(MWU; $U = 4419$; adj. p-value $= 1$)
S3: "*Upon my request, cloud office providers should have to show what they do with my documents and who has or had access.*"
(MWU; $U = 4181$; adj. p-value $= 1$)
S4: "*Cloud office providers must be able to modify or delete any data they have on private individuals.*"
(MWU; $U = 4566$; adj. p-value $= 1$)

and found no significant differences between our U.S. and German participants. Similarly, we compared U.S. and German participants regarding their (dis)comfort with the following statements (Note that the statements were localized, e.g., an U.S. participant would be presented with "US regulation"):

S1: "*Cloud providers can store my documents on servers outside of the US/Germany without legal repercussions.*"
(MWU; $U = 4151$; adj. p-value $= 1$)
S2: "*US/German regulations and laws still apply if the documents are stored on servers outside of the US/Germany.*"
(MWU; $U = 4817$; adj. p-value $= 0.04$)
S3: "*US/German law enforcement can access my cloud documents without a court order.*"
(MWU; $U = 4768$; adj. p-value $= 0.02$)



(a) Survey participants from MTurk (U.S.).



(b) Survey participants from CrowdWorker (German).

Figure 5: Risk of unauthorized parties accessing participants' documents on servers in the U.S. vs. Germany.

S4: "*US/German law enforcement can force me to give up my cloud office password.*"
(MWU; $U = 5104$; adj. p-value < 0.001)

and found significant differences for S2, S3, and S4. These differences can be mostly attributed to U.S. participants being more uncomfortable with privacy violations by their government compared to the Germans (cf. Figure 4).

We further investigated differences between U.S. and German participants by asking them where they do think the risk is higher of different parties obtaining unauthorized access to their documents if they are either stored on a server in the U.S. or Germany (cf. Figure 5).

> Summary: Document Responsibility. While participants from the U.S. and Germany agree on the responsibilities of cloud providers, U.S. participants are comparably more uncomfortable regarding potential privacy violations by the government.

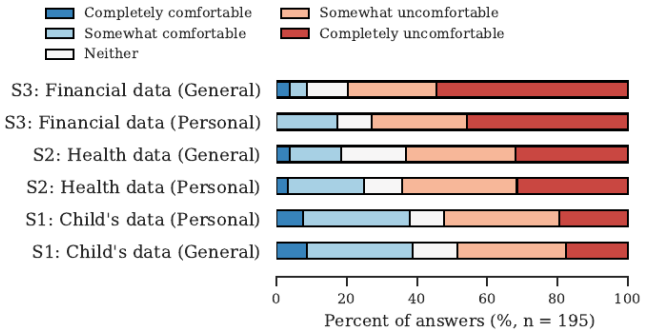


Figure 6: Participants' comfort with three different data scenarios (Financial, Health, and Children) and two different conditions (General and Personal perspective).

Factor	Coef.	C.I.	*p*-value
Scenario: Health	−0.50	[−0.66, −0.33]	$< 0.001*$
Scenario: Financial	−0.88	[−1.05, −0.72]	$< 0.001*$
Condition: Personal	0.03	[−0.25, 0.31]	0.843
Country: Germany	−0.11	[−0.39, 0.17]	0.431

Table 4: Final linear mixed regression model examining the perception of 3 different scenarios in 2 phrasing conditions. "I don't know" answers were omitted. See Section 3.3 and Table 2 for further details.

4.6 Scenario Perception

In this section, we wanted to explore the effect of different conditions and scenarios on how comfortable our participants are with processing documents in the cloud. For this, our participants were presented with three different types of private data stored in cloud documents: children data including names and grades, health data including names and diagnosis, and financial data including names and SSNs. As additional modifier, participants were equally distributed across two conditions: "General" with a more generalized phrasing and "Personal" with a more personalized phrasing (e.g., "a child" vs. "your child").

We explored participants' answers by selecting the best performing model from multiple linear regressions (cf. Table 4). We find that neither the country nor the condition has a significant coefficient in the regression. Both the "Health data" scenario and the "Financial data" scenario are significantly rated as less comfortable by our participants than the "Child data" baseline (cf. Figure 6).

> Summary: Scenario Perception. Our participants are uncomfortable the most with the scenario of processing financial documents in the cloud. Presenting a more personalized scenario nor their country did not significantly affect their comfort level.

4.7 Data Protection

In this section, we asked two free text questions to assess the amount of data our participants think cloud office suite providers collect when processing documents. Additionally, we asked our participants what security measures they think cloud providers deploy to protect their documents. Regarding data collection, most participants thought that cloud office suite providers collected the actual document content and metadata including the time and duration they used the cloud office application, IP addresses and filenames. A few participants were concerned that cloud providers would search their documents for keywords and report them to security agencies and law enforcement, e.g. P96 thinks that providers are "*searching for specific keywords, most notably for US security reasons*" (P96).

Most participants had very specific ideas of what security measures cloud office suite providers would deploy. The majority of our participants were convinced that providers would deploy encryption to protect their stored documents. For example, P74 believes that "*the cloud servers are supposed to be encrypted and follow industry-standard protocols [. . .]*" (P74). Similarly, participants mentioned access control and authentication, e.g., P79 hopes that "*Security is handled by the service provider of the cloud office applications. They probably use complicated passwords and 2 factor authentication.*" (P79). Finally, some participants mentioned firewalls and other network security measures. P111 hopes that "*they are protected by multiple firewalls [and] they are continuously monitored*" (P111).

> Summary: Data Protection. While our participants are aware that the content of their documents might be collected, only few were concerned that specific keywords might be reported to law enforcement. Our participants identified encryption as their preferred security measure their cloud office suite should employ.

5 Related Work

As we conduct surveys investigating end-user security and privacy perceptions, as well as expectations with cloud office suites, we discuss related work in the areas of security & privacy in the cloud and user studies within a context of cloud applications or cloud storage.

Security & Privacy in the Cloud. Past research in the cloud often investigated the privacy of data and sharing, a field still not fully solved judging by the overall unclear or pessimistic views of our participants. The backup and restore performance, liabilities, and problems with data privacy of four cloud storage providers was examined by Hu et al. in 2010 [14]. Also in 2010, Svantesson and Clarke reviewed the terms of use of Google Docs finding that cloud computing is associated with risks to privacy and consumer rights [39]. Johnson proposed in 2017 that the cloud providers should make changes to their terms of service to allow the users better control over their privacy [17], a proposal supported by our work. Similarly, Nestori et al. found in their 2018 paper, that Office 365 is not GDPR compliant [29]. MUBox introduced a meta-cloud storage application to help improve user collaboration on cloud storage services by introducing activity views and Nebeling et al. conducted a user study with 16 participants to examine accuracy and confidence with the activity views [28]. Massey et al. conducted a qualitative study with 27 participants and identified four different strategies that teams used in shared repositories and suggested ways to improve existing tools with new technologies [24].

A number of works concerns client-side encryption or hiding layers to prevent third parties including the cloud office provider from accessing the content of any document edited in the cloud [1, 8, 41], further underlining the need for native encryption, as identified by our participants.

User Studies of the Cloud. Often user surveys in the cloud context focus on the storage aspect: Tan et al. investigated the acceptance of SaaS collaboration tools like Google Docs in an organizational setting and found that their intention to continue using these tools is positively affected by the perceived usefulness and satisfaction [40], which corresponds to our findings regarding ease of use and sharing. Marshall et al. conducted a survey with 106 participants and 19 interviews to understand early user experiences and models of cloud storage systems, finding that users' misconceptions limit the ability to take full advantage of cloud features [23]. Burda et al. developed a technology acceptance model which incorporates users' perception of risk and trust and verified it in a study with 229 cloud storage users. They found evidence that trust in cloud archiving can be increased by a providers' reputation and user satisfaction [3].

Both Clark et al. and Khan et al. explored users' perception of file sharing status over time, finding a mismatch in user expectations and reality [6, 19]. Ramokapane et al. conducted a user study, finding that users struggle to delete their data from the cloud, as incomplete or inaccurate mental models based on a lack of information on deletion lead to a failure to remove the data properly [32]. Mijuskovic et al. conducted a qualitative user study with 28 participants and found that most users are aware of security and privacy risks in the cloud, but lack knowledge to describe potential risks in detail [26]. These previous studies agree with our findings of incomplete mental models, often due to lacking technical knowledge.

We consider the following works by Ion et al. and Arpaci et al. closest to our surveys. Ion et al. studied privacy attitudes and beliefs towards consumer cloud storage by conducting interviews and a survey with end-users in Switzerland and India, finding that requirements for consumer cloud storage differ from those of companies and that end-users prefer local offline storage for sensitive data [16]. Arpaci et al. conducted a study with 200 pre-service teachers to understand the effects

of security and privacy concerns of cloud computing in educational use and proposed a research model that indicates that security and privacy perception has a significant influence on students' attitudes towards cloud services [2].

Compared to these earlier studies consisting mostly of small-scale qualitative studies investigating the acceptance of cloud technology or larger studies focusing on cloud storage, our larger-scale study ($n = 200$) with both qualitative and quantitative parts investigates security and privacy explicitly in the context of cloud office suites.

6 Discussion

In this work, we explored the security and privacy perceptions and expectations of cloud office users, as well as their intuitions for how cloud office suites should ideally handle security and privacy. We performed two online surveys with 200 cloud office users from the U.S. and Germany to explore the following research questions:

RQ1: "*How and why do our participants interact with cloud office applications?*" The fairly recent shift from offline-only tools to cloud office suites includes immense changes of privacy and security implications, although the application design and end user experience remained mostly the same or even included new features. We find that a large majority of our participants regularly work on different document types in cloud office applications. The most common reason for using cloud office applications are the ease of sharing and the ease of use without requiring installation of additional software.

RQ2: "*What are end users' awareness, perceptions, and attitudes about privacy in cloud office applications?*" Users seem to be aware of some general security implications, storage models, and access by others, although some of their threat models seem underdeveloped (e.g., "*I think it is not possible to hack the cloud.*"), likely due to lacking technical knowledge.

RQ3: "*What are participants' mental models regarding protection and security of their cloud documents?*" We find that users' mental models regarding access and sharing are incomplete and their understanding of cloud office security and privacy is limited, likely caused by a lack of transparency of the services' operations.

Our findings suggest that the current state of cloud office suites leaves much to be desired in the eyes of end users. General misconceptions and the unclear responsibilities of cloud providers might result in additional challenges for end user adoption of cloud office suites.

6.1 Recommendations

Based on our findings, we offer recommendations for groups associated with cloud office suites.

For industry: Since our participants were somewhat unsure about who actually has access to their documents (Section 4.3), we recommend changes to user interfaces and sharing policies that will improve their awareness. In case of unauthorized access, we recommend notifications via email, as most of our participants prefer their provider to inform them this way (Section 4.5). Participants also identified encryption as their preferred security measure their cloud office suite should employ for improved security (Section 4.7).

For end users: A number of self-hosted alternatives to cloud office applications, such as Seafile or NextCloud, allow for most of the cloud conveniences while you retain full control of your data (Section 2).

For policy makers: Our participants are somewhat unsure about who actually has access to their documents and about how many copies actually exist on which servers (Section 4.3). Privacy-focused policies such as GDPR could serve as a first step for improving security and privacy considerations for end users and could enable more privacy-friendly applications. In addition, data-at-rest and responsible disclosure policies could help with user wishes such as prefer encryption measures and notifications by email in case of unauthorized access (Sections 4.5, 4.7).

7 Conclusion

This paper provides a comprehensive insight into the awareness, perceptions, and attitudes of cloud office users, their general usage, and basic mental models.

We find that participants commonly use cloud office suites, mainly for the convenience of free access and easy collaboration for remote documents. Compared to local offline office suites, they voice security and privacy concerns, mainly in terms of unauthorized access. They are somewhat uncomfortable with the security implications of processing their documents in the cloud, however, their threat models remain vague. Our participants have strong opinions on how comfortable they are with the access of certain parties, but are somewhat unsure about who actually has access to their documents. In cases of unauthorized access, participants clearly place the responsibility of informing them of the breach on the cloud office provider, preferably via email.

U.S. and German participants' perceptions, awareness and attitudes closely resembled one another, except that U.S. participants were more uncomfortable with government access to their cloud office data.

We think that, in light of the popularity and widespread use of cloud office suites, participants should be able to make informed decisions about their security and privacy.

We hope that our recommendations for different groups associated with cloud office suites, can help inform future standards, regulations, and implementations.

References

[1] Lilian Adkinson-Orellana, Daniel A Rodríguez-Silva, Felipe Gil-Castiñeira, and Juan C Burguillo-Rial. Privacy for Google Docs: Implementing a transparent encryption layer. In *Proc. 2nd Cloud Computing International Conference - CloudViews*, 2010.

[2] Ibrahim Arpaci, Kerem Kilicer, and Salih Bardakci. Effects of security and privacy concerns on educational use of cloud services. *Computers in Human Behavior*, 45:93–98, 2015.

[3] Daniel Burda and Frank Teuteberg. The role of trust and risk perceptions in cloud archiving—results from an empirical study. *The Journal of High Technology Management Research*, 25(2):172–187, 2014.

[4] K. P. Burnham. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304, 2004.

[5] Kathy Charmaz. *Constructing Grounded Theory*. SAGE Publications, 2014.

[6] Jason W Clark, Peter Snyder, Damon McCoy, and Chris Kanich. I saw images I didn't even know I had: Understanding user perceptions of cloud storage privacy. In *Proc. 33rd ACM Conference on Human Factors in Computing Systems (CHI'15)*, pages 1641–1644. ACM, 2015.

[7] Juliet Corbin and Anselm Strauss. Grounded theory research: Procedures, canons and evaluative criteria. *Zeitschrift für Soziologie*, 19(6):418–427, 1990.

[8] Gabriele D'Angelo, Fabio Vitali, and Stefano Zacchiroli. Content Cloaking: Preserving Privacy with Google Docs and Other Web Applications. In *Proc. 25th ACM Symposium on Applied Computing (SAC'10)*, pages 826–830. ACM, 2010.

[9] Esat Dedezade. Microsoft to deliver cloud services from new datacentres in Germany in 2019 to meet evolving customer needs. https://news.microsoft.com/europe/2018/08/31/microsoft-to-deliver-cloud-services-from-new-datacentres-in-germany-in-2019-to-meet-evolving-customer-needs/, August 2019.

[10] D. Dittrich and E. Kenneally. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. Technical report, U.S. Department of Homeland Security, August 2012.

[11] The Document Foundation. LibreOffice Online | LibreOffice - Free Office Suite. https://www.libreoffice.org/download/libreoffice-online/, June 2020.

[12] Google. Google Drive: Free Cloud Storage for Personal Use. https://www.google.com/intl/en/drive/, June 2020.

[13] Nick Heath. From linux to windows 10: Why did munich switch and why does it matter? https://www.techrepublic.com/article/linux-to-windows-10-why-did-munich-switch-and-why-does-it-matter/, November 2017.

[14] Wenjin Hu, Tao Yang, and Jeanna N Matthews. The good, the bad and the ugly of consumer cloud storage. *Operating Systems Review*, 44(3):110–115, 2010.

[15] Apple Inc. iWork - Apple. https://www.apple.com/iwork/, June 2020.

[16] Iulia Ion, Niharika Sachdeva, Ponnurangam Kumaraguru, and Srdjan Čapkun. Home is Safer Than the Cloud!: Privacy Concerns for Consumer Cloud Storage. In *Proc. 7th Symposium on Usable Privacy and Security (SOUPS'11)*. ACM, 2011.

[17] Eric Johnson. Lost in the cloud: Cloud storage, privacy, and suggestions for protecting users' data. *Stan. L. Rev.*, 69:867–909, 2017.

[18] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Ryan Jewell, and Philip Waggoner. The Shape of and Solutions to the MTurk Quality Crisis. *Political Science Research and Methods*, pages 1–16, May 2018.

[19] Mohammad Taha Khan, Maria Hyun, Chris Kanich, and Blase Ur. Forgotten But Not Gone: Identifying the Need for Longitudinal Data Management in Cloud Storage. In *Proc. 36th ACM Conference on Human Factors in Computing Systems (CHI'18)*, pages 543:1–543:12. ACM, 2018.

[20] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology (2nd ed.)*. SAGE Publications, 2004.

[21] Paul Lorimer. Microsoft office 365 and dynamics 365 now available from new german datacenter regions. https://www.microsoft.com/en-us/microsoft-365/blog/2020/02/20/microsoft-office-365-dynamics-365-now-available-from-new-german-datacenter-regions/, February 2020.

[22] Peter V Marsden and James D Wright. *Handbook of survey research*. Emerald Group Publishing, 2010.

[23] Cathy Marshall and John C Tang. That syncing feeling: early user experiences with the cloud. In *Proc. 9th ACM Conference on Designing Interactive Systems (DIS'12)*, pages 544–553. ACM, 2012.

[24] Charlotte Massey, Thomas Lennig, and Steve Whittaker. Cloudy forecast: an exploration of the factors underlying shared repository use. In *Proc. 32nd ACM Conference on Human Factors in Computing Systems (CHI'14)*, pages 2461–2470. ACM, 2014.

[25] Microsoft. Office 365 | Microsoft Office. https://www.office.com/, June 2020.

[26] Adriana Mijuskovic and Mexhid Ferati. User awareness of existing privacy and security risks when storing data in the cloud. In *Proc. International Conference on e-Learning and the Knowledge Society*, pages 268–273. European Commission, 2015.

[27] David Morris. General Dynamics wins $7.6 billion contract to supply Microsoft office software to the Pentagon. https://fortune.com/2019/08/29/general-dynamics-pentagon-contract/, August 2019.

[28] Michael Nebeling, Matthias Geel, Oleksiy Syrotkin, and Moira C Norrie. MUBox: Multi-user aware personal cloud storage. In *Proc. 33rd ACM Conference on Human Factors in Computing Systems (CHI'15)*, pages 1855–1864. ACM, 2015.

[29] Syynimaa Nestori and Viitanen Tessa. Is My Office 365 GDPR Compliant? - Security Issues in Authentication and Administration. In *Proc. 20th International Conference on Enterprise Information Systems, ICEIS'18*, pages 299–305, 2018.

[30] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior Research Methods*, 46, December 2013.

[31] Stanley Presser, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, Jennifer M. Rothgeb, and Eleanor Singer. Methods for testing and evaluating survey questions. *Public Opinion Quarterly*, 68(1):109–130, March 2004.

[32] Kopo Marvin Ramokapane, Awais Rashid, and Jose Miguel Such. "I feel stupid I can't delete ...": A Study of Users' Cloud Deletion Practices and Coping Strategies. In *Proc. 13th Symposium on Usable Privacy and Security (SOUPS'17)*, pages 241–256, 2017.

[33] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How I Learned to Be Secure: A Census-Representative Survey of Security Advice Sources and Behavior. In *Proc. 23nd ACM Conference on Computer and Communication Security (CCS'16)*. ACM, 2016.

[34] Jim Salter. Office 365 declared illegal in german schools due to privacy risks. https://arstechnica.com/information-technology/2019/07/germany-threatens-to-break-up-with-microsoft-office-again/, July 2019.

[35] Nora Cate Schaeffer and Stanley Presser. The science of asking questions. *Annual Review of Sociology*, 29(1):65–88, 2003.

[36] Cathrin Schaer. Microsoft Office 365: Banned in German schools over privacy fears. https://www.zdnet.com/article/microsoft-office-365-banned-in-german-schools-over-privacy-fears/, Jul 2019.

[37] Ascensio System SIA. Cloud online office suite for business management | ONLYOFFICE. https://www.onlyoffice.com/cloud-office.aspx, June 2020.

[38] Anselm Strauss and Juliet M Corbin. *Grounded theory in practice*. SAGE Publications, 1997.

[39] Dan Svantesson and Roger Clarke. Privacy and consumer risks in cloud computing. *Computer Law & Security Review*, 26(4):391–397, 2010.

[40] Xin Tan and Yongbeom Kim. User acceptance of SaaS-based collaboration tools: a case of Google Docs. *Journal of Enterprise Information Management*, 28(3):423–442, 2015.

[41] Marten van Dijk and Ari Juels. On the Impossibility of Cryptography Alone for Privacy-Preserving Cloud Computing. In *Proc. 6th USENIX Workshop on Hot Topics in Security (HotSec'11)*, pages 1–8. USENIX Association, 2010.

A Survey

The following survey is the English version of the survey, the German version followed the same structure with nearly identical questions. Differences in questions included localization changes, e.g., for country-specific agencies and institutions. Question numbers were not displayed to the participants and order of answer options was generally randomized.

Consent Form

[Consent Form with contact information.]

Please indicate, in the box below, that you are at least 18 years old, have read and understood this consent form, and you agree to participate in this online research study.

- □ I am age 18 or older.
- □ I have read this consent form or had it read to me.
- □ I am comfortable using the English language to participate in this study.
- □ I have used cloud office software before (e.g., Google Drive or Microsoft Office 365).
- □ I agree to participate in this research and I want to continue with the study.

Office demographics

For this survey, we are interested in your experience with and use of Cloud Office Suites and applications.
Cloud Office Application or Online Office Application are software that can be used to create office documents in a web browser, without requiring the installation of a dedicated software.
Examples for Cloud Office Suites are Google Docs/Sheets/Slides, Microsoft Office 365, and LibreOffice Online.

Q1.1: Which office suites have you used before?
(Please select all that apply)

- □ Microsoft Office (Offline; Word, Excel, Powerpoint, ...)
- □ Microsoft Office 365 (Cloud-based; Word, Excel, Powerpoint, ...)
- □ LibreOffice (Offline; Writer, Calc, ...)
- □ LibreOffice Online (Cloud-based; Writer, Calc, ...)
- □ Google Drive (Cloud-based; Docs, Sheets, Slides, ...)
- □ Apple's iWork App (Offline; Pages, Numbers. Keynote...)
- □ Apple's iWork Web (Cloud-based; Pages, Numbers. Keynote...)
- □ OnlyOffice
- □ Other (please specify): _____

Q1.2: Which office suites have you used this month?
(Please select all that apply)

- □ Microsoft Office (Offline; Word, Excel, Powerpoint, ...)
- □ Microsoft Office 365 (Cloud-based; Word, Excel, Powerpoint, ...)
- □ LibreOffice (Offline; Writer, Calc, ...)
- □ LibreOffice Online (Cloud-based; Writer, Calc, ...)
- □ Google Drive (Cloud-based; Docs, Sheets, Slides, ...)
- □ Apple's iWork App (Offline; Pages, Numbers. Keynote...)
- □ Apple's iWork Web (Cloud-based; Pages, Numbers. Keynote...)
- □ OnlyOffice
- □ Other (please specify): _____

Q1.3: Does your job involve using office applications on a regular basis?

- ○ Yes
- ○ No
- ○ I don't know
- ○ I'd prefer not to answer

Q1.4: Which types of documents do you process with office suites?
For this question, please give answers both for your job and your personal life.
(Please select all that apply)

- □ Text (Reports, Letters, etc.)
- □ Spreadsheets (Numbers, Dates, etc.)
- □ Presentations
- □ Calendar and Appointments
- □ Emails
- □ Other (please specify): _____

Q1.5: How do you store your documents?
For this question, please give answers for any documents you might store, including personal and work documents, including but not limited to documents that you edit with office applications.
(Please select all that apply)

- □ Locally on my computer
- □ My office suite saves them online automatically.
- □ Dropbox
- □ Google Drive
- □ Network Share
- □ Self-hosted cloud service
- □ OneDrive
- □ iCloud
- □ Other (please specify): _____

Q1.6: Why do you use **cloud** office applications (compared to local office applications)?
(Please select all that apply)

- □ Provided or required by work
- □ Easy remote access (e.g., from multiple devices)
- □ Ease of collaboration
- □ No installation required
- □ Built-in backup of documents
- □ Free / cheap access
- □ Other (please specify): _____

Document Safety

Q2.1: Where do you think your documents are more secure from any unauthorized access?
[Matrix question, the scale for answers is:]

- More secure on my computer
- Somewhat more secure on my computer
- Equally secure
- Somewhat more secure in the cloud
- More secure in the cloud

- I don't know

[The questions are:]

- Word documents
- Presentations
- Spreadsheets
- E-Mails
- Calendar and Appointments

Q2.2: Why (if at all) do you think your documents may be more secure **on your computer**?
[Free text field]

Q2.3: Why (if at all) do you think your documents may be more secure **in the cloud**?
[Free text field]

Document Access

Q3.1: Who else besides yourself might be able to access the documents you edit in cloud office applications?
(Please select all that apply)

- □ People I share the documents with
- □ My employer
- □ My internet provider
- □ The cloud office provider (e.g., Google or Microsoft)
- □ My browser vendor (e.g., Google or Mozilla)
- □ My operating system manufacturer (e.g., Apple or Microsoft)
- □ Cybercriminals (e.g., hackers or organized crime)
- □ Law enforcement or intelligence agencies (e.g., police, FBI or NSA)
- □ Third parties (e.g., online advertisers or plugin developers)
- □ The manufacturer of my computer hardware (e.g., Intel, AMD, Apple, or Lenovo)
- □ Other (please specify): _____

[The following 3 questions are matrix questions with the following options:]

- People I share the documents with
- My employer
- My internet provider
- The cloud office provider (e.g., Google or Microsoft)
- My browser vendor (e.g., Google or Mozilla)
- My operating system manufacturer (e.g., Apple or Microsoft)
- Cybercriminals (e.g., hackers or organized crime)
- Law enforcement or intelligence agencies (e.g., police, FBI or NSA)
- Third parties (e.g., online advertisers or plugin developers)
- The manufacturer of my computer hardware (e.g., Intel, AMD, Apple, or Lenovo)

Q3.2: Where do you think the risk is higher that the following parties can obtain unauthorized access to your cloud office documents?

- Higher risk on my computer
- Somewhat higher risk on my computer
- Equal risk
- Somewhat higher risk in the cloud
- Higher risk in the cloud
- I don't know

Q3.3: Do you think that any of these parties have already accessed your documents?

- Yes
- No
- I don't know

Q3.4: Please rate your level of (dis)comfort with the potential access of these parties to your cloud office documents.

- Completely comfortable
- Somewhat comfortable
- Neither
- Somewhat uncomfortable
- Completely uncomfortable
- I don't know

Q3.5: Who do you think **would** inform you if an unauthorized party or person accessed you documents?
(Please select all that apply)

- □ People I share the documents with
- □ My employer
- □ My internet provider
- □ The cloud office provider (e.g., Google or Microsoft)
- □ My browser vendor (e.g., Google or Mozilla)
- □ My operating system manufacturer (e.g., Apple or Microsoft)
- □ Law enforcement or intelligence agencies (e.g., police, FBI or NSA)
- □ Third parties (e.g., online advertisers or plugin developers)
- □ The manufacturer of my computer hardware (e.g., Intel, AMD, Apple, or Lenovo)
- □ The news
- □ Scientists
- □ Nobody would inform me
- □ Other (please specify): _____

Q3.6: Who do you think **should be responsible** for informing you if an unauthorized party or person accessed your documents?
(Please select all that apply)

- □ People I share the documents with
- □ My employer

- □ My internet provider
- □ The cloud office provider (e.g., Google or Microsoft)
- □ My browser vendor (e.g., Google or Mozilla)
- □ My operating system manufacturer (e.g., Apple or Microsoft)
- □ Law enforcement or intelligence agencies (e.g., police, FBI or NSA)
- □ Third parties (e.g., online advertisers or plugin developers)
- □ The manufacturer of my computer hardware (e.g., Intel, AMD, Apple, or Lenovo)
- □ The news
- □ Scientists
- □ Nobody would inform me
- □ Other (please specify): _____

Q3.7: **How** would you like to be informed if an unauthorized party or person accessed your cloud office documents?
[Free text field]

Document Storage

Q4.1: Do you think that multiple copies of your cloud office documents exist?
These can be documents that are shared with others or private documents.

- ○ Yes
- ○ No
- ○ I don't know
- ○ I'd prefer not to answer

Q4.2: [only shown if Q4.1 = Yes] For which purpose do you think these copies might exist?
[Free text field]

Q4.3: [only shown if Q4.1 = Yes] In which geographic locations do you think your cloud office documents and copies of these are stored?
[Free text field]

Q4.4: [only shown if Q4.1 = Yes] Which of the copies do you think are actually removed if you delete a cloud office document?

- ○ All
- ○ Mine and my collaborators'
- ○ Only mine
- ○ Only my collaborators'
- ○ None
- ○ I don't know
- ○ I'd prefer not to answer
- ○ Other (please specify):

Q4.5: [only shown if Q4.1 = Yes and Q4.4 != All] **Where or with whom** do you think copies remain?
[Free text field]

Q4.6: [only shown if Q4.1 = Yes and Q4.4 != All] For which purpose do you think that the copies remain?
[Free text field]

Q4.7: Who do you think can delete your documents?
(Please select all that apply)

- □ People I share the documents with
- □ My employer
- □ My internet provider
- □ The cloud office provider (e.g., Google or Microsoft)
- □ My browser vendor (e.g., Google or Mozilla)
- □ My operating system manufacturer (e.g., Apple or Microsoft)
- □ Cybercriminals (e.g., hackers or organized crime)
- □ Law enforcement or intelligence agencies (e.g., police, FBI or NSA)
- □ Third parties (e.g., online advertisers or plugin developers)
- □ The manufacturer of my computer hardware (e.g., Intel, AMD, Apple, or Lenovo)
- □ Other (please specify): _____

Q4.8: Who do you think is responsible for protecting your data?
(Please select all that apply)

- □ People I share the documents with
- □ My employer
- □ My internet provider
- □ The cloud office provider (e.g., Google or Microsoft)
- □ My browser vendor (e.g., Google or Mozilla)
- □ My operating system manufacturer (e.g., Apple or Microsoft)
- □ Cybercriminals (e.g., hackers or organized crime)
- □ Law enforcement or intelligence agencies (e.g., police, FBI or NSA)
- □ Third parties (e.g., online advertisers or plugin developers)
- □ The manufacturer of my computer hardware (e.g., Intel, AMD, Apple, or Lenovo)
- □ Myself
- □ The US-Government
- □ Other (please specify): _____

Responsibility

Q5.1: Please indicate your agreement with the following statements:
[5 point-likert scale from Strongly agree to Strongly disagree + I don't know option]

- Cloud office providers should offer adequate protection for cloud office documents (e.g., by encryption and well implemented security practices)
- I should have the right to demand a full overview of my data collected by cloud office providers.

- Upon my request, cloud office providers should have to show what they do with my documents and who has or had access.
- Cloud office providers must be able to modify or delete any data they have on private individuals.

Q5.2: Please indicate your (dis)comfort with the following statements:
[5 point-likert scale from Completely comfortable to Completely uncomfortable + I don't know option]

- Cloud providers can store my documents on servers outside of the US without legal repercussions.
- US regulations and laws still apply if the documents are stored on servers outside of the US.
- US law enforcement can access my cloud documents without a court order.
- US law enforcement can force me to give up my cloud office password.

Q5.3: Where do you think the risk is higher of somebody **obtaining unauthorized access** to your documents if they are either stored on a server in Germany or the US?
[5 point-likert scale from "Higher risk for server in Germany" to "Higher risk for server in the US" + I don't know option]

- My employer
- My internet provider
- The cloud office provider (e.g., Google or Microsoft)
- My browser vendor (e.g., Google or Mozilla)
- My operating system manufacturer (e.g., Apple or Microsoft)
- Cybercriminals (e.g., hackers or organized crime)
- Third parties (e.g., online advertisers or plugin developers)
- The manufacturer of my computer hardware (e.g., Intel, AMD, Apple, or Lenovo)
- US government
- German governments
- Foreign government (neither US nor German)

Personal Perception - Scenario A - Personalized Scenario

[Only scenario block A or B was randomly shown to the participants]
[Question order was randomized]
Below are listed three different scenarios. How comfortable do you feel with each approach?

Q6.A.1: Your child is required by the school to use a cloud office suite for tasks. The processed documents include private information such as your child's name and grades.

- ○ Completely comfortable
- ○ Somewhat comfortable
- ○ Neither
- ○ Somewhat uncomfortable
- ○ Completely uncomfortable
- ○ I don't know

Q6.A.2: Your general practitioner uses a cloud office suite to process patient data. The processed documents include private information such as your name, age, weight, diagnosis, and treatment plan.

- ○ Completely comfortable
- ○ Somewhat comfortable
- ○ Neither
- ○ Somewhat uncomfortable
- ○ Completely uncomfortable
- ○ I don't know

Q6.A.3: Your financial advisor uses a cloud office suite to process client data. The processed documents include private information such as your name, SSN, and financial information.

- ○ Completely comfortable
- ○ Somewhat comfortable
- ○ Neither
- ○ Somewhat uncomfortable
- ○ Completely uncomfortable
- ○ I don't know

Personal Perception - Scenario B - Generalized Scenario

[Only scenario block A or B was randomly shown to the participants]
[Question order was randomized]
Below are listed three different scenarios. How comfortable do you feel with each approach?

Q6.B.1: A school requires children to use a cloud office suite for tasks. The processed documents include private information such as children names and grades.

- ○ Completely comfortable
- ○ Somewhat comfortable
- ○ Neither
- ○ Somewhat uncomfortable
- ○ Completely uncomfortable
- ○ I don't know

Q6.B.2: A doctor's office uses a cloud office suite to process patient data. The processed documents include private information such as name, age, weight, diagnosis, and treatment plans.

- ○ Completely comfortable
- ○ Somewhat comfortable
- ○ Neither
- ○ Somewhat uncomfortable
- ○ Completely uncomfortable
- ○ I don't know

Q6.B.3: A financial advisor's office uses a cloud office suite to process client data. The processed documents include private information such as name, SSN, and financial information.

- ○ Completely comfortable
- ○ Somewhat comfortable
- ○ Neither
- ○ Somewhat uncomfortable
- ○ Completely uncomfortable
- ○ I don't know

Data Protection

Q7.1: What do you think — what data does the cloud office application collect when you process documents with it?
[Free text field]

Q7.2: How do you think documents processed by cloud office applications are protected?
[Free text field]

GDPR

Q8.1: Do you know what the GDPR is?

- ○ A data protection regulation in EU law
- ○ A plugin for Google Drive
- ○ A cloud office provider
- ○ A counter terrorism act in US law
- ○ I don't know
- ○ I'd prefer not to answer

Q8.2: [Only shown if Q8.1 = A data protection regulation in EU law] What do you think does the GDPR protect?
[Free text field]

Demographics

[We administered demographic questions at the end of the questionnaire to prevent stereotype bias.]

Q9.1: How old are you? (in years, e.g. 42. Optional)
[Free text field]

Q9.2: As which gender do you identify?

- ○ Male
- ○ Female
- ○ [Free text field]
- ○ I'd prefer not to answer

Q9.3: Do you have formal education (Bachelor's degree or higher) in computer science, information technology, or a related field?

- ○ Yes
- ○ No
- ○ I'd prefer not to answer

Q9.4: Have you held a job in computer science, information technology, or a related field?

- ○ Yes
- ○ No
- ○ I'd prefer not to answer

Q9.5: Do you have any feedback or additional comments for us? (completely optional)
[Free text field]

From Intent to Action: Nudging Users Towards Secure Mobile Payments

Peter Story
Carnegie Mellon University
pstory@andrew.cmu.edu

Daniel Smullen
Carnegie Mellon University
dsmullen@cs.cmu.edu

Alessandro Acquisti
Carnegie Mellon University
acquisti@andrew.cmu.edu

Lorrie Faith Cranor
Carnegie Mellon University
lorrie@cmu.edu

Norman Sadeh
Carnegie Mellon University
sadeh@cs.cmu.edu

Florian Schaub
University of Michigan
fschaub@umich.edu

Abstract

Despite experts agreeing on many security best practices, there remains a gap between their advice and users' behavior. One example is the low adoption of secure mobile payments in the United States, despite widespread prevalence of credit and debit card fraud. Prior work has proposed nudging interventions to help users adopt security experts' recommendations. We designed and tested nudging interventions based on *protection motivation theory* (PMT) and *implementation intentions* (II) to encourage participants to use secure mobile payments. We designed the interventions using an interview study with 20 participants, and then tested them in a longitudinal, between-subjects field experiment with 411 participants. In one condition, drawing on PMT, we informed participants about the threat of card fraud and the protection offered by mobile payments. In a second condition, we combined the PMT intervention with an II-based intervention, and asked participants to formulate a plan to make a mobile payment in the week ahead. A third condition acted as a control. Both PMT-only and PMT+II interventions made participants more likely to make mobile payments relative to the control group. The results suggest that PMT and implementation intention-based nudges can help people translate their desire to behave securely into actual behavior.

1 Introduction

Survey research consistently shows that people are concerned about their security and privacy [36, 53]. However, research also shows that people often do not take steps to protect themselves [14, 45]. A prime example is the continued prevalence of payment card usage (and the elevated risk of fraud associated with them) in the United States, despite the availability of alternative payment technologies that could limit fraud. Card fraud remains enormously lucrative for criminals, and compromised point-of-sale terminals are a major source of card information [29, 30]. Mobile payments (e.g., Apple Pay) incorporate security measures that protect against this threat (i.e., payment tokenization) [10, 21, 58], and are widely available [51]. Despite this, adoption of mobile payments in the United States lags far behind other countries [50, 57]. The reasons for this lag may be multiple and various, from status quo bias, to faith in card issuers' willingness to cover losses, to lack of awareness about the threat of card fraud and the protection offered by using mobile payments.

In the context of security and privacy, behavioral economics provides convincing explanations for why people sometimes act in manners that subtly diverge from their expressed preferences, including failing to protect their security and privacy even when users claim security and privacy to be important to them [2, 3]. Researchers have proposed a variety of ways to help users protect themselves, and frameworks incorporating *nudges* are especially promising [4]. Nudges are carefully crafted interventions that help users to act in ways that align with their stated preferences. The privacy and security literature demonstrates the effectiveness of a number of different nudges in this domain [4–6, 8, 17]. However, it is notable that *implementation intentions* are largely unstudied in the field of security and privacy, despite decades of strong support for their effectiveness in the medical domain [19, 32, 56]. Implementation intentions are contextually activated mental plans that help people follow through on their goals. Implementation intentions appear to be well-suited to situations where a person needs to remember to take an action to protect themselves, such as remembering to pay using a secure mobile payment system instead of swiping their card [22]. Prior work suggests that implementation intentions are effective when paired with *protection motivation theory*-based (PMT) interventions [42]. Protection motivation theory claims that users'

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

protection-seeking actions are based on their perceptions of threats and possible responses [43, 54, 62]. In the context of mobile payments, our interventions aimed to educate participants about the threat of card fraud and the protection offered by using mobile payments instead of physical cards.

To investigate the impact of PMT and implementation intention nudges on the adoption of mobile payments, we first conducted a series of interviews with 20 participants to understand people's thoughts about card fraud, mobile payments, and our implementation intention plan (§ 3). Next, informed by the findings from our interviews, we conducted a longitudinal, randomized controlled experiment with 411 participants to measure the effect of two interventions on participants' use of a specific payment system (Apple Pay) (§ 4). We preregistered our study design prior to collecting any data. Our results showed that participants in our PMT-only and PMT with implementation intention treatment groups were 2.4x ($p = 0.020$) and 3.9x ($p < 0.001$) more likely to use Apple Pay than were participants in our control group, respectively. Our findings further suggest that adding an implementation intention to our PMT-only treatment increased its efficacy (1.7x more likely, $p = 0.085$). These results show that PMT and implementation intention-based nudges can be a powerful tool for helping people translate their intention to behave securely into actual behavior. Finally, we discuss the implications of our findings (§ 5) and our conclusions (§ 6).

2 Related Work

In this section we describe the rationale for nudging people to adopt mobile payments (§ 2.1), previous use of nudges in the context of security and privacy (§ 2.2), and how protection motivation theory and implementation intentions informed the design of our nudges (§ 2.3).

2.1 Card Fraud and Payment System Security

Credit and debit card fraud takes place when a criminal either obtains a physical card or obtains information about a card, and then initiates a transaction that the card owner did not consent to. Two major sources of card information are compromised point of sale terminals (POSs) [29, 30] and compromised retail websites [28]. The interventions we study focus on protecting against compromised POSs. Criminals can compromise POSs by remotely installing malware on them [30], but physical compromise is also possible (e.g., using skimmers) [26]. When a card's magnetic stripe is swiped on a compromised POS, the information on the magnetic stripe can be recorded and used to make counterfeit cards. When a card's chip is inserted into a compromised POS, in the best case the only useful information that can be stolen are the card number, cardholder name, and expiration date [27]. However, chips implement a variety of EMV protocols, some of which are susceptible to card cloning (e.g., static data card authentication) and have other known weaknesses [25, 69]. Contactless card transactions suffer from similar weaknesses [69]. Thus, compromised POSs can trivially steal valuable magnetic stripe data and still pose a threat to EMV cards.

Apple Pay, Google Pay, and Samsung Pay allow users to register credit or debit cards on their phone, so that payments can be made through the phone rather than with physical cards. These mobile payments systems can protect against compromised POSs. First, Pay[1] always uses transaction-specific codes [10, 21, 58], making re-use of transaction information theoretically impossible. Second, Pay uses a device-specific card number in transactions, so a compromised POS cannot steal the original card number. Thus, mobile payments are a strong protection against compromised POSs. In addition, unlike most EMV cards in the United States, a biometric or PIN is used to authenticate the user before a transaction can be made, making it much more difficult for a thief to use a stolen device with Pay.

Despite their security benefits, fewer than half of Americans use mobile payments regularly [50, 51, 57]. As noted in the Introduction, the reasons behind the lag in adoption may be various. For instance, research by Pew and Huh et al. suggest that it is partly due to people's belief that smartphone-based payment systems are less secure than paying with physical cards [22, 50]. Furthermore, Huh et al.'s participants reported that lack of availability, convenience, and forgetfulness were reasons why they did not use Apple Pay or Android Pay [22]. These findings led us to test using nudges to correct people's misconceptions and to encourage them to start using Pay.

2.2 Nudges for Security and Privacy

A large body of research examines ways to help people protect their security and privacy. For example, researchers have studied how to improve privacy notices [24], how to guide people towards choices that fit their preferences [34], and how a lack of usability can inhibit adoption of security tools [31, 72]. This varied research is unified by the acknowledgment that people have limited cognitive resources and suffer from behavioral biases.

Inspired by work in psychology and behavioral economics [67], researchers are increasingly studying how *nudges* can improve design for security and privacy [4]. Nudges are design elements that help people overcome their cognitive and behavioral biases in order to make decisions which align with their stated preferences. For example, Almuhimedi et al. used nudges to mitigate the information asymmetry between users and the behaviors of apps on their devices [7, 8]. Their nudges were successful at encouraging users to reassess and restrict permissions settings. Frik et at. tested using nudges to overcome present bias [17]. They

[1] In the rest of this manuscript, we use *Pay* to refer to Apple Pay, Google Pay, and Samsung Pay generically.

found that users given the option to be reminded later were less likely to completely dismiss prompts for security updates and 2FA configuration. Albayram et al. and Al Qahtani et al. used educational videos to motivate participants to enable lock screens on their smartphones [5, 6]. In both studies, the videos successfully motivated many participants to enable secure lock screens. However, nudges are not always effective [4], which shows the value of empirical research like ours.

2.3 Nudges to Protect Against Card Fraud

Our goal was to test whether nudges that focus on security can induce individuals to consider using mobile payments to protect themselves, and whether those nudges can help them translate that intention into actual behavior. *Implementation intentions* are designed to help people translate intention into behavior, especially in cases when people must remember to take some action (i.e., use Pay instead of swiping a card). An implementation intention is a concrete, contextually activated plan to achieve a goal [19]. The plan should be specific (e.g., specifying location or time), which facilitates the plan being triggered in the planner's mind by contextual factors (e.g., when they arrive at a certain location). In our study, we encouraged participants to fill in a plan template detailing up to three locations where they would make a mobile payment in the week ahead (Figure 12).

Research in the medical domain has shown that encouraging people to form implementation intentions can have a powerful effect on people achieving their goals. For example, Milne et al. successfully used implementation intentions to encourage young people to exercise in order reduce their future risk of heart disease; 91% of the participants who formed implementation intentions exercised in the week after treatment, as compared to only 35% of the group that was only exposed to motivational materials [42]. Implementation intentions have been shown to be effective in many other contexts [40, 41, 44, 46, 47]. However, with the exception of a study design described by Liao et al. [33], we are unaware of explicit application of implementation intentions in the domain of security and privacy. Thus, a contribution of our work is bringing awareness of implementation intentions to our research community.

In a review of studies of implementation intentions, Gollwitzer explains why and when implementation intentions are effective [19]. Implementation intentions are effective because they help people remember to perform their planned action. Also, planning the details of the action reduces the amount of conscious effort needed when it comes time to perform the action. Implementation intentions are most likely to be effective when the person has a strong commitment to both their plan and to the goal that motivates the plan. Thus, implementation intentions to protect security and privacy should have an effect when users are motivated to take action to protect themselves. In order to motivate participants in our study, we draw on Protection Motivation Theory (PMT) [37, 54, 55]. PMT has been applied in both the medical domain [43, 74] and in computer security [5, 6, 12, 60–62]. PMT proposes that people are more likely to take action to protect themselves from a threat when they perceive that the threat is severe (i.e., greater perception of *threat severity*), that they are susceptible to the threat (i.e., greater perception of *threat susceptibility*), that the action they could take is not too difficult to perform (i.e., greater perception of *self-efficacy*), and that the action they could take will be effective in protecting against the threat (i.e., greater perception of *response efficacy*) [43, 74]. PMT has been effectively combined with implementation intentions in domains outside of security and privacy. For example, in their study of implementation intentions for exercise, Milne et al. used PMT to motivate participants [42].

In summary, in our study we designed and tested nudges to help participants protect themselves from card fraud by adopting mobile payments. In particular, we designed our nudges based on PMT and implementation intentions, a combination which, to the best of our knowledge, we are the first to test in the domain of security and privacy.

3 Qualitative Interviews

The first part of our study focused on gathering qualitative information on people's thoughts about the threat of card fraud, the use of Pay[1] to protect against card fraud, and people's experiences forming implementation intention plans to use Pay. We conducted a series of surveys and interviews to gather longitudinal self-reported data about participants' experiences. Our findings informed the design of our controlled experiment (§ 4), allowing us to refine our interventions, correct common misconceptions, and to understand some of the limitations of our approach. All of our study protocols were approved by Carnegie Mellon University's IRB.

3.1 Protocol

This portion of our study included three surveys and two interviews (illustrated in Figure 7 in the appendix). We recruited participants from Craigslist and Carnegie Mellon University's participant pool. Survey #1 gathered information about users' devices, prior use of payment methods, perceptions of the likelihood and severity of card information theft and fraud, prior experience with card information theft and fraud, and demographic information. We reasoned that our nudges would have the largest impact on people who were not already using Pay,[1] but whose phones were compatible with Pay and were likely to have opportunities to use Pay. Thus, we screened out participants who reported having used Pay in a physical location in the past month, we required that participants had made at least one payment with a credit or debit card in a physical location in the past month, and we required that their

smartphone be compatible with either Apple Pay, Google Pay, or Samsung Pay.

We invited a diverse subset of qualifying participants to participate in a semi-structured interview (Interview #1). In accordance with purposive sampling, our attempt to balance several factors of interest (e.g., phone type, age, occupation, fraud-related perceptions, etc.) influenced our choice of who to invite to the interview. 20 participants attended Interview #1. 75% of our participants were female, their median age was 26.5, 55% had iPhones, 25% had Samsung phones, and 20% had other Android phones. The interview started with a discussion of prior experiences with card fraud, card information theft, and prior experiences with Pay. Next, the interviewer described recent cases of card information theft from major retailers, and the potential consequences of such theft for the participant. This intervention was included in order to help participants develop an accurate perception of their *susceptibility* to card fraud and the potential *severity* of card fraud, two elements of *threat appraisal* that protection motivation theory (PMT) suggests are associated with protective behavior [43]. Then, the interviewer described how Pay may protect against card information theft, presented the participant with instructions for setting up and using Pay, and gave the participant the opportunity to set up Pay if they wanted to. This intervention was included in order to help participants understand how Pay may help protect them from card fraud and to give them confidence that they can use Pay, influencing perceptions of *response efficacy* and *self-efficacy*, two additional elements of PMT. Next, participants were given an opportunity to form an implementation intention by filling out a paper template. The template encouraged participants to plan where they might use Pay in the coming week and to mentally rehearse using Pay in these locations. These activities were designed to help mentally activate participants' plans to use Pay when they were in these locations [19]. Finally, participants were given the opportunity to express a strong commitment to their plan, which prior work suggests increases the efficacy of implementation intentions [19, 63]. The template was similar in content to the template in our controlled experiment (see Figure 12).

One week after completing Interview #1, participants were sent Survey #2, which asked whether participants had set up Pay after the interview, whether they had tried to use Pay, and whether they had successfully used Pay. Participants who completed Interview #1 and Survey #2 were compensated with a $15 Amazon e-gift card.

Participants who completed Survey #2 and who had set up Pay on their phones were invited to Interview #2, which was designed to understand people's experiences using Pay or their reasons for not using it. We also asked questions about whether participants followed their implementation intention plans and whether they found the plans to be helpful. Participants in Interview #2 were compensated with an additional $15 Amazon e-gift card.

Four weeks after completing Survey #2, participants who had set up Pay on their phones were invited to take Survey #3, which asked whether participants had used Pay in the past week. We also asked whether participants thought they were likely to use Pay in the future. Our surveys and interview scripts are included in the appendix (§ 8.1-8.5).

3.2 Analysis

We used thematic coding to analyze transcripts of our interviews and our survey's open-text responses. Two of the authors reviewed these materials together and collaboratively developed a codebook. To ensure that the codes we developed later were consistently applied to the materials we analyzed earlier, one author then re-reviewed all the materials. Since our goal for this portion of our study was to gather rich, qualitative data, we did not attempt to calculate measures of annotator reliability [39]. Table 4 in the appendix contains our final codebook and the frequencies of our codes.

3.3 Results

Below we summarize key takeaways from our survey and interview data. Although in some cases we report the frequency of codes, due to our use of purposive sampling in selecting participants, it would be inappropriate to assume that these frequencies correspond to the frequencies that might be observed in the general population.

Use of Pay

We received 288 complete responses to Survey #1. Among these respondents, only 34.7% reported using Pay sometime in the past, and a mere 23.6% reported using Pay in the past month. We recruited only respondents who had not used Pay in the last month for Interview #1. In Interview #1, nearly all participants (19/20) said they had heard of Pay before our study, but only one participant reported using it to pay in a physical location before. Multiple participants mentioned seeing Pay in advertisements, seeing it on their phone, seeing it as a payment option, using it for digital purchases, or knowing that friends or family use it. This widespread awareness of Pay makes sense, considering that many smartphones come with Pay preinstalled [20, 59] and that iPhones include persistent reminders to set up Apple Pay [23].

Prior to Interview #2, 11/20 participants had Pay set up on their phone, and so could have used it before Interview #2. One participant had set up Pay prior to Interview #1, four set it up in Interview #1, and six set it up after Interview #1. Participants gave a variety of different reasons for not setting up Pay including being too busy, not thinking they needed it, wanting to do more research, and wanting to consult their partner. Between Interview #1 and Interview #2, seven

participants used Pay. Three of these participants used Pay successfully at at least one of the locations in their plan.

To understand whether participants were likely to continue using Pay after our study, four weeks after completing Survey #2 we sent Survey #3 to all participants who had set up Pay. In response to Survey #3, three participants indicated that they had successfully used Pay to make a payment in a physical location in the prior week.

Our results suggest that despite widespread awareness of Pay, most people are not using it regularly. However, after being exposed to our nudges in Interview #1, a substantial percentage of participants (35%) used Pay at least once during the remainder of our study. Furthermore, responses to Survey #3 indicate that our nudges may increase use of Pay long after the initial intervention. These results encouraged us to move forward with the controlled experiment described in § 4.

Perceptions of Threat Susceptibility and Severity

All but one participant recounted their own or others' experiences with card fraud. Fewer participants (10/20) recounted experiences with card information theft. When asked to describe experiences with card information theft, seven participants instead described cases of card fraud or theft. This makes sense, given that card information theft can be difficult for individuals to detect directly.

After we described recent hacks in which credit and debit card information was stolen and the possible consequences of having one's information stolen, we asked questions to gauge participants' levels of concern about and perceptions of susceptibility to card fraud and information theft.

Participants expressed varied opinions about their susceptibility to these threats. Eleven participants expressed that information theft happened frequently (P14: "It just seems like it does happen so frequently..."), but three participants said such occurrences were infrequent (P6: "Cause even like ... the hacking things you mentioned, I mean they're once in a blue moon."). Three participants said their behavior made theft or fraud more likely, but nine others thought their behavior lowered their likelihood of suffering from card information theft or fraud.

Participants described a number of negative outcomes associated with card theft and fraud, including the hassle and stress of dealing with it, feelings of anger and helplessness, loss of money due to theft or overdraft fees, and the fear that additional bad things might happen to them. Participants also mentioned that their level of concern would depend on the size of the fraudulent purchase and whether the purchase was on their credit or debit card. Ten participants expressed confidence that their card issuer would help them resolve fraudulent purchases, and two even thought they would be refunded under all circumstances. It is potentially a misconception to believe that fraudulent charges will be refunded in all cases, since U.S. law does not require this of card issuers [13].

Our takeaway is that while most participants have a high level of awareness of the possibility of card fraud, some people remain under-informed and might benefit from additional information.

Perceptions of Self-efficacy

Some participants thought Pay setup was easy, but others encountered difficulties. In particular, two participants were confused by Apple Pay's ability to automatically add card details using the phone's camera and three mentioned interacting with their bank to approve registering their card as a challenge. Additionally, two participants found that certain cards simply could not be added to Pay. Seven participants said that setup or use would be a challenge, and would require practice, learning, or attention to detail.

Eleven participants said they did not (or might not) have opportunities to use Pay because they did not go shopping, did not have enough money, or due to other reasons.

Participants described different challenges they might (or did) encounter in stores using Pay. First, stores might or might not accept Pay. Second, participants might not remember to use Pay, suggesting an opportunity for implementation intentions to help in this area. Third, participants might experience difficulty using Pay. Despite our written instructions, some participants still had questions about how to use Pay. Thus, we included a short video alongside written instructions in our controlled experiment (§ 4). Participants also described positive aspects of Pay. Some participants expressed that Pay was easy to use, that it would allow them to not carry or take out their cards or wallet, that it would be a good backup option if they didn't have a card, and that it would be fun to try something new.

Two usability challenges in particular may be of interest, due to their potential generalizability: the case of accidental activation and the case of failure to activate. Four of our participants who set up Apple Pay described accidentally activating it and not knowing why this was happening. Not understanding this accidental activation alarmed at least one of our participants (P19: "The credit thing keeps popping up whenever I angle my phone a certain direction. I wonder where it's sending my credit info each time."). It is possible to open Apple Pay by either double-clicking the home button when the phone is locked or by bringing the phone in proximity to an NFC reader (even if the NFC reader is not a payment terminal). To address some of this confusion, we added the double-clicking functionality to our instructions in subsequent interviews and in our experiment. One of these same participants (P11) also experienced the problem of Apple Pay not activating. At one location, this participant reported having to scan their phone twice before it worked. At another location, the participant was ultimately unsuccessful using Apple Pay, concluding that it must not have been supported and expressing frustration with this failure mode:

"What happens when it doesn't work is nothing happens. It just sits there. And it doesn't even apologize. You know it doesn't say anything on it. 'Oops, sorry. Try again.' Nothing like that." Unfortunately, due to the lack of an NFC signal in the case when a terminal does not support NFC payments, it is hard to imagine a technical solution to this kind of silent failure mode. Thus, while some of these usability challenges may be addressable through education, some may be inherent to the technology.

Perceptions of Response Efficacy

Most participants (14/20) expressed some confidence in the security properties of Pay that we described. However, nine participants also expressed concern about Apple, Google, or their phone being hacked. P16 cited their previous experience having their iTunes account hacked as a reason for not believing that Apple Pay would protect their card information: "[T]he only time I've been hacked was with an Apple product. That's the only reason. ... [T]he only time I had a fraudulent charge was when I was with an Apple product." Interestingly, this participant also recognized that the hack was likely due to their choice of a weak password, saying: "I guess my password wasn't as secure as I thought it was." P11 said that "I feel as if the phone is more vulnerable than the computer." P8 expressed a more concrete concern about NFC signal skimming, expressing concern that "...in a physical store ... the person behind you can actually take your information if they know what they're doing on the phone."

Despite participants' concerns about hacking, Apple Pay is designed to be resistant to hacking: card information is not stored with Apple after the initial enrollment process, mitigating the risk of data breach, device-specific Device Account Numbers are stored on each phone's Secure Element, protecting against phones being compromised, and user interaction is required before making payments [10]. Google Pay and Samsung Pay employ similar protections [21, 58]. Of course, attacks that can thwart these protections are possible (e.g., a persistent threat on Apple's servers), but such attacks would require substantially more resources than simply adding card skimmers to point of sale terminals. Communicating useful mental models to non-technical users remains an open research area [71]. Our participants' responses point to the challenge of communicating complex threat models to a general audience.

Awareness of Protection Actions

Participants demonstrated awareness of many different ways they could protect themselves from credit and debit card information theft and fraud. The most prevalent actions involved working with one's card issuer, such as reporting fraudulent purchases or receiving a new card. Actions involving physical awareness (e.g., looking for card skimmers), monitoring card statements for unauthorized transactions, protecting access to one's account (e.g., with a strong password), using cash, or using a credit card (e.g., due to liability protections) were also common. Interestingly, two participants brought up the possibility of using Apple Pay to protect themselves before we had described it as being a secure payment method (but after we had asked them whether they had used it). P18 even gave an accurate explanation of why Apple Pay might be more secure: "Maybe I could use Apple Pay or something. Then if I don't give my card information directly to these companies or grocery stores, if I go via a secure party like Apple Pay, it should be a good option."

Our overall takeaway is that most participants are aware of some ways they can protect themselves from card information theft and fraud. Unfortunately, prior work and the continued profitability of card fraud suggest that people's ability to protect themselves is limited (e.g., password re-use is prevalent [48]). In addition, most participants seemed unaware that Pay could protect them before we explained that it could, suggesting our information about Pay may be helpful.

Effectiveness of Implementation Intentions

All participants were given the opportunity to form an implementation intention plan to help them remember to use Pay. 16/20 participants wrote or described at least one location where they might use Pay. About half of participants checked or otherwise indicated that they performed at least one mental rehearsal activity. As we conducted interviews, we refined the way we introduced the plan to communicate that filling out and following the plan were not mandatory, but that filling out the plan was encouraged if the participant wanted to remember to use Pay. Participants described several obstacles to forming an implementation intention, including not being able to think of places they would visit, not having decided whether they wanted to use Pay, and simply thinking the plan wouldn't be helpful for them. In addition, four participants had at least some difficulty remembering their plan in Interview #2. The act of forming a plan seemed to help four participants understand where Pay could be used. For example, P11 realized that it would be difficult to use Pay at a restaurant where waiters collect cards for payment processing. Participants also described other things that could remind them to setup or use Pay, including receiving notifications from Google Pay about availability, adding Pay to their shopping list, and putting the Pay app on their home screen.

Of the seven participants who used Pay between Interview #1 and Interview #2, three used Pay successfully at at least one of the locations in their plan. As the majority of participants were able to form plans, and some of the participants who formed plans went on to use Pay at their planned locations, we thought that our implementation intention plan template was worth testing in our controlled experiment (§ 4). At the same time, our plan may be unhelpful to participants who

have difficulty thinking of locations they are likely to make payments in the coming week, and is almost certain to be unhelpful for participants who simply decide not to use Pay. Since Pay is not available in all locations, it is unsurprising that many participants had questions about where they could use Pay. As part of our description of Pay, we described just four popular locations where Pay can be used in our city. With a more comprehensive list of locations, it might be possible to develop an interactive plan template which could contribute to greater awareness of where Pay is available.

Misconceptions and Other Concerns

Our interviews helped us identify a number of misconceptions related to Pay. For example, four participants thought Pay might interfere with their credit card rewards (P5, P9, P12, P16), four participants thought our study was affiliated with Apple (P11, P16, P19, P20), three participants wondered if Pay cost something (P5, P8, P15), and one participant thought Pay might prevent them from getting receipts (P12). In addition, P7 thought Samsung Pay was a credit card and two participants confused Apple Pay with iPad-based point of sale terminals (e.g., Square). We addressed several of these misconceptions in a "Frequently Asked Questions" section in our controlled experiment (Figure 11).

Pay on watches offers the same level of security as on phones, but with potentially greater convenience. Thus, we were surprised that all three of the participants we spoke with about their smartwatches expressed skepticism about using Pay with their watches. P1 thought they would start using Apple Pay on their iPhone, because they thought they would need to practice the motion of making payments with their Apple Watch. P12 thought Apple Pay would be less secure on their Apple Watch than on their iPhone because their Apple Watch did not have a fingerprint reader. Neither P1 nor P12 set up Apple Pay during our study. In Interview #1, P15 was worried that setting up Samsung Pay might allow transactions to be made through their Samsung Watch without their knowledge, due to the fact that their watch did not have a PIN. In Interview #2, P15 said they had figured out how to add a PIN to their watch, and after doing so they proceeded with the setup of Samsung Pay.

3.4 Limitations

To protect external validity, it was important that participants understood that they were not required to set up Pay, use Pay, form an implementation intention, or follow their implementation intention. We iterated on the design of our interview protocol until we arrived at language which we thought communicated this clearly to participants. However, although setting up Pay was not required to receive compensation for Interview #1 and Survey #2, participants who never set up Pay were not invited to Interview #2 or Survey #3. Although we tried to disguise the qualification criteria for Interview #2 and Survey #3 from participants, participants may have inferred that some action on their part would be required to qualify, and some asked us directly in the interview. To ensure this was not a threat to validity in our controlled experiment, we emphasized that participants' compensation would not be affected by their use or non-use of Pay.

The generalizability of our findings might be impaired by our relatively small sample size ($n = 20$) and recruitment from the geographic area around our institution. To mitigate this, we used purposive sampling to recruit a diverse set of participants. Further, we recruited a much larger set of participants in our controlled experiment (§ 4).

4 Controlled Experiment

The primary goal of the second part of our study was to determine whether participants presented with a PMT-only nudge and a PMT with implementation intention nudge would be more likely to use mobile payments than those who were not presented with these nudges. Thus, we designed and conducted a randomized controlled experiment with a sufficient number of participants ($n = 411$) to determine statistical significance. Our experimental design was influenced by the results of our qualitative interviews. In particular, over the course of our interviews we iterated on the design of our nudges and we compiled a list of common questions and misconceptions which we sought to correct in our experiment. For ease of recruitment and to reduce the complexity of our protocol, we choose to focus on Apple Pay.

4.1 Protocol

Our design included three experimental conditions. In our control group, we did not try to motivate participants to use Apple Pay. In our PMT group, we presented participants with information about the threat of card fraud (Figure 9) and the mitigation of using Apple Pay (Figure 10 and Figure 11) in order to motivate them to use Apple Pay. This motivational intervention was based on protection motivation theory [43], as described in § 3.1. In our PMT+II group, we presented participants with the motivational intervention of the PMT group in addition to an opportunity to form an implementation intention. This opportunity took the form of a template we designed to help participants plan where they could use Apple Pay, as shown in Figure 12 in the appendix. We did not test an implementation intention intervention without a PMT intervention because the literature suggests that implementation intentions are only effective when participants are motivated [19].

Our study consisted of three surveys hosted on Qualtrics using recruitment from Prolific (see Figure 8 and § 8.6–8.8 in the appendix). Survey #1 was designed to determine eligibility for Survey #2 and Survey #3. The only requirements

for taking Survey #1 were that participants live in the United States, speak English, be at least 18 years old, and have an iPhone. We thought our nudges would have the largest impact on people who were not actively using Apple Pay, but whose phones were compatible with Apple Pay and who were likely to have opportunities to use Apple Pay in the week ahead. Thus, to be eligible for participation in Survey #2 and #3, participants must have purchased their iPhone in the United States, owned an iPhone model compatible with Apple Pay (iPhone 6 or newer), must have had a version of iOS compatible with Apple Pay (iOS 12.2 or higher), in the last week must have made an in-person payment in a physical location using their credit or debit card, in the last week must not have made an in-person payment in a physical location using Apple Pay, and they must have passed a simple attention check.

Shortly after completing Survey #1, participants were invited to Survey #2, which contained our randomly assigned experimental conditions. The control group saw only a short description of Apple Pay. The PMT group was provided with a description of the threat of credit and debit card information theft and fraud, and information about the mitigation of using Apple Pay. This information included written instructions about how to set up and use Apple Pay, a short video showing how to use Apple Pay, and an FAQ addressing questions participants asked in our qualitative interviews. We encouraged participants to set up Apple Pay if they wanted to, but we reassured participants that their compensation would not be affected if they did not set it up. The PMT+II group received the same information as the PMT group, but was also given a chance to form a plan to use Apple Pay. Near the end of the survey, participants in the treatment groups were given links to the information about Apple Pay and their plan for using Apple Pay, with the option to request that these links be sent to them via Prolific. Participants in all treatment groups were asked demographic questions and questions related to their perceptions of Apple Pay and card fraud.

Survey #3 was sent to participants one week after they completed Survey #2 in order to measure whether they had used Apple Pay. We asked participants whether they had registered a card in Apple Pay, whether they had made an in-person payment using Apple Pay, and about other details related to their use of Apple Pay and other payment methods.

Our goal was to pay participants $12/hour, so compensation was determined based on estimated duration of our surveys. Survey #1 was estimated to take five minutes, so compensation was $1. Survey #2 was estimated to take up to 30 minutes (accounting for time potentially spent outside the survey setting up Apple Pay), so compensation was $6. Survey #3 was estimated to take five minutes, so compensation was $1. Participants only received compensation for Survey #2 and Survey #3 if they completed both surveys within three days of being invited.

We conducted an a priori power analysis using G*Power to determine our target number of participants [16]. We planned three chi-square tests of independence to compare the use of Apple Pay between the three treatment groups. In order to detect a small to medium effect size ($w = 0.2357$, informed by the effect size seen in our interviews), with a Bonferroni corrected $\alpha = 0.05 \div 3 = 0.01667$, power=0.9, and df=1, we determined that we needed 122 participants in each treatment.

We preregistered our protocol on The Open Science Framework prior to collecting any data (§ 7).

Treatments	% that used Pay	Odds Ratio	p-value
Control vs PMT+II	8.7% vs 27.2%	3.92	<0.001
Control vs PMT	8.7% vs 18.3%	2.35	0.020
PMT vs PMT+II	18.3% vs 27.2%	1.67	0.085

Table 1: Comparisons between the percent of participants who reported using Apple Pay in each of our treatment conditions. Per convention, the reported odds ratios correspond to large, medium, and small effect sizes, respectively [66].

4.2 Analysis

We collected 670 valid responses to Survey #1, and invited 430 qualifying participants to participate in Survey #2. Of the 430 participants invited to Survey #2, 418 completed Survey #2 and 411 went on to complete Survey #3, for an overall dropout rate of 4%.

After completing data collection, we conducted our preregistered hypothesis tests to compare use of Apple Pay between our three treatment conditions, and we report these findings in § 4.3. Next, we conducted a series of exploratory analyses, which we report in § 4.4.

4.3 Results

We conducted three chi-square tests of independence to compare the use of Apple Pay between our three treatment groups, as shown in Table 1. We used the Holm-Bonferroni method to control Type I error.[2] Participants in the PMT+II group, who saw our PMT with implementation intention nudge, were 3.92x more likely to use Apple Pay than our control group ($p < 0.001$). Participants in the PMT group, who saw only the PMT nudge, were 2.35x more likely to use Apple Pay than our control group ($p = 0.020$). Both of these differences were statistically significant at $\alpha = 0.05$. However, we did not find a statistically significant difference in use of Apple Pay between the PMT and PMT+II groups ($p = 0.085$).

Therefore, we have evidence that our interventions in both the PMT+II and PMT groups had large and medium effects

[2] In our preregistration, we described using a Bonferroni correction. We switched to the Holm-Bonferroni method because it controls the experiment's Type I error rate at the same level as a Bonferroni correction, while having a lower Type II error rate [1, 73]. Using a simple Bonferroni correction, only our Control vs PMT+II comparison would have been found significant. See [11] for further discussion of the Bonferroni correction.

on participants' use of Apple Pay, respectively. Since the treatment conditions only differed in their inclusion of our educational materials (Figures 9, 10, and 11) and our implementation intention template (Figure 12), we can conclude that these differences are what made participants more likely to report using Apple Pay. Although we did not find statistically significant differences between the PMT and PMT+II groups, our findings suggest ($p = 0.085$) that the inclusion of the implementation intention plan had a small effect on increasing the PMT+II participants' use of Apple Pay.

4.4 Results of Exploratory Analyses

Although our primary research questions were about the effect of our nudges on participants' use of Apple Pay, the data we collected gave us the opportunity to explore additional questions. Note that these exploratory analyses were not part of our preregistered study design. In this section, we describe the effect of our nudges on participants' attitudes, how expressed intention differed from reported behavior, exactly when participants reported setting up Apple Pay, and additional factors associated with use of Apple Pay.

Effects of Interventions on Attitudes

After testing for the effect of our interventions on participants' use of Apple Pay (§ 4.3), we decided to test for other potential effects, as shown in Table 2. We used Kruskal-Wallis tests for all variables except whether participants registered a card, where we used a chi-square test of independence. Details of the statistically significant results are shown in Figures 1, 2, and 3. Effect sizes are given as epsilon-squared (ε^2) estimates [38, 68]. Insignificant results are included in Figures 13–17 in the appendix. Post-hoc Dunn tests significant at $\alpha = 0.05$ after Holm-Bonferroni correction are bolded.

As shown in Figure 1, our treatments had a dramatic effect on participants' agreement that Apple Pay would protect them from card fraud ($\varepsilon^2 = 0.241$, $p < 0.001$). In the control group, only 37% of participants agreed that Apple Pay would protect them, whereas in both treatment groups over 84% agreed. Thus, we have strong evidence that our information was effective at correcting people's misconceptions about Apple Pay's security [22]. As illustrated in Figure 2, our treatments increased participants' expressed intentions to use Apple Pay, and implementation intentions were even more effective at increasing intention than PMT alone ($\varepsilon^2 = 0.172$, $p < 0.001$). Finally, Figure 3 shows that our treatments had a small effect on participants' belief that Apple Pay would be useful for making payments ($\varepsilon^2 = 0.015$, $p = 0.047$).

Intention vs Behavior

Comparing participants' Survey #2 responses to their Survey #3 responses gave us insight into how participants' stated intentions to act did or did not translate to actual behavior.

Variable	p-value
Perception of threat severity	0.932
Perception of threat susceptibility	0.881
Perception of self-efficacy	0.523
Perception of response-efficacy	**<0.001**
Expressed intention to use Pay	**<0.001**
Perception of Pay's usefulness	**0.047**
Self-consciousness using Pay	0.628
Registered card by end of study	0.237

Table 2: The results of hypothesis tests measuring whether these variables differed between our treatment groups. p-values significant at $\alpha = 0.05$ are bolded, representing tests where the null hypothesis was rejected.

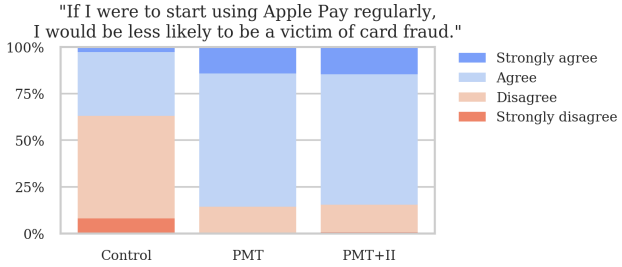


Figure 1: Participants in our treatment groups expressed greater agreement that Apple Pay would protect them from card fraud (i.e., response efficacy). Post-hoc tests: **Control vs PMT**, $\mathbf{p < 0.001}$; **Control vs PMT+II**, $\mathbf{p < 0.001}$; PMT vs PMT+II, $p = 0.880$.

First, we measured how stated intention to register a credit or debit card in Apple Pay translated to actually setting up Apple Pay. As shown in Figure 4, while half of those who expressed a strong intention to register a card did so, those who expressed weaker intentions were correspondingly less likely to register a card. In particular, note that less than half of the participants who responded with "Agree" actually set up Apple Pay by the time of Survey #3.

Next, we compared stated intention to use Apple Pay to actual use of Apple Pay. We performed a chi-square test of independence and found that those who indicated they intended to use Apple Pay in the week ahead were more likely to use Apple Pay than those who did not ($p < 0.001$). However, as shown in Figure 5, many participants who expressed an intention to use Apple Pay did not do so. This reinforces our belief that it is important to ask participants about their actual behavior, rather than only measuring their intentions.

Finally, we took a closer look at the behavior of participants in the PMT+II group, who were given the opportunity to make a plan for using Apple Pay. 96.3% of participants in the PMT+II group wrote plans in Survey #2. Of those who wrote plans, 88.5% visited a location in their plan, 25.2%

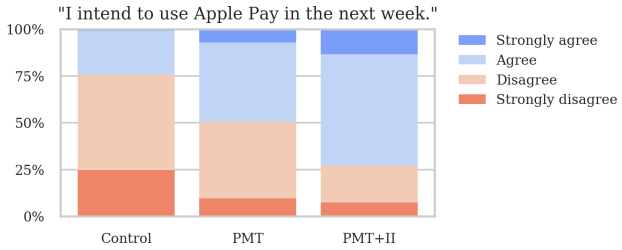


Figure 2: Participants in our treatment groups expressed stronger intentions to use Apple Pay. Further, participants who received the implementation intention treatment expressed even stronger intentions to use Apple Pay than did participants who only received the PMT treatment. Post-hoc tests: **Control vs PMT**, $\mathbf{p < 0.001}$; **Control vs PMT+II**, $\mathbf{p < 0.001}$; **PMT vs PMT+II**, $\mathbf{p = 0.001}$.

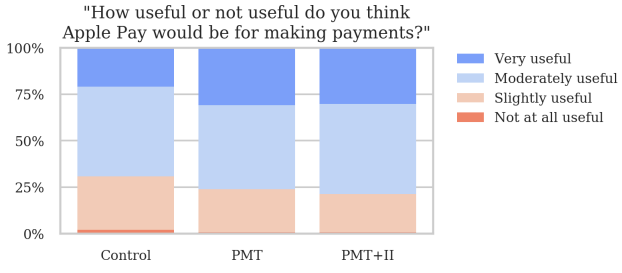


Figure 3: Our treatments had an effect on participants' belief that Apple Pay would be useful for making payments. Post-hoc tests: Control vs PMT, $p = 0.046$; Control vs PMT+II, $p = 0.026$; PMT vs PMT+II, $p = 0.856$.

used Apple Pay at a location in their plan, and 87% used other payment methods at a location in their plan. Of those who wrote plans, 83.2% checked a box indicating "I strongly intend to try to use Apple Pay at these locations!" Of these participants, 89.9% visited a location in their plan, 30.3% used Apple Pay at a location in their plan, and 87.2% used other payment methods at a location in their plan.

In conclusion, although intention to set up and use Apple Pay was associated with actually doing so, many participants who expressed intentions did not follow through. This suggests nudges like implementation intentions may help participants follow through on their intentions. This also demonstrates the importance of measuring actual behavior in addition to intention when evaluating the effectiveness of nudging techniques.

When Did Participants Set Up Apple Pay?

As shown in Figure 6, 35% of participants had set up Apple Pay before Survey #2. In Survey #2, we encouraged the participants in our treatment groups to set up Apple Pay, but

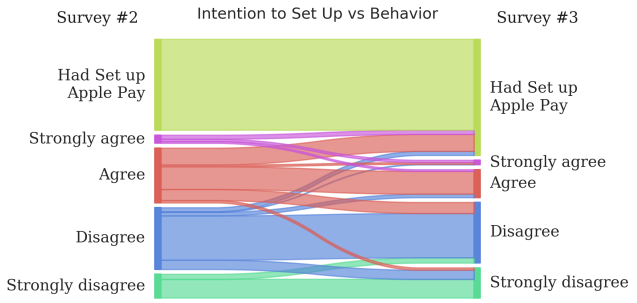


Figure 4: In both Survey #2 and Survey #3, we asked participants whether they had registered a card in Apple Pay. Those who had not were asked to rate their level of disagreement or agreement with the statement: "I intend to register a credit or debit card in Apple Pay in the next week."

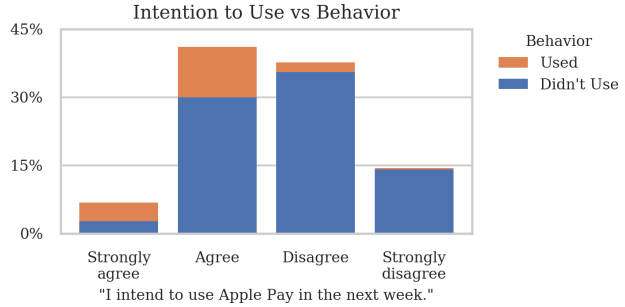


Figure 5: In Survey #2 we asked participants to rate their intention to use Apple Pay in the week ahead. We compared those responses to whether participants reported using Apple Pay in Survey #3.

only 2.9% reported setting it up during Survey #2. However, an additional 10.5% reported setting up Apple Pay when we asked again in Survey #3. Overall, about half of participants had Apple Pay set up by the end of our study.

Note that most of the participants who set up Apple Pay during our study did so after completing Survey #2. The same pattern held in our qualitative interviews (§ 3.3). This suggests the importance of an experimental design like ours, in which information is given to participants, but participants are allowed time to think about that information and potentially conduct additional research before taking action.

Factors Associated with Use of Apple Pay

Having found that our treatments were associated with participants using Apple Pay, we trained three logistic regression models to identify additional factors associated with using Apple Pay.

First, we trained a model on all participants who completed all three of our surveys ($n = 411$). Our model contains the

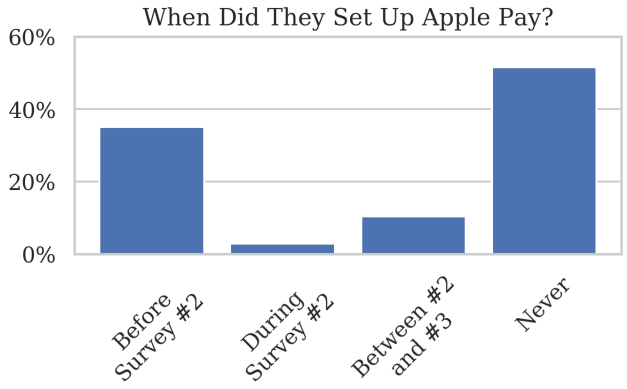


Figure 6: Participants could choose to set up Apple Pay at any point in our study, or not at all. More participants set up Apple Pay after Survey #2 than during Survey #2, suggesting the importance of giving participants time to think about the information we gave them.

following 17 variables: treatment condition, security attitudes (SA-6) [15], age, Computer Science (CS) background, prior experience with card fraud, phone biometric (Face ID or Touch ID), gender, expressed intention to use Apple Pay, whether the participant knew anyone who used Apple Pay, whether the participant owned an Apple Watch, whether the participant had used Apple Pay before the study, and the participants' perceptions of response efficacy, self-efficacy, threat severity, threat susceptibility, self-consciousness, and Apple Pay's usefulness. Our model is shown in Table 6 in the appendix. The model suggests that those with a computer science background and those who have experienced card fraud before are less likely to use Apple Pay (0.24x and 0.45x as likely, respectively). Perhaps those with a computer science background generally know more about Apple Pay, making those eligible for our experiment more likely to have consciously decided not to use it in advance of our interventions. This possibility is supported by Survey #1 from the qualitative interviews showing a positive association between having a CS background and having previously used Pay. The model also suggests that those whose phones are compatible with Face ID (2.1x), those who are non-female (2.4x), those who have used Apple Pay before (3.7x), and those who express an intention to use Apple Pay (6.1x) are more likely to use it.

Next, we trained a model on only the participants in the PMT+II group ($n = 136$). Our model contains the same variables as our first model, with the removal of treatment and the addition of these variables: whether the participant checked the box indicating that they strongly intended to follow their plan, whether the participant requested they be sent information about Apple Pay, whether the participant requested they be sent their plan, and whether the participant visited at least one location in their plan. Our model is shown in Table 7

Variable	β	e^{β}	p-value
age	-0.002	0.998	0.813
Face_ID	0.365	1.441	0.080
own_watch	1.035	2.815	**<0.001**
Intercept	-1.539	0.215	**<0.001**

Table 3: The variables in our regression model for predicting use of Apple Pay in the week prior to Survey #1 ($n = 590$). e^{β} indicates the change in odds of using Apple Pay for a one unit change in the variable (or when the variable is true). p-values significant at $\alpha = 0.05$ are bolded. Cox & Snell $R^2 = 0.051$.

in the appendix. Like our first model, this model suggests that those who experienced card fraud before are less likely to use Apple Pay (0.22x), and that those who used Apple Pay before are more likely to use it again (4.0x). Perhaps counterintuitively, the model also suggests that those who express self-consciousness about using Apple Pay in public are *more likely* to use it (5.1x). It is possible that participants' increased self-consciousness was due to their greater engagement with the plan, which could have made them more likely to use Apple Pay. There is also some evidence that whether the participant visited a location in their plan was associated with using Apple Pay (30x, $p = 0.058$).

Finally, we trained a model on the data we collected in Survey #1 to identify factors associated with people having used Apple Pay in the week before our study. We eliminated participants whose phones were incompatible with Apple Pay and who failed our attention check, leaving us with 590 participants. Due to the limited number of questions we asked participants in Survey #1, our model only contains age, phone compatibility with either Face ID or Touch ID, and Apple Watch ownership. The variables in our model are shown in Table 3. Overall, 23.7% of participants reported using Apple Pay in the past week. Our model shows a strong association between owning an Apple Watch and using Apple Pay, with Apple Watch owners being more than 2.8x more likely to use Apple Pay than non-owners. It is difficult to know the reason for this association, but one possible explanation might be that it's easier to use Apple Pay with an Apple Watch.

4.5 Limitations

One limitation of our study is our reliance on self-reported data. In particular, it is possible that participants did not accurately report whether they used Apple Pay between taking Survey #2 and #3. To encourage honesty, at the beginning of Survey #2 and Survey #3 we included text which encouraged participants to answer honestly and reassured them that there were no right or wrong answers. We also included attention checks in all our surveys. Fifteen participants (2%) failed our Survey #1 attention check and so were not invited to the subsequent surveys, but no participants failed our Survey #2

or Survey #3 attention checks. Another threat to validity is the possibility that some participants may have thought that setting up or using Apple Pay was not optional. To avoid that misconception, we included text assuring participants that setting up or using Apple Pay was not required and would have no effect on their compensation. One threat to the generalizability of our findings is the fact that crowd workers have been shown to differ from the general population. Our use of Prolific was informed by recent findings that Prolific workers are more diverse and honest than Mechanical Turk workers [49]. See Table 5 in the appendix for a summary of demographic information about our participants.

5 Discussion and Future Work

Our results have implications for both practitioners and researchers. First, banks, card issuers, and mobile payment operators could use our nudges to increase use of mobile payments instead of less secure, physical card payments. More widespread adoption of secure mobile payments has the potential to reduce card fraud, saving companies and customers both time and money. Second, our findings advance the field of nudging research. Our PMT and implementation intention nudges corrected participants' misconceptions and increased intention to and actual use of mobile payments. In particular, we believe our PMT-inspired description of Apple Pay's security (Figure 10) was instrumental in correcting participants' misconception that mobile payments are less secure than physical card payments. Our implementation intention nudge was designed to help participants remember to use mobile payments when they visited certain locations. Although we did not find sufficient evidence to conclude that our implementation intention nudge increased *use of Apple Pay* compared to the PMT nudge (Table 1), we did find strong evidence that it increased *intention to use Apple Pay* (Figure 2). This discrepancy is due to the fact that many participants who expressed an intention to use Apple Pay did not actually use it (Figure 5). This shows the importance of an experimental design which measures both intention to use and actual behavior, as we did in our study. Our results also show the need for additional research into techniques that may help people translate their intentions to act in a secure manner into actual behavior.

Our study suggests several possible areas for future work. First, it would be useful to compare our PMT and implementation intention nudges in an experiment with a larger sample size. This would allow us to conclusively determine whether implementation intentions yield improvements over PMT alone. Second, testing variations of PMT and implementation intention nudges could yield insight into what exactly makes these nudges effective. Knowing the most essential elements of these nudges could help translate them into a form suitable for large-scale messaging campaigns. Relatedly, people's receptivity to such messaging campaigns may depend on the entities conducting the campaigns, making a study of such messenger effects worthwhile. Third, PMT and implementation intentions should be tested for their potential to increase adoption of other secure technologies and for encouraging adherence to security best practices.

6 Conclusions

Despite the security benefits they offer, adoption of mobile payments in the United States remains low, at least in part due to the belief that mobile payments are less secure than payments with physical cards [22, 50]. Our nudges addressed this misconception and increased adoption of mobile payments: participants in our PMT and PMT with implementation intention treatment groups were 2.4x and 3.9x more likely, respectively, to use Apple Pay than those in our control group. Our qualitative interviews suggested additional factors which may limit adoption of mobile payments, including lack of availability and usability challenges.

Our findings show that it is possible to increase real-world adoption of security-enhancing technologies using carefully crafted informational interventions. At the same time, many people who express an intention to adopt such technologies may fail to do so. This suggests the need for further research into interventions which can help people translate intention into action. Implementation intentions are designed to do this. In our study, we found only weak evidence of a small improvement (1.67x) from adding implementation intentions to our PMT intervention. However, implementation intentions might become more helpful as mobile payments become more available and other barriers to adoption are removed. Clearly, there is no single solution for increasing adoption of security-enhancing technologies, but PMT and implementation intention nudges are two tools that may help.

7 Preregistration and Materials

We preregistered our controlled experiment on the Open Science Framework [64]. After preregistering but before collecting any data, we made two small edits to the survey text. Also, before collecting any Survey #3 data, we added a "using another payment method" option to Q18, Q19, and Q20 in Survey #3. In our preregistration, we described using a Bonferroni correction, but switched to the Holm-Bonferroni method as it controls the experiment's Type I error rate at the same level as a Bonferroni correction, while having a lower Type II error rate [1, 73]. Otherwise, we conducted our study as preregistered. To view the final version of all study materials, see our study page [65].

Acknowledgments

We thank Yuanyuan Feng, Hana Habib, Linda Moreci, Yaxing Yao, and Yixin Zou for their assistance with our work. This

research was supported in part by grants from the National Science Foundation's Secure and Trustworthy Computing program (CNS-1330596, SES-1513957, CNS-1801316). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the NSF or the U.S. Government.

References

[1] Hervé Abdi. Holm's Sequential Bonferroni Procedure. *Encyclopedia of research design*, pages 1–8, 2010.

[2] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, January 2015.

[3] Alessandro Acquisti and Jens Grossklags. Privacy and Rationality in Individual Decision Making. *IEEE Security & Privacy*, 3(1):26–33, 2005.

[4] Alessandro Acquisti, Manya Sleeper, Yang Wang, Shomir Wilson, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, and Florian Schaub. Nudges for Privacy and Security. *ACM Computing Surveys*, 50(3):1–41, August 2017.

[5] Elham Al Qahtani, Mohamed Shehab, and Abrar Aljohani. The Effectiveness of Fear Appeals in Increasing Smartphone Locking Behavior among Saudi Arabians. *SOUPS @ USENIX Security Symposium*, 2018.

[6] Yusuf Albayram, Mohammad Maifi Hasan Khan, Theodore Jensen, and Nhan Nguyen. "...better to use a lock screen than to worry about saving a few seconds of time" - Effect of Fear Appeal in the Context of Smartphone Locking Behavior. *Symposium on Usable Privacy and Security*, 2017.

[7] Hazim Almuhimedi. Helping Smartphone Users Manage their Privacy through Nudges. Technical Report CMU-ISR-17-111, December 2017.

[8] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Cranor, and Yuvraj Agarwal. Your location has been shared 5,398 times!: A field study on mobile app privacy nudging. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015.

[9] Apple. Apple Pay coming to Target, Taco Bell and more top US retail locations, Jan 2019. https://www.apple.com/newsroom/2019/01/apple-pay-coming-to-target-taco-bell-and-more-top-us-retail-locations.

[10] Apple. Apple Pay security and privacy overview, Oct 2019. https://support.apple.com/en-us/HT203027.

[11] Richard A Armstrong. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, 34(5):502–508, April 2014.

[12] Pamela Briggs, Debbie Jeske, and Lynne Coventry. Behavior change interventions for cybersecurity. In *Behavior Change Research and Theory*, pages 115–136. Elsevier, 2017.

[13] Kevin Cash. Credit card vs. debit card: Which is safer online?, Sep 2015. https://www.nerdwallet.com/blog/credit-cards/credit-card-vs-debit-card-safer-online-purchases/.

[14] Michael Fagan and Mohammad Maifi Hasan Khan. Why Do They Do What They Do? - A Study of What Motivates Users to (Not) Follow Computer Security Advice. *Symposium on Usable Privacy and Security*, 2016.

[15] Cori Faklaris, Laura A Dabbish, and Jason I Hong. A Self-Report Measure of End-User Security Attitudes (SA-6). *SOUPS @ USENIX Security Symposium*, 2019.

[16] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, May 2007.

[17] Alisa Frik, Nathan Malkin, Marian Harbach, Eyal Peer, and Serge Egelman. A Promise Is A Promise. In *the 2019 CHI Conference*, pages 1–12, New York, New York, USA, 2019. ACM Press.

[18] Vindu Goel and Rachel Abrams. Card Data Stolen From 5 Million Saks and Lord & Taylor Customers, Apr 2018. https://www.nytimes.com/2018/04/01/technology/saks-lord-taylor-credit-cards.html.

[19] Peter M Gollwitzer. Implementation intentions: strong effects of simple plans. *American Psychologist*, 54(7), 1999.

[20] Google. Google Pay may be preinstalled on your phone, May 2020. https://support.google.com/pay/answer/7644010.

[21] Google. How payments work, Jan 2020. https://support.google.com/pay/merchants/answer/6345242?hl=en.

[22] Jun Ho Huh, Saurabh Verma, Swathi Sri V Rayala, Rakesh B Bobba, Konstantin Beznosov, and Hyoungshick Kim. I Don't Use Apple Pay because it's less secure...: perception of security and usability in mobile tap-and-pay. *Proceedings of the Workshop on Usable Security*, 2017.

[23] Joseph Keller. How to stop Apple Pay from pestering you into signing up, Apr 2018. https://www.imore.com/how-stop-apple-pay-pestering-you.

[24] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. Standardizing privacy notices - an online study of the nutrition label approach. *CHI*, page 1573, 2010.

[25] Brian Krebs. Chip Card ATM 'Shimmer' Found in Mexico, Aug 2015. https://krebsonsecurity.com/2015/08/chip-card-atm-shimmer-found-in-mexico/.

[26] Brian Krebs. All About Skimmers, Jun 2016. https://krebsonsecurity.com/all-about-skimmers/.

[27] Brian Krebs. Comment on Credit Card Breach at Buckle Stores, Jun 2017. https://krebsonsecurity.com/2017/06/credit-card-breach-at-buckle-stores/#comment-433940.

[28] Brian Krebs. Data: E-Retail Hacks More Lucrative Than Ever, Apr 2019. https://krebsonsecurity.com/2019/04/data-e-retail-hacks-more-lucrative-than-ever/.

[29] Brian Krebs. Sale of 4 Million Stolen Cards Tied to Breaches at 4 Restaurant Chains, Nov 2019. https://krebsonsecurity.com/2019/11/sale-of-4-million-stolen-cards-tied-to-breaches-at-4-restaurant-chains/.

[30] Brian Krebs. Wawa Breach May Have Compromised More Than 30 Million Payment Cards, Jan 2020. https://krebsonsecurity.com/2020/01/wawa-breach-may-have-compromised-more-than-30-million-payment-cards/.

[31] Linda Lee, David Fifield, Nathan Malkin, Ganesh Iyer, Serge Egelman, and David Wagner. A Usability Evaluation of Tor Launcher. *Proceedings on Privacy Enhancing Technologies*, 2017(3):257–20, June 2017.

[32] Howard Leventhal, Robert Singer, and Susan Jones. Effects of fear and specificity of recommendation upon attitudes and behavior. *American Psychologist*, 2(1):20–29, 1965.

[33] Gen-Yih Liao and Chia-Min Wang. Exploring the Influences of Implementation Intention on Information Security Behaviors. *AMCIS*, 2011.

[34] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun Zhang, Norman Sadeh, Alessandro Acquisti, and Yuvraj Agarwal. Follow My Recommendations - A Personalized Privacy Assistant for Mobile App Permissions. *Symposium on Usable Privacy and Security*, 2016.

[35] Ben Luthi. Is Apple Pay Safe?, Apr 2019. https://creditcards.usnews.com/articles/is-apple-pay-safe.

[36] Mary Madden and L Rainie. Americans' attitudes about privacy, security and surveillance. Pew Research Center, May 2015.

[37] James E Maddux and Ronald W Rogers. Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change. *Journal of experimental social psychology*, 19(5):469–479, 1983.

[38] Salvatore S. Mangiafico. Kruskal–wallis test, Feb 2020. https://rcompanion.org/handbook/F_08.html.

[39] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM Human-Computer Interaction*, pages 1–23, August 2019.

[40] Katherine L Milkman, John Beshears, James J Choi, David Laibson, and Brigitte C Madrian. Using implementation intentions prompts to enhance influenza vaccination rates. *Proceedings of the National Academy of Sciences*, 108(26), 2011.

[41] Katherine L Milkman, John Beshears, James J Choi, David Laibson, and Brigitte C Madrian. Planning prompts as a means of increasing preventive screening rates. *Preventive Medicine*, 56(1):92–93, January 2013.

[42] Sarah Milne, Sheina Orbell, and Paschal Sheeran. Combining motivational and volitional interventions to promote exercise participation: Protection motivation theory and implementation intentions. *British Journal of Health Psychology*, 7(2):163–184, May 2002.

[43] Sarah Milne, Paschal Sheeran, and Sheina Orbell. Prediction and Intervention in Health-Related Behavior: A Meta-Analytic Review of Protection Motivation Theory. *Journal of Applied Social Psychology*, 30(1):106–143, January 2000.

[44] David W Nickerson and Todd Rogers. Do You Have a Voting Plan? *Psychological Science*, 21(2):194–199, January 2010.

[45] Patricia A Norberg, Daniel R Horne, and David A Horne. The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors. *Journal of Consumer Affairs*, 41(1):100–126, June 2007.

[46] Sheina Orbell, Sarah Hodgkins, and Paschal Sheeran. Implementation intentions and the theory of planned behavior. *Personality and Social Psychology Review*, 23(9):945–954, 1997.

[47] Sheina Orbell and Paschal Sheeran. Motivational and Volitional Processes in Action Initiation: A Field Study of the Role of Implementation Intentions1. *Journal of Applied Social Psychology*, 30(4):780–797, 2000.

[48] Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. Let's Go in for a Closer Look. In *the 2017 ACM SIGSAC Conference*, pages 295–310, New York, New York, USA, 2017. ACM Press.

[49] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70(C):153–163, May 2017.

[50] Pew. Are Americans Embracing Mobile Payments? pages 1–22, October 2019.

[51] PYMNTS.com. Apple Pay Adoption Stats, Jan 2020. https://www.pymnts.com/apple-pay-adoption/.

[52] Steve Ragan. What you need to know about the Home Depot data breach, Sep 2014. https://www.csoonline.com/article/2604320/what-you-need-to-know-about-the-home-depot-data-breach.html.

[53] Lee Rainie. Americans' complicated feelings about social media in an era of privacy concerns. *Pew Research Center*, March 2018.

[54] Ronald W Rogers. A Protection Motivation Theory of Fear Appeals and Attitude Change. *The Journal of Psychology*, 91(1):93–114, 1975.

[55] Ronald W Rogers and Steven Prentice-Dunn. Protection motivation theory. 1997.

[56] Todd Rogers, Katherine L Milkman, Leslie K John, and Michael I Norton. Beyond good intentions: Prompting people to make plans improves follow-through on important tasks. *Behavioral Science & Policy*, 1(2):33–41, 2015.

[57] Kate Rooney. Mobile payments have barely caught on in the US, despite the rise of smartphones, Aug 2019. https://www.cnbc.com/2019/08/29/why-mobile-payments-have-barely-caught-on-in-the-us.html.

[58] Samsung. How secure is Samsung Pay?, Jan 2020. https://www.samsung.com/us/support/answer/ANS00043932/.

[59] Samsung. Set up Samsung Pay on your phone, May 2020. https://www.samsung.com/us/support/answer/ANS00045081/.

[60] Ruth Shillair, Shelia R Cotten, Hsin-Yi Sandy Tsai, Saleem Alhabash, Robert LaRose, and Nora J Rifon. Online safety begins with you and me: Convincing Internet users to protect themselves. *Computers in Human Behavior*, 48(C):199–207, July 2015.

[61] Mikko Siponen, M Adam Mahmood, and Seppo Pahnila. Employees' adherence to information security policies: An exploratory field study. *Information & Management*, 51(2):217–224, March 2014.

[62] Teodor Sommestad, Henrik Karlzén, and Jonas Hallberg. A Meta-Analysis of Studies on Protection Motivation Theory and Information Security Behavior. *Dewald Roode Information Security Research Workshop*, pages 1–32, May 2014.

[63] Birgit Steller. *Vorsätze und die Wahrnehmung günstiger Gelegenheiten [Implementation intentions and the detection of good opportunities to act]*. tuduv-Verlag-Ges., 1992.

[64] Peter Story and Daniel Smullen. Apple services study: Preregistration, Dec 2019. https://osf.io/k3nrd.

[65] Peter Story and Daniel Smullen. Apple services study: Study page, Dec 2019. https://osf.io/srwyk.

[66] Gail M Sullivan and Richard Feinn. Using Effect Size—or Why the PValue Is Not Enough. *Journal of Graduate Medical Education*, 4(3):279–282, September 2012.

[67] Richard H Thaler and Cass R Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. J. Wiley and Sons, 2008.

[68] Maciej Tomczak and Ewa Tomczak. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. pages 1–7, July 2014.

[69] Jordi van den Breekel, Diego A Ortiz-Yepes, Erik Poll, and Joeri de Ruiter. EMV in a nutshell. Technical report, June 2016.

[70] Gregory Wallace. Target credit card hack: What you need to know, Dec 2013. https://money.cnn.com/2013/12/22/news/companies/target-credit-card-hack/index.html.

[71] Rick Wash and Emilee J Rader. Influencing mental models of security - a research agenda. *NSPW*, page 57, 2011.

[72] Alma Whitten and J D Tygar. Why Johnny can't encrypt: a usability case study of PGP 5. *USENIX Security Symposium*, August 1999.

[73] Wikipedia. Holm–bonferroni method, Nov 2019. https://en.wikipedia.org/wiki/Holm-Bonferroni_method.

[74] Kim Witte and Mike Allen. A meta-analysis of fear appeals: Implications for effective public health campaigns. *Personality and Social Psychology Review*, 27(5):591–615, 2000.

8 Appendix

Table 4: Final Codebook With Code Frequencies

Code	Description	Interview #1 (n=20)	Interview #2 (n=10)	Survey #2 (n=20)	Survey #3 (n=10)	Overall (n=20)
accidental_activation	Accidentally activating Pay (e.g., by double-tapping the home button, proximity of NFC devices, etc.).	1	3	1	0	4
additional_research	Performing additional research about Pay (e.g., asking others for their opinion about it, doing Google searches, etc.).	5	3	1	0	7
curiosity_availability	Wondering which places will accept Pay.	10	0	0	0	10
curiosity_information_theft	Wondering how their card information was stolen or could be stolen, how fraud occurred, why a data breach occurred, etc.	9	1	0	0	9
curiosity_reviewing_ transactions	Wondering whether they will still be able to review their past transactions if they start using Pay.	1	0	0	0	1
curiosity_technology	Wondering about specific technologies behind Pay (e.g., how NFC works, how the cryptography works, etc.), what cards can be added, how to activate it, how it works, its business model, etc.	16	4	3	1	16
experience_card_fraud	People's own (or others') experiences with card fraud. Any fraudulent purchase made to a card is card fraud.	19	1	0	0	19
experience_card_information_ theft	People's own (or others') experiences with card info theft.	10	0	0	0	10
experience_card_theft	People's own (or others') experiences with their physical card being stolen.	5	0	0	0	5
experience_no_card_fraud	People having no experiences of their own (or others') to recount about card fraud.	1	0	0	0	1
experience_no_card_ information_theft	People having no experiences of their own (or others') to recount about card info theft.	5	0	0	0	5
experience_other	Security-related experiences that don't fit into the other codes.	2	1	0	0	3
experience_unsure	People saying they are unsure whether their card information has been stolen or whether they have been the victim of fraud.	4	0	0	0	4

Continued on the next page

Code	Description	Interview #1 (n=20)	Interview #2 (n=10)	Survey #2 (n=20)	Survey #3 (n=10)	Overall (n=20)
implementation_intention_ clarified_understanding	Forming the implementation intention clarified the person's understanding of Pay (e.g., realizing it won't work at gas stations).	4	0	0	0	4
implementation_intention_ forgotten	The participant not being able to remember their plan.	0	4	0	0	4
implementation_intention_ helpful	Participants' reasons why the implementation intention would or did help them remember to set up or use Pay.	10	6	1	0	13
implementation_intention_ remembered	The participant remembering their plan.	0	8	1	0	8
implementation_intention_ unhelpful	Participants' reasons why the implementation intention would not help them remember to use Pay, why it is hard to form a plan, etc.	12	7	0	0	15
influenced_positive_self_report	The participant saying that the interview made them more likely to use or set up Pay.	10	1	0	0	10
misconception_affiliation	Thinking that we are working for or being funded by a company behind one of the technologies we're discussing (e.g., are you guys working for Google?).	4	0	0	0	4
misconception_always_resolved	Thinking that fraudulent purchases will always be resolved (e.g., they will always get their money back).	2	0	0	0	2
misconception_cost	Thinking or wondering if Pay costs something to use.	3	0	0	0	3
misconception_opening_app	Thinking that using Pay requires opening the Pay app by tapping on its icon.	1	0	0	0	1
misconception_other	Other misconceptions.	2	2	0	0	4
misconception_required	Thinking that using Pay or following the plan is a required part of the study.	1	0	0	0	1
misconception_rewards	Thinking that they won't get rewards, points, or cash back if they use Pay.	4	0	0	0	4
misconception_screen_scan	Thinking that Pay works by scanning the user's phone or watch screen, rather than by using NFC.	4	0	0	0	4
misconception_square_pos	Thinking that Pay only works at Square POSs, or that Pay is the software running on those Square POS iPads. It is not a misconception that Pay works at most Square POSs.	2	0	0	0	2

Continued on the next page

Code	Description	Interview #1 (n=20)	Interview #2 (n=10)	Survey #2 (n=20)	Survey #3 (n=10)	Overall (n=20)
mitigation_description_helpful	Participants' reasons why the description of Pay or the instructions for how to set up and use Pay are helpful to them.	6	2	1	0	6
protection_action_RFID_wallet	Using an RFID-blocking wallet to protect your card information.	1	0	0	0	1
protection_action_account_access	Protect access to your account (e.g., password, 2FA).	6	2	1	0	8
protection_action_avoid_disclosure	Avoid giving information to others, whether prompted or not; avoiding falling for phishing, etc.	5	0	0	0	5
protection_action_avoid_merchant	Avoid transactions at untrusted merchants, only use trusted merchants, etc.	4	0	0	0	4
protection_action_avoid_online	Avoid making purchases online, avoid putting card information online, etc.	3	0	0	0	3
protection_action_certification_logo	Looking for certification logos (e.g., Trustee, Verisign, McAfee), browser plugin indicators (e.g., Web of Trust), TLS certificates, or any other symbols that attest to security in some way.	4	0	0	0	4
protection_action_corporate_resolution	Reporting fraudulent purchases to the card issuer, getting a new card, etc.	19	0	0	0	19
protection_action_data_retention	Preventing a card from being saved on a website either in whole or in part (e.g., not allowing the CVC to be saved).	2	0	0	0	2
protection_action_law_enforcement	Reporting card fraud or theft to law enforcement.	1	0	0	0	1
protection_action_monitor_statements	Looking for unauthorized transactions on card statements.	8	0	0	0	8
protection_action_monitoring_service	Lifelock, credit monitoring, etc.	4	2	2	1	4
protection_action_network	Using a secure network connection (e.g., home Wi-Fi, a VPN when on public Wi-Fi, etc.), avoiding insecure networks, avoiding public computers, etc.	2	0	0	0	2
protection_action_other	Other actions people take to protect themselves from card info theft and fraud.	7	1	0	0	8
protection_action_physical_awareness	Looking for card skimmers, hiding PIN, putting things in a place so they won't be stolen, paying close attention to what a shopkeeper does, checking receipts, etc.	10	1	0	1	11

Continued on the next page

Code	Description	Interview #1 (n=20)	Interview #2 (n=10)	Survey #2 (n=20)	Survey #3 (n=10)	Overall (n=20)
protection_action_use_cash	Using cash.	6	0	0	0	6
protection_action_use_chip	Using the chip in their card (as opposed to the magnetic stripe).	1	0	0	0	1
protection_action_use_credit	Using a credit card, since getting refunded is easier, etc.	6	1	1	0	7
protection_action_use_debit	Using a debit card or using a debit card as a credit card.	1	0	0	0	1
protection_action_use_other_payment_service	Using PayPal, Venmo, or another payment service other than Pay.	2	1	1	1	4
protection_action_use_pay	Using Apple Pay, Google Pay, or Samsung Pay (coded only when brought up prior to us suggesting that they use Pay).	2	0	0	0	2
reasons_for_not_setting_up	People's reasons for not setting up Pay.	14	0	9	0	16
reasons_for_not_using	People's reasons why they don't want to or did not use Pay.	7	1	4	6	12
response_efficacy_security_convinced	Reasons why participants are convinced that Pay will protect them.	12	4	2	1	14
response_efficacy_security_unconvinced	Reasons why participants think Pay will not protect them.	8	2	1	0	8
response_efficacy_security_unsure	Reasons why participants are unsure whether Pay will protect them.	12	4	0	0	14
self_efficacy_negative_battery	Using Pay requires a charged phone.	1	0	0	0	1
self_efficacy_negative_learning	Using or setting up Pay requires practice, learning, or attention to detail.	7	1	1	0	7
self_efficacy_negative_limited_availability	Not all places accept Pay. It may be unclear whether a given place accepts it.	7	7	2	0	12
self_efficacy_negative_limited_card_compatibility	Not all cards can be added to Pay.	1	2	0	0	3
self_efficacy_negative_opportunities	Not going shopping, not having any money, etc., and so not having opportunities to use Pay.	4	4	5	3	11
self_efficacy_negative_other	Other challenges to using Pay, negative experiences using it, and things that make using it more difficult.	7	5	2	2	9
self_efficacy_negative_overspending	The convenience of Pay makes the participant more inclined to wastefully or accidentally spend money.	4	0	0	0	4

Continued on the next page

Code	Description	Interview #1 (n=20)	Interview #2 (n=10)	Survey #2 (n=20)	Survey #3 (n=10)	Overall (n=20)
self_efficacy_negative_payment_failure	Pay payments not going through or taking too long/timing out.	2	6	2	2	6
self_efficacy_negative_remembering	Difficulty remembering to use Pay.	5	3	2	1	9
self_efficacy_negative_setup	Difficulty or irritation setting up Pay.	9	2	6	0	14
self_efficacy_negative_time	It taking too long or a long time to use Pay.	1	4	1	2	6
self_efficacy_other	Other comments about Pay usability, that are neither positive nor negative.	2	2	0	0	4
self_efficacy_positive_easy_to_use	Fast, simple, convenient, etc. to make transactions.	14	6	4	1	15
self_efficacy_positive_extensive_availability	Many or enough places accept Pay.	4	1	0	0	4
self_efficacy_positive_initiative	People taking the initiative to determine whether Pay is accepted (e.g., asking if Pay is accepted, or attempting to use it if they're unsure). Not coded if people said they didn't take the initiative.	0	1	0	0	1
self_efficacy_positive_no_wallet	If you use Pay, you won't have to carry your wallet, carry your cards, or pull out your cards or wallet.	7	5	2	3	10
self_efficacy_positive_novelty	Using or setting up Pay due to curiosity, wanting to see if it works.	9	2	5	1	10
self_efficacy_positive_only_option	Being more likely to use or using Pay because it's an option if you forget another payment method, another payment method doesn't work, you don't have your cards with you, etc. Also includes making Pay more accessible than cards (e.g., by burying cards in your purse and leaving phone on top).	7	5	1	0	10
self_efficacy_positive_opportunities	Going shopping, etc., and so having opportunities to use Pay. Includes inferred opportunities (e.g., if someone says they used Pay, that implies they had opportunities).	15	10	5	3	18

Continued on the next page

Code	Description	Interview #1 (n=20)	Interview #2 (n=10)	Survey #2 (n=20)	Survey #3 (n=10)	Overall (n=20)
self_efficacy_positive_other	Other non-security perks to using Pay, positive experiences using it, good things about Pay, etc.	4	1	0	0	5
self_efficacy_positive_other_reminders	Other things reminding people to use Pay. Not including the setup instructions or implementation intention plan template we offer users. Not including it being the only option.	1	6	1	0	7
self_efficacy_positive_setup	Positive things said about the setup process (easy, etc.).	14	4	1	0	17
self_efficacy_practice	Wanting to practice (or actually practicing) using Pay in a low-pressure situation (e.g., a vending machine, a self-checkout, etc.).	0	1	1	0	2
threat_severity_card_type	The severity of fraud would depend on what type of card was affected by the fraud (e.g., fraud on credit vs debit card).	2	0	0	0	2
threat_severity_high_concern_gets_worse	When fraud or information theft occurs, this might be a precursor to something worse (e.g., a worse hack, more lost money, etc.).	10	0	0	0	10
threat_severity_high_concern_hassle	Resolving the situation would be time-consuming, stressful, irritating, etc.	10	0	0	0	10
threat_severity_high_concern_lost_money	Being concerned about losing money, either from purchases not being refunded, or not being refunded for overdraft or other fees.	6	0	0	0	6
threat_severity_high_concern_other	Other reasons why people perceive the severity to be higher.	7	0	0	0	7
threat_severity_high_concern_violation	People feel violated, helpless, angry, etc. when they suffer from card fraud or information theft.	3	0	0	0	3
threat_severity_low_concern_other	Other reasons why people perceive the severity to be lower.	2	0	0	0	2
threat_severity_low_concern_resolution	It would be possible to resolve the situation.	11	0	0	0	11
threat_severity_other	Other things that impact perceptions of threat severity.	1	0	0	0	1
threat_severity_purchase_size	The severity of fraud would depend on the size of the fraudulent purchase which was made.	6	0	0	0	6
threat_susceptibility_comparison	Participants comparing the relative likelihood of one type of card (information) theft/fraud to another type of event. For example, it being more likely for debit information to be stolen than credit information.	11	0	0	0	11

Continued on the next page

Code	Description	Interview #1 (n=20)	Interview #2 (n=10)	Survey #2 (n=20)	Survey #3 (n=10)	Overall (n=20)
threat_susceptibility_high_ likelihood	Reasons participants perceive the likelihood of encountering the threat to be higher.	12	0	0	0	12
threat_susceptibility_low_ likelihood	Reasons participants perceive the likelihood of encountering the threat to be lower or non-existent (e.g., it's never happened to me before, it's never going to happen, etc.).	9	0	0	0	9
threat_susceptibility_other	Other things that impact perceptions of threat susceptibility. Also includes participants expressing that they are unsure about their threat susceptibility.	2	0	0	0	2

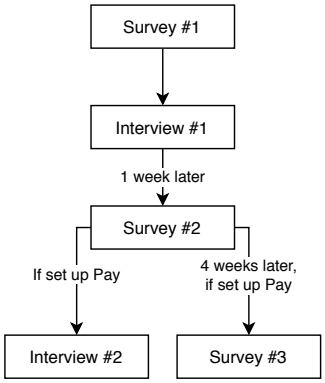


Figure 7: The process of administering surveys and interviews in the qualitative portion of our study (§ 3).

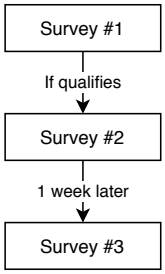


Figure 8: Our controlled experiment contained three online surveys (§ 4).

Demographic	**Values**	
Age	Minimum	18
	Median	32
	Mean	34.7
	Maximum	71
Gender	Female	58%
	Male	41%
	Other	1%
Employment	Working	74%
	Student	11%
	Not employed	10%
	Other	6%
Education	High school or less	18%
	College or associate	56%
	Graduate degree	18%
	Professional degree	4%
	Other	3%
Worked or Studied in a Computer-related Field	Yes	25%
	No	75%
Household Income	Median	$60,000 to $79,999

Table 5: Demographics for the 411 participants who completed all parts of our controlled experiment.

There have been many big hacks where credit and debit card information was stolen from retailers. For example, Target [70] was hacked in 2013, Home Depot [52] was hacked in 2014, and Saks Fifth Avenue [18] was hacked in 2018. Information about millions of cards was stolen in these hacks. If criminals get your credit or debit card information, they might use that information to make fraudulent purchases. If you notice fraudulent purchases on your credit card, you can probably get refunded. But if the purchases are made on your debit card, you might not be able to get your money back [13]. In any case, you would need to get a replacement card with a new number, which would be inconvenient.

Figure 9: In our experiment, participants in the PMT and PMT+II groups were shown this text to inform them about the threat of card fraud. This text was included in order to help participants develop accurate perceptions of threat susceptibility and threat severity, two elements of PMT [43].

Thankfully, there are steps you can take to prevent your card information from being stolen and to protect yourself from card fraud. One of the best things you can do is to start using Apple Pay. Instead of paying by swiping or inserting your card, you can make payments through your phone, which adds an extra layer of security. Payments made with Apple Pay will still be charged to your credit or debit card, but because the payments go through Apple Pay, your card number is not shown to or recorded by retailers [10]. This means that your card number cannot be stolen from transactions made with Apple Pay. If your phone is stolen, the thief will not be able to make payments because Apple Pay is protected by your fingerprint and lock screen PIN. Although no system is perfectly secure, security experts generally agree that Apple Pay is more secure than paying with credit or debit cards [35]. Apple Pay takes just a few minutes to set up, and is widely accepted. As of this year, Apple Pay is accepted in 65% of retail locations [9] in the United States. For example, ALDI grocery, CVS pharmacy, and Starbucks all accept Apple Pay.

Figure 10: In our experiment, participants in the PMT and PMT+II groups were shown this text to inform them about how using Apple Pay can protect them from card fraud. This text was included in order to help participants understand how Apple Pay may help protect them from card fraud and to give them confidence that they would have opportunities to use Apple Pay, influencing perceptions of response efficacy and self-efficacy, two additional elements of PMT [43].

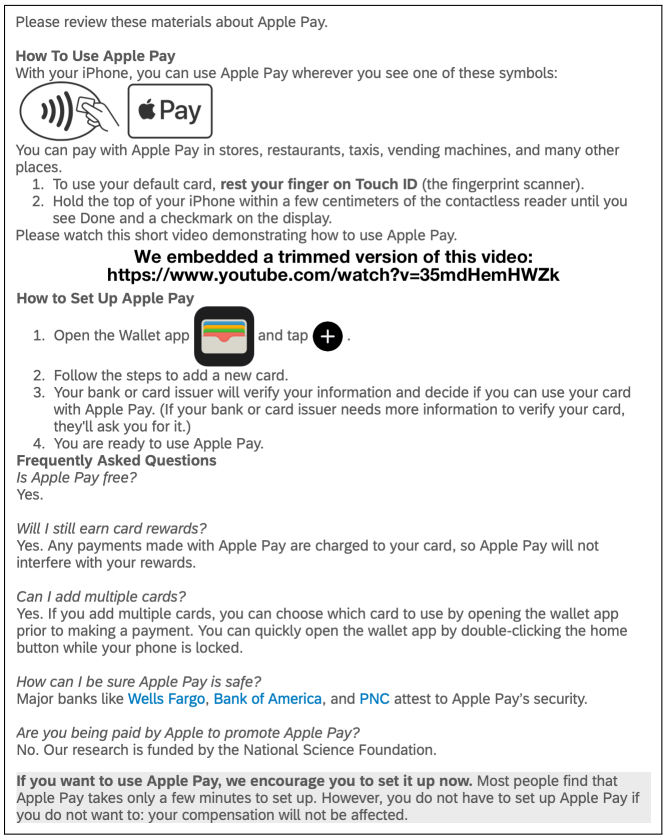


Figure 11: In our experiment, participants in the PMT and PMT+II groups were shown these details about Apple Pay. The instructions contained information about either Touch ID or Face ID, based on which technology the participant's phone was compatible with. These instructions were designed to positively influence perceptions of self-efficacy.

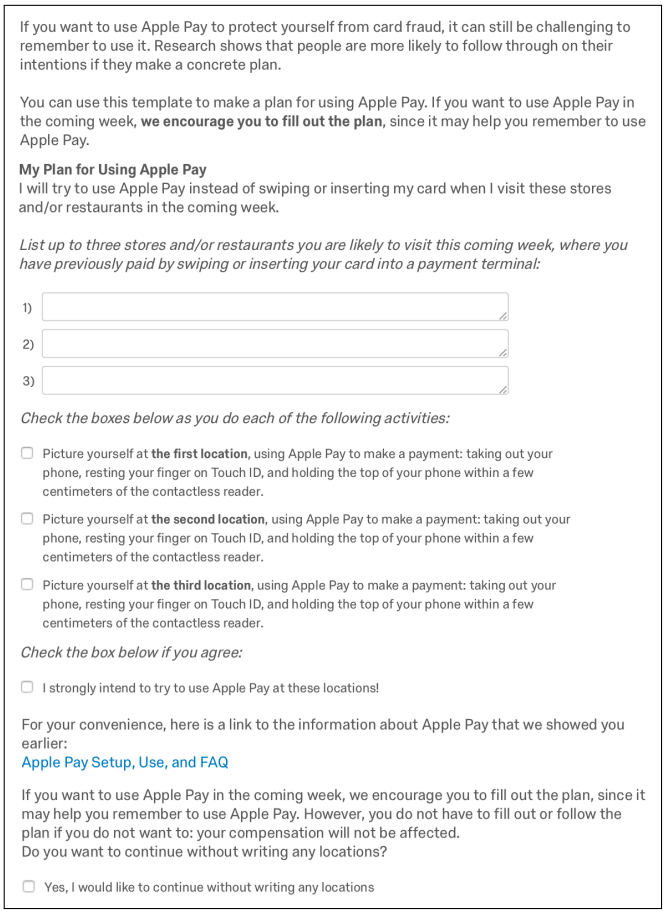


Figure 12: In our experiment, participants in the PMT+II group were shown this implementation intention template. The template encourages participants to plan where they might use Apple Pay in the coming week and to mentally rehearse using Apple Pay in these locations. These activities should help mentally activate participants' plans to use Apple Pay when they are in these locations [19]. Finally, participants are given the opportunity to strongly commit to their plan [19, 63].

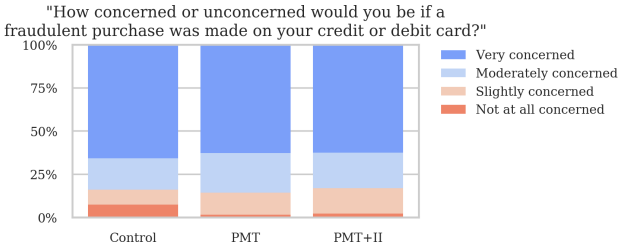


Figure 13: We did not find statistically significant evidence that our treatments affected perception of threat severity ($p = 0.932$).

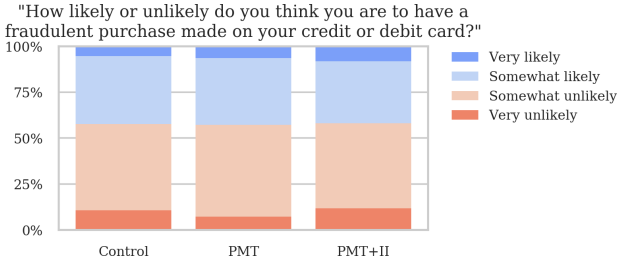


Figure 14: We did not find statistically significant evidence that our treatments affected perception of threat susceptibility ($p = 0.881$).

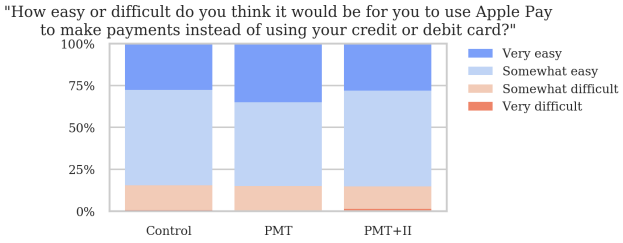


Figure 15: We did not find statistically significant evidence that our treatments affected perception of self-efficacy ($p = 0.523$).

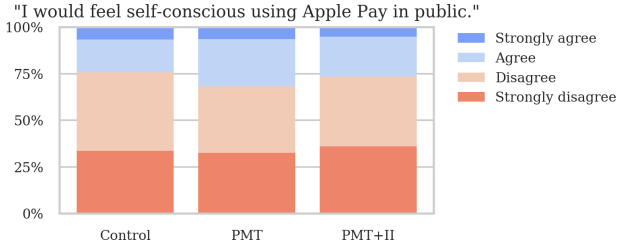


Figure 16: We did not find statistically significant evidence that our treatments affected self-consciousness ($p = 0.628$).

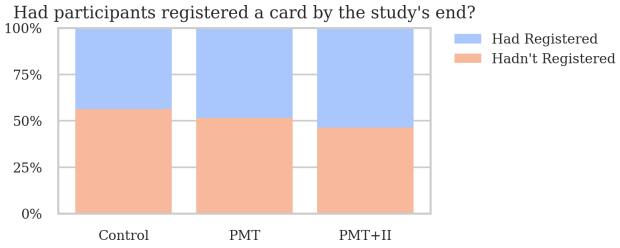


Figure 17: We did not find statistically significant evidence that our treatments affected whether participants would have a card registered in Apple Pay by the end of our study ($p = 0.237$).

Variable	β	e^{β}	p-value
CS_background	-1.438	0.237	**0.001**
experienced_fraud	-0.799	0.450	**0.024**
threat_severity	-0.353	0.703	0.414
knows_users	-0.041	0.960	0.911
age	0.010	1.010	0.527
SA6	0.042	1.043	0.259
response_efficacy	0.146	1.157	0.783
self_conscious	0.281	1.324	0.451
usefulness	0.368	1.445	0.509
self_efficacy	0.377	1.458	0.501
threat_susceptibility	0.383	1.466	0.296
own_watch	0.483	1.621	0.198
Face_ID	0.745	2.106	**0.022**
non-female_gender	0.870	2.387	**0.009**
prior_use	1.295	3.653	**<0.001**
intention	1.804	6.077	**<0.001**
treatment			0.297
PMT	0.390	1.477	0.394
PMT+II	0.698	2.010	0.123
Intercept	-4.856	0.008	**<0.001**

Table 6: Our logistic regression model for predicting use of Apple Pay by those who completed Survey #1, #2, and #3 ($n = 411$). e^{β} indicates the change in odds of using Apple Pay when the variable is true. p-values significant at $\alpha = 0.05$ are bolded. Cox & Snell $R^2 = 0.238$.

Variable	β	e^{β}	p-value
CS_background	-1.635	0.195	0.059
experienced_fraud	-1.519	0.219	**0.039**
response_efficacy	-1.326	0.266	0.245
messaged_info	-0.607	0.545	0.451
usefulness	-0.598	0.550	0.598
non-female_gender	-0.563	0.569	0.378
own_watch	-0.243	0.784	0.728
threat_severity	-0.118	0.889	0.904
SA6	-0.011	0.989	0.870
age	0.008	1.009	0.752
checked_intention	0.165	1.179	0.923
knows_users	0.463	1.589	0.461
Face_ID	0.902	2.464	0.207
messaged_plan	0.936	2.549	0.245
self_efficacy	1.032	2.805	0.267
prior_use	1.377	3.964	**0.028**
threat_susceptibility	1.458	4.297	0.058
self_conscious	1.639	5.148	**0.020**
visited_location	3.414	30.374	0.052
intention	20.768	1045582764.370	0.997
Intercept	-24.824	0.000	0.997

Table 7: Our logistic regression model for predicting whether those who received our implementation intention treatment used Apple Pay ($n = 136$). e^{β} indicates the change in odds of using Apple Pay when the variable is true. p-values significant at $\alpha = 0.05$ are bolded. Cox & Snell $R^2 = 0.385$.

8.1 Qualitative Interviews, Survey #1

Researchers at OMITTED are conducting a study to understand people's use of smartphones, credit cards, and debit cards to make payments.

All participants are asked to answer the screening questions below.

Based on your answers to the screening questions, we will determine your eligibility for our preliminary survey. If you are eligible, the preliminary survey will take about 10 minutes to complete. Only some of the participants who take this survey will be invited to participate in subsequent interviews and follow-up surveys. Participants will not be compensated for completing this survey: participants will only be compensated if they are selected to participate in subsequent parts of this study.

Do you live in the United States of America?
(Yes, No)

Do you speak English?
(Yes, No)

What is your age in years?
___

Are you able to visit OMITTED's campus for an interview?
(Yes, No)

Please review the details below:
[Consent Form]

Have you read and understood the information above?
(Yes, No)

Do you want to participate in this research and continue with the survey?
(Yes, No)

Do you use a smartphone?
(Yes, No, I don't know)

In which country did you purchase your smartphone?
(The United States, Other: ___, I don't know)

What kind of smartphone do you have? If you have multiple phones, answer based on the phone you use the most.
(iPhone, Samsung phone, Other Android phone, Other: ___, I don't know)

[Here we show the iPhone-specific text, but users saw text appropriate to the type of phone they selected.]
What **model of iPhone** do you have? For example, iPhone 4S, iPhone 5, etc. You can check your phone's model by opening the "Settings" app, going to "General", then "About". Your phone's "Model Name" should be listed on the "About" page.
_____

What **version of iOS** is running on your phone? For example, 7.9, 10.3, etc. You can check your phone's iOS software version by opening the "Settings" app, going to "General", then "About". Your phone's "Software Version" should be listed on the "About" page.
_____

Do you have an Apple Watch?
(Yes, No)

We would like to understand how you make payments at brick and mortar stores, restaurants, or other physical locations.

Do you have a **credit card**?
(Yes, No)

Do you have a **debit card**?
(Yes, No)

Please select **all options** which accurately complete the following statement: "**Sometime in the past**, I have made in-person payments in physical locations..."
... using cash
... using my credit card
... using my debit card
... using Apple Pay. Apple Pay allows you to make payments using your smartphone.
... using Google Pay. Google Pay allows you to make payments using your smartphone.
... using Samsung Pay. Samsung Pay allows you to make payments using your smartphone.

Please select **all options** which accurately complete the following statement: "**In the past month**, I have made in-person payments in physical locations..."
... using cash
... using my credit card
... using my debit card
... using Apple Pay. Apple Pay allows you to make payments using your smartphone.
... using Google Pay. Google Pay allows you to make payments using your smartphone.
... using Samsung Pay. Samsung Pay allows you to make payments using your smartphone.

Has your **credit or debit card information** ever been stolen?
(Yes, No, I don't know)

How **concerned or unconcerned** would you be if your credit or debit card information was stolen in the future?
(Not at all concerned, Slightly concerned, Moderately concerned, Very concerned)

How **likely or unlikely** do you think you are to have your credit or debit card information stolen in the future?
(Very unlikely, Somewhat unlikely, Somewhat likely, Very likely)

Has a **fraudulent purchase** ever been made on your credit or debit card?
(Yes, No, I don't know)

How **concerned or unconcerned** would you be if a fraudulent purchase was made on your credit or debit card in the future?
(Not at all concerned, Slightly concerned, Moderately concerned, Very concerned)

How **likely or unlikely** do you think you are to have a fraudulent purchase made on your credit or debit card in the future?
(Very unlikely, Somewhat unlikely, Somewhat likely, Very likely)

How did you find this study?
(OMITTED Participation Pool, Craigslist, Other: ___)

What gender do you identify with?
(Male, Female, Non-binary, Other: ___, Prefer not to answer)

What best describes your employment status?
(Working, paid employee; Working, self employed; Student; Not employed; Retired; Prefer not to answer)

Have you ever worked in or studied in a computer-related field? (Computer Science, IT support, etc.)
(Yes, No)

What is the highest level of school you have completed or degree you have earned?
(Less than high school, High school or equivalent, College or associate degree, Master's degree, Doctoral degree, Professional degree, Other: ___, Prefer not to answer)

Please estimate what your total household income will be for this year:
(Less than $10,000; $10,000 - $19,999; $20,000 - $39,999; $40,000 - $59,999; $60,000 - $79,999; $80,000 - $99,999; $100,000 or more; Prefer not to answer)

Have you ever lived outside the United States for more than 1 month?
(Yes, No, Prefer not to answer)

Where outside the United States have you lived the longest?

_____

If you are eligible for participation in this study, we may email you with an invitation to participate in the study. Because we have a limited number of interview slots available, we may not be able to interview all eligible candidates.

Name:

_____

Email address:

_____

8.2 Qualitative Interviews, Interview #1 Script

Hello XXX, my name is YYY [and my assistant's name is ZZZ]. Thank you for agreeing to participate in Interview #1. [I will be asking most of the questions, and ZZZ will be taking notes.] [I am/We are] very interested in your thoughts about credit cards, debits cards, and smartphones. This interview will be recorded, but the audio will not be shared with the public. Your responses will be kept anonymous, but quotes from your responses may be shared with the public.

Prior to completing Survey #1, you expressed your consent to participate in this study. However, the interview is completely voluntary, and you are free to end it at any time. The interview will take up to an hour. Is it alright if I start the audio recording now?

Great! I will start the audio recording now.

Alright, let's get started! Remember that there are no right or wrong answers to any of my questions.

Could you explain how you typically pay when you make a purchase in a physical location, like a brick and mortar store or restaurant?

In the survey, you also indicated that you used [a credit card][a debit card][credit and debit cards] to make purchases.

If has credit and debit: Is there a reason why you use one card instead of another?

If fraudulent purchase: In the survey, you wrote that a fraudulent purchase had been made on your credit or debit card. What happened? [Was it your credit or debit card?] [What did you do?] [How do you think it happened?]

If no fraudulent purchase: In the survey, you wrote that a fraudulent purchase had not been made on your credit or

debit card. Do you know anyone who has had a fraudulent purchase made on their credit or debit card? What happened?

If don't know: In the survey, you wrote that you weren't sure if a fraudulent purchase had been made on your credit or debit card. What did you mean by that? [What did you do?]

If card info was stolen: In the survey, you wrote that your credit or debit card information had been stolen before. What happened? [Was it your credit or debit card?] [What did you do?] [How do you think it happened?]

If card info wasn't stolen: In the survey, you wrote that your credit or debit card information had not been stolen before. Do you know anyone who has had their credit or debit card information stolen? What happened?

If don't know: In the survey, you wrote that you weren't sure if your credit or debit card information had been stolen. What did you mean by that? [What did you do?]

[I think most people carry their smartphones all the time, but this is a sanity check.]

What kind of smartphone do you use?

Do you carry your smartphone with you every day?

Are there any times when you do go out without your smartphone?

If Apple Watch: In the survey, you indicated that you have an Apple Watch. Do you wear it every day?

If they have an iPhone: [Pay] = [Apple Pay]

If they have a non-Samsung Android phone: [Pay] = [Google Pay]

If they have a Samsung phone:

If they previously used Google Pay and Samsung Pay:

In the survey, you said that you had previously used Google Pay and Samsung Pay, but haven't used either to pay in a physical location in the last month.

If you were going to use one of them again, which would you use? [Why?] [If you don't have a preference, that's okay, too.]

If they previously used Google Pay xor Samsung pay:

In the survey, you said that you had previously used [Google Pay][Samsung Pay], but haven't used it to pay in a physical location in the last month. Your phone is also compatible with [Samsung Pay][Google Pay], which can also be used to make payments through your phone.

If you were going to use Google Pay or Samsung Pay in the future, which would you use? [Why?] [If you don't have a preference, that's okay, too.]

If they haven't previously used Google Pay or Samsung pay:

In the survey, you indicated that you hadn't used either Google Pay or Samsung Pay to pay in a physical location before. Google Pay and Samsung Pay are both mobile payments systems that allow you to make payments in stores through your phone. Your phone is compatible with both Google Pay and Samsung Pay.

If you were going to start using one, which would you choose? [Why?] [If you don't know enough to choose, that's okay, too.]

If Samsung Pay: [Pay] = [Samsung Pay]
If Google Pay: [Pay] = [Google Pay]
Else: [Pay] = [Samsung Pay]

In that case, let's focus on [Pay] for the rest of the interview.

If they previously used [Pay], but haven't used it recently:

Omit if asked above: In the survey, you said that you had previously used [Pay], but haven't used it to pay in a physical location in the last month.

Tell me about your experiences using [Pay]. [When did you first use it? For how long did you use it? Was your experience using [Pay] good or bad?]

Is there a reason why you haven't used [Pay] recently?

If they have never used [Pay]:

Omit if asked above: In the survey, you indicated that you hadn't used [Pay] to pay in a physical location before. [Pay] is a mobile payments system that allows you to make payments in stores through your phone [Apple watch: or watch].

Had you heard of [Pay] before taking the survey?

If yes: How did you hear about [Pay]? Have you set it up on your phone [or watch]?

If yes: Have you tried using [Pay] before? Is there any reason why you haven't used it to make a payment before?

If no: Is there any reason why you haven't set it up?

There have been many big hacks where credit and debit card information was stolen from retailers. For example, Target was hacked in 2013, Home Depot was hacked in 2014, and Saks Fifth Avenue was hacked last year. Information about millions of cards was stolen in these hacks. If criminals get your credit or debit card information, they might use that information to make fraudulent purchases. If you notice fraudulent purchases on your credit card, you can probably get refunded. But if the purchases are made on your debit card, you might not be able to get your money back. In any case, you would need to get a replacement card with a new number, which would be inconvenient.

How concerned or unconcerned would you be if a fraudulent purchase was made on your credit or debit card [again]? Why?

[Concern Likert] On this scale, which option best reflects your answer?

How likely or unlikely do you think you are to have a fraudulent purchase made on your credit or debit card [again]? Why?

[Likelihood Likert] On this scale, which option best reflects your answer?

How concerned or unconcerned would you be if your credit or debit card information was stolen [again]? Why?
[Concern Likert] On this scale, which option best reflects your answer?

How likely or unlikely do you think you are to have your credit or debit card information stolen [again]? Why?
[Likelihood Likert] On this scale, which option best reflects your answer?

Do you know of anything you can do to prevent your credit or debit card information from being stolen? [Have you done anything to protect your card information?]

Thankfully, there are steps you can take to prevent your card information from being stolen and to protect yourself from card fraud. One of the best things you can do is to start using [Pay]. Instead of paying by swiping or inserting your card, you can make payments through your phone [or watch], which adds an extra layer of security. Payments made with [Pay] will still be charged to your credit or debit card, but because the payments go through [Pay], your card number is not shown to or recorded by retailers. This means that your card number cannot be stolen from transactions made with [Pay]. If your phone [or watch] is stolen, the thief will not be able to make payments because [Pay] is protected by your [Apple: fingerprint/Face ID and lock screen PIN][Other: lock screen]. Although no system is perfectly secure, security experts generally agree that [Pay] is more secure than paying with credit or debit cards. [Pay] takes just a few minutes to set up, and is widely accepted. Apple Pay: As of this year, Apple Pay is accepted in 65% of retail locations in the United States. For example, Giant Eagle, ALDI, Dunkin' Donuts, and CVS all accept Apple Pay. Google Pay: Google Pay is accepted at millions of locations. For example, Giant Eagle, ALDI, Dunkin' Donuts, and CVS all accept Google Pay. Samsung Pay: Samsung Pay is accepted at most retail locations in the United States.

These instructions show you how to set up [Pay] on your phone and how to make payments in stores.

If Apple Watch: Since you wear an Apple Watch, you might also be interested in the instructions for using Apple Pay on your watch. Using your watch might be more convenient than using your phone, and it's just as secure.

Please take a minute to review the instructions. If you want to set up [Pay], feel free to try it right now. If you run into any trouble, I would be happy to help you set it up. However, you do not have to set up [Pay] if you do not want to.

[Pass the handout to the participant]

[If they make a phone call: pause the recording to avoid recording their card number, SSN, or other sensitive information]

[Note whether they simply read the instructions, or tried to set up Pay. Ask if it's unclear.]

[Ask or observe what the participant had to do to verify their card (e.g., whether they had to call their bank, open the bank's app, etc.)]

[After pausing for at least 30 seconds, or however long it takes them to start setting up Pay]

[Remember to resume the recording, if it was paused]

After reviewing the instructions, do you have any questions about [Pay]?

If they simply reviewed the instructions:
How easy or difficult do you think it would be for you to set up [Pay]? Why?
[Difficulty Likert] On this scale, which option best reflects your answer?
Do you plan to try to set up [Pay] later, or would you rather not? Why?

If they tried to set up Pay:
Were you able to complete the setup of [Pay]?
If yes: How easy or difficult was it for you to set up [Pay]? Why?
[Difficulty Likert] On this scale, which option best reflects your answer?
If no: How easy or difficult was it for you try to set up [Pay]? Why?
[Difficulty Likert] On this scale, which option best reflects your answer?
Do you plan to try to set up [Pay] later, or would you rather not? Why?

How easy or difficult do you think it would be for you to use [Pay] to make payments instead of using your credit or debit card? Why?
[Difficulty Likert] On this scale, which option best reflects your answer?

[Agreement Likert] On this scale, please rate your level of disagreement or agreement with the following statement:
"If I were to start using [Pay], I would be less likely to have my card information stolen."
[And why do you choose that option?]

[Interest Likert] And on this scale, could you show me how interested or uninterested you are in using [Pay]? Why?

[Based on the person's stated level of interest and why they feel that way, I may skip the entire implementation intention section below.]

[To determine which handout to give the person. If they are ambivalent:
If they set up Apple Pay: handout corresponding to where they set it up
If they wear the Apple Watch all the time: watch handout
Else: iPhone handout]

Apple Watch: If you were going to start using Apple Pay, do you think you would be more likely to pay with your phone or with your watch? [Why?]

If you plan to use [Pay] in order to protect your credit or debit card information, one challenge might be simply remembering to use [Pay]. Forming a simple, concrete plan to use [Pay] can help you remember. If you like, you can use the plan template I have written on this handout.

[Hand the appropriate handout to the person]

Please take a minute to read through the plan. If you want to use [Pay] in the coming week, I encourage you to fill out the plan, since it may help you remember to use [Pay]. However, you do not have to fill out or follow the plan.

[Note the number of locations the person wrote and which boxes they checked]
[Number of locations written: ___]
[Number of boxes checked: ___ / 3]
[Final box checked? ___]

You are welcome to keep the plan and the instructions for using [Pay].

Do you want to use [Pay] in the coming week?

If they did not fill out the plan:
Is there a reason why you didn't fill out the plan?

If they did fill out the plan:
In the coming week, how likely or unlikely are you to visit at least one of the locations you listed? Why?
[Likelihood Likert] On this scale, which option best reflects your answer?

How likely or unlikely are you to try to use [Pay] at these locations? Why?
[Likelihood Likert] On this scale, which option best reflects your answer?

Do you think this plan will or will not help you remember to use [Pay]? Why?

Before we conclude the interview, do you have any other thoughts or questions?

Thank you for participating in this interview! In about one week, I will send you a short follow-up survey. After you complete that survey, I will email you a $15 Amazon e-Gift Card.

8.3 Qualitative Interviews, Survey #2

This survey is Survey #2 in the study "Use of Smartphones, Credit Cards, and Debit Cards" that you previously gave your consent to participate in. It will take about 10 minutes to complete this survey. If you complete this survey, we will email you a $15 Amazon e-Gift Card for your participation in our study.

Please answer the following questions about your experiences since our interview. There are no right or wrong answers to any of these questions, so please answer honestly.

You did not set up $PAY during our interview. Did you set up $PAY after the interview?
(Yes, No)

Please write a few sentences explaining why you [set up][did not set up] $PAY.

When did you set up $PAY?
(Today, Yesterday, A few days ago, Right after the interview)

Since our interview, have you **tried to use** $PAY to make a payment in a physical location?
(Yes, No)

Please write a few sentences explaining why you [tried][did not try] to use $PAY.

_____

Since our interview, have you **successfully used** $PAY to make a payment in a physical location?
(Yes, No)

Please write a few sentences describing your experience [using][trying to use] $PAY.

Since our interview, have you done anything else to protect your credit or debit card information from being stolen, or to protect yourself from credit or debit card fraud?
(Yes, No)

Please write a few sentences explaining what other steps you have taken to protect yourself from card information theft or card fraud.

_____

Are you interested in meeting for an additional 30 minute interview? If you participate in this interview, you will be compensated with an additional $15 Amazon e-Gift Card.
(Yes, No)

8.4 Qualitative Interviews, Interview #2 Script

Hello XXX, my name is YYY and my assistant's name is ZZZ. Thank you for coming to Interview #2. This interview is focused on your experiences since Interview #1. I will be asking most of the questions, and ZZZ will be taking notes. This interview will be recorded, but the audio will not be shared with the public. Your responses will be kept anonymous,

but quotes from your responses may be shared with the public.

Prior to completing Survey #1, you expressed your consent to participate in this study. However, the interview is completely voluntary, and you are free to end it at any time. The interview will take roughly 30 minutes. Is it alright if I start the audio recording now?

Great! I will start the audio recording now.

Alright, let's get started! Remember that there are no right or wrong answers to any of my questions.

In interview #1, we discussed [Pay].

If setup after Interview #1: During our last interview, you did not [setup][complete the setup of] [Pay].

When did you set up [Pay]? What reminded you to set up [Pay]?

What did you have to do to set up [Pay]? [Did you have to call your bank?]

Did you try to use [Pay] since our last interview?

If yes:

These instructions show how to review the transactions you made with [Pay]. Please take a minute to review the transactions you made since our interview. [Hands handout]

What were your experiences trying to use [Pay]?

Where did using [Pay] work the best? What happened?

Where was using [Pay] the most difficult? What happened?

Are there any other experiences you'd like to share?

If no:

Did you have any opportunities to use [Pay]?

Did you visit any stores, restaurants, or other locations where you thought Pay might be accepted?

Why did you not end up using [Pay]?

Did you use [Pay] [if yes: more or] less than you thought you would?

Did anything about [Pay] surprise you?

Did you encounter any challenges trying to use [Pay]?

Do you plan to use [Pay] in the future? Why?

What is your overall impression of [Pay]?

During our last interview, we discussed making a concrete plan to help you remember to use [Pay].

If filled out in interview: You filled out the plan template during our last interview. Do you remember what the plan was?

If not filled out in interview: You did not fill out the plan template during our last interview. Did you fill out the plan after the interview? Do you remember what the plan was?

If filled out at some point: Part of the plan was listing three stores or restaurants you thought you might visit. Do you remember what stores or restaurants you listed?

If yes: Did you visit any of those locations? Did you try to use [Pay] there? Did you try to use [Pay] at any other locations?

Did you find the plan to be helpful or not helpful? [Did the plan help you remember to use [Pay]?] [Was the plan more or less helpful than you thought it would be?] [Do you think you would have remembered to use Pay if you hadn't made the plan? Why?] [Can you think of any other strategies to help you remember to use Pay?]

Did you do anything else to help you remember to use [Pay]?

Potentially ask for clarification about free-text responses to survey.

Before we conclude the interview, do you have any other thoughts or questions?

Thank you for participating in this interview! In the next couple days, I will email you a $15 Amazon e-Gift Card.

8.5 Qualitative Interviews, Survey #3

This survey is Survey #3 in the study "Use of Smartphones, Credit Cards, and Debit Cards" that you previously gave your consent to participate in. It will take about 10 minutes to complete this survey. If you complete this survey, we will email you a $5 Amazon e-Gift Card.

Please answer the following questions about your experiences in the past week. There are no right or wrong answers to any of these questions, so please answer honestly.

In the past week, did you **try to use** $PAY to make a payment in a physical location?
(Yes, No)

Please write a few sentences explaining why you [tried][did not try] to use $PAY.

_____

In the past week, did you **successfully use** $PAY to make a payment in a physical location?
(Yes, No)

Please write a few sentences describing your experience [using][trying to use] $PAY.

_____

How likely or unlikely are you to use $PAY in the future?
(Very unlikely, Somewhat unlikely, Somewhat likely, Very likely)

In the past week, have you done anything else to protect your credit or debit card information from being stolen, or to

protect yourself from credit or debit card fraud?
(Yes, No)

Please write a few sentences explaining what other steps you have taken to protect yourself from card information theft or card fraud. _____

8.6 Controlled Experiment, Survey #1

Researchers at OMITTED are conducting a study to understand people's use of Apple services.

All participants are asked to answer the screening questions below.

Based on your answers to the screening questions, we will determine your eligibility for our Survey #1. If you are eligible, Survey #1 will take about 5 minutes to complete. Only some of the participants who take Survey #1 will be invited to participate in two follow-up surveys (Surveys #2 and #3).

In what country do you currently reside?
(United States, Other country)

What operating system (OS) does your primary mobile phone have?
(iOS (iPhone), Other, I don't know)

Do you speak English?
(Yes, No)

What is your age in years?
___

Based on your answers to our screening questions, we have determined that you are eligible for Survey #1.

Please review the details below:
[Consent Form]

Have you read and understood the information above?
(Yes, No)

Do you want to participate in this research and continue with the survey?
(Yes, No)

In which country did you purchase your iPhone?
(United States, Other country ___, I don't know)

What model of iPhone do you have? For example, iPhone 4S, iPhone 5, etc. You can check your phone's model by opening the "Settings" app, going to "General", then "About". Your phone's "Model Name" should be listed on the "About" page.
(Original iPhone, iPhone 3G, ..., iPhone 11 (or 11 Pro or 11 Pro Max))

What version of iOS is running on your phone? For example, 7.9, 10.3, etc. You can check your phone's iOS software version by opening the "Settings" app, going to "General", then "About". Your phone's "Software Version" should be listed on the "About" page.
_____

Do you own an Apple Watch?
(Yes, No)

Please select **all options** which accurately complete the following statement: "**Sometime in the past**, I have made in-person payments in physical locations..."
... using cash.
... using my credit card.
... using my debit card.
... using Apple Pay. Apple Pay allows you to make payments using your iPhone.

Please select **all options** which accurately complete the following statement: "**In the past week**, I have made in-person payments in physical locations..."
... using cash.
... using my credit card.
... using my debit card.
... using Apple Pay. Apple Pay allows you to make payments using your iPhone.

8.7 Controlled Experiment, Survey #2

Researchers at OMITTED are conducting a study to understand people's use of Apple services.

This survey is Survey #2 in the "Apple Services Study" that you previously gave your consent to participate in. It will take up to 30 minutes to complete this survey. If you complete **both** Survey #2 and Survey #3 **within 3 days of each survey invitation**, you will be compensated $7 total. We will invite you to Survey #3 one week after you complete Survey #2.

There are no right or wrong answers to any of our questions, so please answer honestly. Also, **please take the time to read the information in this survey carefully**.

[Control Group]
Apple Pay allows you to make payments in stores using your iPhone. Payments made with Apple Pay are charged to credit or debit cards that have been registered in Apple Pay.

[PMT and PMT+II Groups]
There have been many big hacks where credit and debit card information was stolen from retailers. For example, Target [70] was hacked in 2013, Home Depot [52] was hacked in 2014, and Saks Fifth Avenue [18] was hacked in 2018.

Information about millions of cards was stolen in these hacks. If criminals get your credit or debit card information, they might use that information to make fraudulent purchases. If you notice fraudulent purchases on your credit card, you can probably get refunded. But if the purchases are made on your debit card, you might not be able to get your money back [13]. In any case, you would need to get a replacement card with a new number, which would be inconvenient.

Thankfully, there are steps you can take to prevent your card information from being stolen and to protect yourself from card fraud. One of the best things you can do is to start using Apple Pay. Instead of paying by swiping or inserting your card, you can make payments through your phone, which adds an extra layer of security. Payments made with Apple Pay will still be charged to your credit or debit card, but because the payments go through Apple Pay, your card number is not shown to or recorded by retailers [10]. This means that your card number cannot be stolen from transactions made with Apple Pay. If your phone is stolen, the thief will not be able to make payments because Apple Pay is protected by your fingerprint and lock screen PIN. Although no system is perfectly secure, security experts generally agree that Apple Pay is more secure than paying with credit or debit cards [35]. Apple Pay takes just a few minutes to set up, and is widely accepted. As of this year, Apple Pay is accepted in 65% of retail locations [9] in the United States. For example, ALDI grocery, CVS pharmacy, and Starbucks all accept Apple Pay.

[See Figure 11]

[See Figure 12]

Please explain why you did not fill out the plan.

_____

How **concerned or unconcerned** would you be if a fraudulent purchase was made on your credit or debit card?
(Not at all concerned, Slightly concerned, Moderately concerned, Very concerned)

How **likely or unlikely** do you think you are to have a fraudulent purchase made on your credit or debit card?
(Very unlikely, Somewhat unlikely, Somewhat likely, Very likely)

How **easy or difficult** do you think it would be for you to use Apple Pay to make payments instead of using your credit or debit card?
(Very difficult, Somewhat difficult, Somewhat easy, Very easy)

Rate your level of **disagreement or agreement** with the following statement: "If I were to start using Apple Pay regularly, I would be **less likely** to be a victim of card fraud."
(Strongly disagree, Disagree, Agree, Strongly agree)

How **useful or not useful** do you think Apple Pay would be for making payments?
(Not at all useful, Slightly useful, Moderately useful, Very useful)

Rate your level of **disagreement or agreement** with the following statement: "I would feel self-conscious using Apple Pay in public."
(Strongly disagree, Disagree, Agree, Strongly agree)

Do you know anyone who uses Apple Pay?
(Yes, No, I'm not sure)

Do you have a credit or debit card registered in Apple Pay?
(Yes, No, I don't know)

When did you register a card in Apple Pay?
(Prior to taking this survey, While taking this survey)

Please explain why you do not know whether you have a credit or debit card registered in Apple Pay.

_____

Rate your level of disagreement or agreement with the following statement: "I intend **to register** a credit or debit card in Apple Pay in the next week." (Strongly disagree, Disagree, Agree, Strongly agree)

Rate your level of disagreement or agreement with the following statement: "I intend **to use** Apple Pay in the next week."
(Strongly disagree, Disagree, Agree, Strongly agree)

What is your overall opinion of Apple Pay? (Please write a few sentences)

_____

This is a link to the information about Apple Pay that we showed you earlier:
Apple Pay Setup, Use, and FAQ
Would you like us to send you a message on Prolific containing this link?
(Yes, No)

This is a link to your plan for using Apple Pay:
My Plan for Using Apple Pay
Would you like us to send you a message on Prolific containing this link?
(Yes, No)

Has a fraudulent purchase ever been made on your credit

or debit card?
(Yes, No, I don't know)

What gender do you identify with?
(Male, Female, Non-binary, Other: ___, Prefer not to answer)

What best describes your employment status?
(Working, paid employee; Working, self employed; Student; Not employed; Retired; Prefer not to answer)

Have you ever worked in or studied in a computer-related field? (Computer Science, IT support, etc.)
(Yes, No)

What is the highest level of school you have completed or degree you have earned?
(Less than high school, High school or equivalent, College or associate degree, Master's degree, Doctoral degree, Professional degree, Other: ___, Prefer not to answer)

Please estimate what your total household income will be for this year:
(Less than $10,000; $10,000 - $19,999; $20,000 - $39,999; $40,000 - $59,999; $60,000 - $79,999; $80,000 - $99,999; $100,000 or more; Prefer not to answer)

Each statement below describes how a person might feel about the use of security measures. Examples of security measures are laptop or tablet passwords, spam email reporting tools, software updates, secure web browsers, fingerprint ID, and anti-virus software.

Please indicate the degree to which you agree or disagree with each statement. In each case, make your choice in terms of how you feel **right now**, not what you have felt in the past or would like to feel.

There are no wrong answers.
(Strongly disagree, Somewhat disagree, Neither disagree nor agree, Somewhat agree, Strongly agree)

I seek out opportunities to learn about security measures that are relevant to me.

I am extremely motivated to take all the steps needed to keep my online data and accounts safe.

Generally, I diligently follow a routine about security practices.

I often am interested in articles about security threats.

I always pay attention to experts' advice about the steps I need to take to keep my online data and accounts safe.

I am extremely knowledgeable about all the steps needed to keep my online data and accounts safe.

8.8 Controlled Experiment, Survey #3

Researchers at OMITTED are conducting a study to understand people's use of Apple services.

This survey is Survey #3 in the "Apple Services Study" that you previously gave your consent to participate in. It will take up to 5 minutes to complete this survey. If you complete this survey **within 3 days of the survey invitation**, you will be compensated $7 total for completing Survey #2 and Survey #3.

There are no right or wrong answers to any of our questions, so please answer honestly. Also, **please take the time to read the information in this survey carefully**.

In Survey #2, you indicated that you [did not have][did not know whether you had] a credit or debit card registered in Apple Pay.

Since taking Survey #2 on $DATE, **have you registered a credit or debit card in Apple Pay**?
(Yes, No)

Please explain why you did not register a credit or debit card in Apple Pay.

_____

Rate your level of disagreement or agreement with the following statement: "I intend **to register** a credit or debit card in Apple Pay in the next week."
(Strongly disagree, Disagree, Agree, Strongly agree)

Since completing Survey #2 on $DATE, have you made an in-person payment in a physical location using Apple Pay?
(Yes, No, I don't know)

Since completing Survey #2 on $DATE, how many payments have you made **with Apple Pay** in physical locations?

___

Please explain why you [used][did not use][do not know whether you used] Apple Pay.

_____

Did you use Apple Pay in a location where you had previously paid with a credit or debit card?
(Yes, No, I don't know)

[PMT+II Group, if they wrote at least one location]
In Survey #2, you made a plan to use Apple Pay.

Since completing Survey #2 on $DATE, which of the locations in your plan, if any, **have you visited**?
($LOCATION_1, $LOCATION_2, $LOCATION_3)

Please select all options which accurately complete the following statement: "**Since completing Survey #2 on**

$DATE, I have made in-person payments at **$LOCATION_N**..."
...using cash
...using my credit card
...using my debit card
...using Apple Pay. Apple Pay allows you to make payments using your iPhone.
...using another payment method. Please specify: ___

How **concerned or unconcerned** would you be if a fraudulent purchase was made on your credit or debit card?
(Not at all concerned, Slightly concerned, Moderately concerned, Very concerned)

How **likely or unlikely** do you think you are to have a fraudulent purchase made on your credit or debit card?
(Very unlikely, Somewhat unlikely, Somewhat likely, Very likely)

How **easy or difficult** do you think it would be for you to use Apple Pay to make payments instead of using your credit or debit card?
(Very difficult, Somewhat difficult, Somewhat easy, Very easy)

Rate your level of **disagreement or agreement** with the following statement: "If I were to start using Apple Pay regularly, I would be **less likely** to be a victim of card fraud."
(Strongly disagree, Disagree, Agree, Strongly agree)

How **useful or not useful** do you think Apple Pay would be for making payments?
(Not at all useful, Slightly useful, Moderately useful, Very useful)

Rate your level of **disagreement or agreement** with the following statement: "I would feel self-conscious using Apple Pay in public."
(Strongly disagree, Disagree, Agree, Strongly agree)

Rate your level of disagreement or agreement with the following statement: "I intend **to use** Apple Pay in the next week."
(Strongly disagree, Disagree, Agree, Strongly agree)

Do Privacy and Security Matter to Everyone? Quantifying and Clustering User-Centric Considerations About Smart Home Device Adoption

Natã M. Barbosa, Zhuohao Zhang, and Yang Wang

University of Illinois at Urbana-Champaign
{natamb2,zhuohao4,yvw}@illinois.edu

Abstract

While consumer adoption of smart home devices continues to grow, privacy concerns reportedly remain a roadblock to mass adoption. However, it is unclear exactly how the interplay between privacy and other factors affect consumers' purchase decisions, and whether such considerations are held only by certain consumer groups but not others. In order to unpack the decision-making process of smart home device adoption, we conducted a mixed-method analysis using online survey data collected from 631 US participants. Our analysis uncovered motivators and blockers of purchase decisions, along with their relative importance. We found that consumers can be segmented based on their considerations into three clusters: affordability-oriented, privacy-oriented, and reliability-oriented. We present an in-depth quantification of consumer considerations on smart home device adoption along with desired privacy and security features consumers wish to use to protect their privacy in the smart home.

1 Introduction

Consumer adoption of smart home devices continues to see steady growth. A recent study in the US by Statista [34] reports 41 million homes with at least one smart home device in 2020. This figure represents a 32.4% household penetration in the US, an increase of 18.7% from the previous year. Such growth continues despite consumer privacy and security concerns reportedly remaining a roadblock to mass adoption [20]. Taken together, these reports pose an interesting conundrum: why do we see increased adoption despite widespread privacy and security concerns? Unpacking this question means understanding how much considerations of privacy and security weigh into purchase decisions alongside other factors, and whether such considerations may be held only by certain consumer groups but not others.

From an economic perspective, consumers consider privacy alongside other factors, and make a decision based on their calculus of whether expected benefits would outweigh expected costs (e.g., associated privacy risks) [1]. Previous work found that privacy is an important factor in the purchase decisions of Internet of Things (IoT) devices for many users, standing only behind features and price [14]. However, it is unclear to what extent considerations of privacy and security stand against the many factors that may motivate or prevent consumers from adopting such devices. For example, would the expected convenience outweigh the privacy concerns? Is the presence of privacy and security features more important than the absence of them? In addition, while many consumers claim to have privacy concerns, the growing number of devices installed each year is evidence that either privacy and security concerns do not stop many from adopting, or such concerns are outweighed by other factors when it comes to actual adoption. Either way, adoption statistics suggest that privacy and security considerations may play a different role to different people, and people could be segmented based on their purchase considerations, since privacy or security may not be a pre-purchase consideration for some people [14].

To address this conundrum, we report on data from a US-based survey with 631 participants where half of all participants reported having a smart home device and the other half did not. Participants were asked what could motivate or prevent them from adopting smart home devices, separately. Our mixed-methods analyses quantify and cluster motivators and blockers in order to provide an in-depth understanding of consumers' decision-making process in smart home device purchases. Participants were also asked about what privacy and security protections they desire for smart home devices.

Our findings show that good privacy or security practice was considered as a motivator only by 11% of the partici-

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

pants, while privacy or security concerns/risks being a blocker for half of the participants. Across all participants, the top considerations are ranked as follows, with "+" for motivators and "-" for blockers: +*convenience*, -*privacy*, -*price*, -*security*, +*cost-saving*, -*risk*, -*reliability*, and +*control*. Our clustering analysis reveals three groups of consumers: (1) affordability-oriented; (2) privacy-oriented; and (3) reliability-oriented. We also built a decision tree to predict which consumer cluster a person belongs to based on the person's purchase considerations. The decision tree can help explain what matters in people's purchase decision-making process. We discuss implications of our results on the use, privacy control and regulation of smart home devices.

The contributions of this research include (1) quantified relative importance of different factors as motivators and blockers in people's smart home device adoption considerations; (2) consumer segmentation based on these factors; (3) a list of desired privacy protections for smart homes; and (4) actionable recommendations based on our findings. In summary, our work helps unpack the smart home adoption conundrum and provide guidance on how companies and policymakers could enable consumers to make more informed decisions about smart home device purchases.

2 Related Work

2.1 Users' IoT Privacy Considerations

A great number of qualitative studies have looked into privacy and security concerns and expectations of IoT and smart home users (e.g., [6–8, 10, 21, 23–25, 38–42]). These studies have revealed privacy and security concerns among both owners and non-owners of smart home devices. More importantly, a recurring theme are the trade-offs between price, functionality, and privacy [8, 10, 40], where some users have reported prioritizing functionality over privacy and *vice-versa*.

A recent interview study by Emami-Naeini *et al.* [14] on user privacy and security considerations on IoT device purchase revealed most device owners did not consider privacy or security prior to purchasing, but did become concerned once the devices were installed in their homes. The attributed reason being lack of access to or information about privacy and security of the devices. Through their interviews, the authors also identified 16 factors that influenced users to purchase smart home devices, and later asked 200 survey respondents to rate the influence of the 16 factors on a 5-point Likert-scale. Their survey results revealed privacy as the third-most influencing factor on participant decisions, standing only after features and price. One of the findings from this study is that privacy and security may not be a consideration for many consumers, and that some would only consider it after being exposed to such considerations. This finding motivated the authors to design and evaluate privacy labels for IoT devices [13, 14] in order to inform and educate consumers about privacy and security aspects before purchasing IoT devices.

Prior qualitative interview studies have also explored user-centric privacy behaviors, tools and protections (e.g., [14, 39]), revealing strategies such as frequent updates, strong access controls, device re-positioning, using separate routers, data localization, disconnection from the Internet, and the use of a private mode.

2.2 Privacy Value in Emerging Technologies

From an economic perspective, if personal and societal privacy are considerations held by consumers, they are pondered as part of a calculus that will inform and mediate decision-making with regard to the adoption of emerging technologies involving privacy risks [1, 4, 32]. Accordingly, numerous studies have looked at consumer considerations of privacy in technological economic transactions (e.g., [2, 9, 11, 19, 29–31]). Notably, Egelman *et al.* [11]'s experiment on smartphone app installs showed that given the appropriate choice architecture highlighting app permissions, users are willing to pay a premium for privacy, potentially leading to more rational decisions. Danezis *et al.*'s study [9] revealed that smartphone users may be willing to allow their location to be monitored for a given price, it being higher when users traveled frequently or communicated with partners using their phone.

Behavioral studies such as these suggest consumers engage in risk assessments heavily controlled by the underlying context, which can be influenced by biases, heuristics, and framing effects [1]. These studies also often point to a potential paradox where people's stated privacy attitudes and preferences deviate from their observed behavior, commonly referred to as the "privacy paradox." [26]. However, contemporary views on privacy decision-making provide possible explanations of and even refute the existence of such a paradox. For example, Adjerid *et al.* [3] argues that the constancy of normative factors (e.g., privacy preferences, settings, regulation) and behavioral effects (e.g., choice framing, defaults) must be challenged in hypothetical and actual choice settings, since "consumers may both overestimate their response to normative factors and underestimate their response to behavioral factors." Solove [32] refutes the existence of a privacy paradox altogether, arguing that behavior is better understood as "choices about risk in specific contexts" and attitudes as "people's broader valuation of privacy, often across many contexts." Solove argues that privacy decision-making should be instead approached as behavior that involves risk in which people's decisions are influenced by many factors. Such an approach is also more suitable to the reality of today, where new products are increasingly made to be Internet-connected and escaping the associated privacy and security risks becomes increasingly difficult. According to Solove, this approach also stands in contrast with the assumptions made in the privacy paradox, which often stems from "leaps in logic" that generalize from specific contexts to broad attitudes across contexts [32].

Arguably, reports indicating privacy is a blocker to mass-adoption of smart home devices (e.g. [20]), yet devices gaining significant adoption in the past few years [34] hint at a potential privacy paradox taking place in the smart home. However, user studies paint a more nuanced picture about smart home device adoption considerations. For instance, some consumers indeed care more about privacy than others and choose not to buy smart home devices [10]. Some consumers do not consider privacy or security before purchasing [14], or deem them less important than other factors [14, 40]. Consumers' expectations and concerns about privacy and security are also shaped by their preceding experiences with computing technologies and their underlying organizations, causing them not to expect privacy by default [35]. Such expectations pose consequences to individual privacy valuations in smart home device purchases, given that consumers may be loss-averse with regard to personal privacy, giving it more value when their current stance "includes" it and lower value otherwise [5].

These studies point to a complex setting where it is unclear where privacy and security considerations may stand in smart home device purchase decisions, suggesting a nuanced assessment of the interplay between functionality, price, convenience, and privacy and security risks. Such a setting also poses an interesting opportunity for the study of privacy given the long-established association of a home and privacy. To the best of our knowledge, no prior works have tried to unpack this interplay and quantify considerations at a large scale. Findings from previous works also point to different, perhaps segmented consumer priorities when considering the purchase of smart home devices, where some may prioritize privacy while others may not, yet this potential segmentation has not been explored.

2.3 Distinctions from Prior Work

Interview studies and surveys have found that privacy and security risks are a major consideration with regard to smart home device adoption. These studies have also shown that some consumers may prioritize price over privacy while others will not, and that there is a tension between privacy and functionality, both of which are common considerations. However, to the best of our knowledge, there has not been an attempt to systematically quantify to what extent privacy and security are important alongside other factors such as price, expected convenience, and interoperability. For example, is privacy more of a blocker than price is a motivator? Answering these questions will give developers and researchers a more contextualized understanding of what considerations are being made alongside privacy, giving them knowledge to design effective privacy features and tools.

More importantly, prior interviews, surveys, and smart home adoption statistics suggest privacy and security may not be considerations held by all consumers, and based on our quantification, we conducted a clustering analysis of consumers. Our goal was to understand whether consumers can be segmented with regard to their priorities, and reveal whether there would be one segment of consumers who would be privacy-oriented, and if so, what might the other competing clusters be. This is important to understand given that interview studies with smart home device owners revealed that they did not pay attention to privacy before purchasing, but only became concerned after the purchase. [14]. Uncovering potential segments will provide opportunities for education and awareness to cater to consumers where they stand with respect to privacy and security. For example, if low price is a strong motivator for a segment of consumers, educating them on potential privacy/security risks with Internet-connected products could prevent undesired privacy/security outcomes.

While privacy concerns, expectations, and valuations around the smart home devices have been studied, potential privacy behaviors, tools, and protections remain largely under explored. Such exploration could, for example, identify user-centered privacy features within the context of the smart home. To address this, we present a ranked, comprehensive list of privacy tools, behaviors, and features of smart home devices desired by survey participants.

3 Method

In this section, we present details of the survey design and the data analysis procedure.

3.1 Survey

The present work encompasses a data analysis from a survey conducted on Amazon Mechanical Turk (AMT) introduced in a previous paper of our authorship [5]. In the present work, we report on a different portion of the survey, not previously reported on. One of the survey's goals was to collect user preferences for different smart home information flows, then create machine learning models to predict such preferences, which was the contribution of the previous paper. Another goal of this survey was to unpack the decision-making process of consumers on smart home device adoption, focusing on where privacy and security concerns and desired protections stand within such considerations. The latter goal had not been addressed before and is the topic of the present work.

The survey presented randomly generated vignette scenarios combining different attributes and purposes of use in the template "The manufacturer/developer of your smart home device is accessing or inferring [attribute]. They are using this information for [stated purpose]," asking participants to provide their comfort levels and preferences on whether they would allow or deny a given information flow. Each participant was presented with four of such scenarios. In addition, for each scenario, participants were asked to review and select up to three out of 14 transmission principles that could make them more or less comfortable with the original

scenario. In [5], we also present an analysis and modeling from an economics-related question about the purchase of a voice assistant, in which participants were asked to specify how much they would be willing to receive or pay extra for privacy protections in the purchase of such voice assistant. The survey received responses from 698 participants, with a median completion time of 19 minutes and compensation of $1.50 (USD), plus equivalent bonuses for participants who took longer than average to finish. The survey was approved by Syracuse University's IRB and the survey protocol can be found in Appendix A.1. More details about the scenario and economics-related questions of the survey can be found on the previous paper [5].

Participants of the survey were required to have 95% of all-time approval rate on previously submitted work and be based in the US. Participants also had to pass a manual qualification task which required them to select three out of six devices they believed to be smart home devices after having read briefly about them. A brief explanation with three short paragraphs largely based on the Wikipedia definition for smart home devices preceded the qualification task, along with three pictures, one with an Ecobee smart thermostat, another with an Alexa smart speaker, and a third one with a Nest smart camera. Then users were asked to select three images containing a smart home device, out of images (with alternative text) of a smart thermostat, a voice assistant, a blender, a DSLR camera, a desk lamp, and a smart bulb. Only users who selected the three smart home devices, namely the smart thermostat, the voice assistant, and the smart bulb, were allowed to proceed.

In the present work, we analyze the data from three open-ended questions asked in the survey that have not been previously reported on. Two of the three questions were presented *immediately after* the qualification task and *immediately before* the four random scenarios. The two first questions followed an introductory question posed to elicit participants' thought processes: *What factors do you consider when making decisions about adopting smart home devices? Please answer below.* Then two follow-up questions were asked:

1. *For example, what are factors that could motivate you to purchase smart home products?*

2. *Similarly, what are factors that could keep you from adopting the technology?*

Privacy or security were not mentioned during the qualification step nor in the two questions as not to prime participants. This helped avoid any privacy or security-related bias in participants' thought processes when answering the questions.

The third question was presented *immediately after* participant responses to the four scenarios and *immediately before* the economics-related question reported in the previous paper:

> 3. *What privacy behaviors you would like to be able to adopt in the context of smart home devices? For example, would there be any privacy-protecting tools, configurations, and techniques you would like to use?*

Answers to the three questions were mandatory, and we did not use or report on the data collected from any of these three questions before. We acknowledge that answers to the third question could include biases resulting from the four scenarios presented earlier in the survey, and this is a limitation of the answers to this question. For example, participants who responded to a scenario where the purpose of use was targeted advertisement may have been primed to mention protections against secondary use. We still analyze and report the data given that (1) participants had a comparable experience because scenarios were created randomly; (2) the analysis on this question is a secondary contribution of our work and (3) the answers still provide valuable insights that have resulted from participant's engagement with a survey focused on potential information flows of the smart home. The survey scenarios also indirectly provided a broad grounding around the potential privacy and security risks associated with using a smart home device, enabling them to provide contextualized and meaningful responses, as evidenced by the level of articulation observed in participants' responses to question 3 (see Table 2 in Appendix).

Following the third question within the scope of the present work (i.e., the question about privacy tools, configurations and techniques), participants were asked the economics-related question presented in [5], questions from the Awareness, Control, and Collection dimensions of the Internet User Internet Privacy Concerns (IUIPC) scale [22], and demographics: gender identity, age bracket, hours spent on the Internet weekly, whether they owned a smart home device, how many smart home devices owned, what specific devices were owned (from a list of 16 types), occupation, education level, income bracket, size of household, whether the participant had children, and marital status. We tested whether any of these demographics would be associated with mentioning privacy or security as a motivator or a blocker in questions 1 and 2.

We initially read each response to check the quality of the responses. We manually inspected the answers to each question and removed responses from 67 participants (9.6%) due to their answers not being meaningful and/or being random copy/paste. Our cleaned up data set resulted in responses from 631 participants. This cleaning process generated the data set used in our qualitative and quantitative analyses.

3.1.1 Participant Demographics

Gender Identity and Age 48.8% identified as female (50.7% male, 0.5% other), 44% as 26-35 years-old, 21% as 36-45 years-old, 16% as 18-25 years old, 9% as 46-55 years old, and 10% over 56.

Education and Income 39% of participants reported having a Bachelor's degree, followed by some college but no degree (21%), master's (14%), associate (13%), high school

(9%), professional (3%) and doctoral (1%). 24% of participants reported earning no more than $30k, 34% no more than $60k, 19% no more than $90k, and 22% over $100k.

Household Size The average household size was 2.73 (Mdn=3, SD=1.38). 45% reported being married, 45% single, 8% divorced, 2% separated, 1% widowed. 44% of respondents reported having children.

Occupation Participants reported a diverse set of occupations, including agriculture, sales, therapist, teacher, attorney, software engineer, student, insurance worker, and accountant. 9% of respondents provided an IT-related occupation.

Device Ownership 48% reported owning a smart home device. The most popular type of device owned was voice assistant, followed by security camera, smart lighting, audio/speakers, and thermostat.

IUIPC Scores We added up the score for the responses to the questions within each corresponding dimension. The average Awareness score was 19.23 (Mdn=21, SD=2.68, Min=7, Max=21). The average Control score was 18.4 (Mdn=19, SD=2.8, Min=6, Max=21). The average Collection score was 23.68 (Mdn=25, SD=4.56, Min=6, Max=21).

3.2 Data Analysis

3.2.1 Privacy/Security and IT Annotation

With a focus on privacy and security, we annotated each row with whether the participant referred to privacy or security as a motivator or a blocker. As a first step in our analysis, we grouped privacy and security responses together because prior works found that IoT users had limited knowledge of privacy and security and often could not distinguish between them [14]. This grouping also enabled us to start with a high-level analysis involving descriptive and test statistics. Examples of when we flagged privacy or security are if participants mentioned "privacy concerns" or said "hacking, " "tracking and monitoring," or "stalking me to market to me." We later used this annotation to generate descriptive statistics about the overall number of responses mentioning privacy or security, in addition to conducting statistical significance tests for relationships with demographics. Although we grouped privacy and security responses for a high-level analysis, we considered them separately during the coding and segmentation analyses. We also report the statistical test results when considering privacy and security separately.

Additionally, we annotated whether each participant's occupation was related to IT. We did this because we wanted to be aware in our analysis when participants could have heightened technical expertise. Some occupation examples where participants were marked include "IT Help Desk Analyst," "Software QA," "Programmer," and "Computer Technician." Only 9% of participants reported an occupation related to IT.

3.2.2 Coding

We conducted inductive coding on the open-ended participant responses, coded by two researchers. We first read each answer in order to get acquainted with the responses and underlying, recurring themes. Then, we drew a random sample of 15% of responses and coded them individually. After coding the sample individually, the two researchers met in person to review their individual codes, discuss, and converge into a code book that would be used for the remaining of the responses. A code named "other" was created for which answers not belonging to any of the codes in the final code book were assigned. The final code book contained 23 categories for motivators, 20 categories for blockers, and 19 categories for privacy tools and behaviors. Using this code book, the two researchers coded the remaining 85% of the responses. In our coding procedure, each answer was allowed to have more than one category. We calculated inter-coder agreement between the two coders using Cohen's Kappa: 87% for motivators, 91% for blockers, and 88% for privacy tools. These values indicate excellent agreement between the two coders [15].

3.2.3 Quantitative Analysis

We merged the coded data sets resulting from the coding procedure based only on the agreements between the two coders. For example, if both coders assigned the same given category to a response, then the category was assigned in the final data set, indicated with a value of 1, otherwise this value was 0. Once our final data set was generated, our quantitative analysis was divided into three parts.

The first part consisted of testing relationships of demographics with whether participants reported privacy or security as a motivator or blocker. We used Chi-square association tests for categorical variables such as gender identity, or owning a smart home device, and logistic regression for numerical and ordinal variables, such as the age bracket, education level, and IUIPC awareness, control, and collection dimensions.

The second part consisted of analyzing the frequency of each factor either as a motivator or a blocker, and quantifying the relative importance of the factors side-by-side. The latter task involved creating a wide data set with each column representing a factor mentioned either as a motivator or a blocker from the coded and merged data set. If the factor was mentioned in the motivator question, the column was given the value of 1. If the factor was mentioned in the blocker question, the column was given the value of -1, and 0 otherwise (i.e., not being mentioned in either). This allowed us to compare motivators and blockers side by side, surfacing whether each factor is more of a motivator or a blocker, as determined by their calculated average values. For instance, a device being privacy-invasive might be a blocker whereas not being privacy-invasive might not be a motivator.

The third part consisted of conducting a clustering analysis with the considerations and creating a decision tree model to

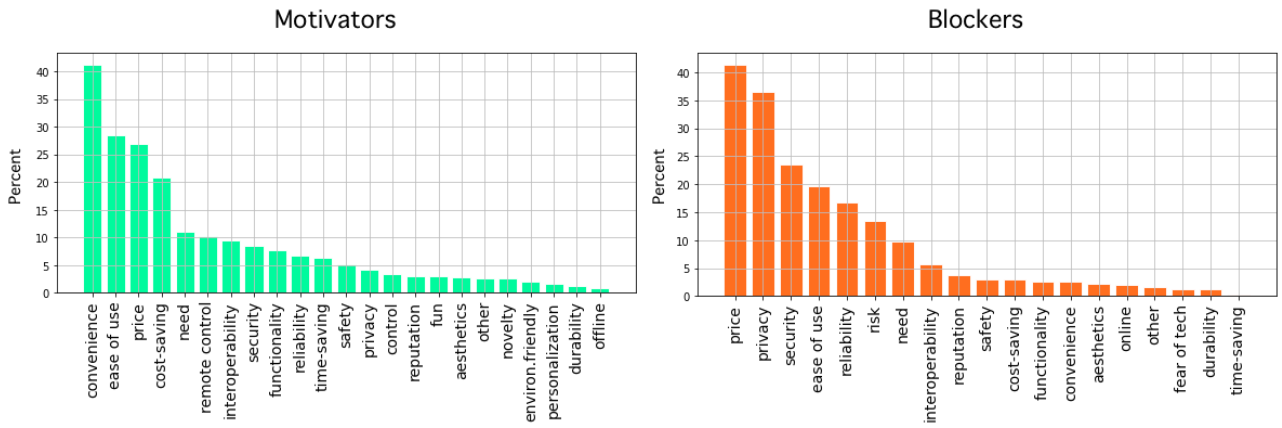


Figure 1: Frequency (%) of motivators (left) and blockers (right) across all survey respondents. Top motivators are convenience, ease of use, price, cost-saving, and need. Top blockers are price, privacy, security, ease of use, and reliability.

predict the assigned cluster based on participant considerations. With the wide data set encoded with -1,0,1 columns, we generated a dendrogram to visualize hierarchical clusters. Then, we assigned each response to a cluster with k-means clustering with $k = 3$. This segmented the participants based on their purchase considerations. We report the ranking of motivators and blockers for each of the three participant clusters and cross-checking of the clusters with specific demographics, such as having technical background, owning a device, or the reported gender identity. The last step in this part consisted of creating a classifier to predict the consumer clusters. We created a decision tree classifier and evaluated it with 10-fold cross validation. We report our results, along with the resulting decision tree of the trained model.

4 Results

4.1 Privacy and Security Considerations

Across all participants, only 11% mentioned privacy and security among the factors which would *motivate* them to adopt smart home devices (separately, privacy=4.12%, security=8.56%). For example, participants mentioned *"if it's non-intrusive"* and *"the security of the system and how protected it is from outside tampering,"* as motivators related to privacy and security. In comparison, 50% mentioned privacy or security as something that would *prevent* them from adopting (separately, privacy=36.61%, security=23.61%). For instance, participants responded with *"the security of the item, could it be hacked? Could it have a camera that could turn on and be hacked? Would my personal information be safe?"* and *"mostly, companies obtaining information on my personal life. Just because consumers buy from a company doesn't mean the company can own them."* Among participants who reported owning a smart home device, 44% mentioned privacy or security concerns as a blocker (separately, privacy=16%, security=10%), while this number was 56% for participants who reported not having a smart home device (separately, privacy=21%, security=14%). A Chi-square association test examining the relationship between having or not having a smart home device and mentioning or not mentioning privacy or security as a blocker produces a statistically significant result: χ^2 (1, N = 631) = 8.901, $p < 0.001$, suggesting that people who have privacy and security concerns are less likely to be associated with having a smart home device. When testing for privacy and security separately, this relationship is also significant: privacy χ^2 (1, N = 631) = 4.041, $p < 0.05$, and security χ^2 (1, N = 631) = 3.8942, $p < 0.05$. A Chi-square association test showed no difference between having one *versus* multiple devices.

Further, we investigated the relationship between privacy or security considerations and the demographics collected in our study. Namely, we tested gender identity, age bracket, whether participants had an IT-related occupation, education, income bracket, household size, whether participants had children, and marital status. Given that there were no pre-planned hypotheses or theoretical model for testing these demographics, we applied Bonferroni correction to control family-wise Type I errors, thus taking .00625 as our significance level considering 8 tests. None of the tests yielded statistically significant results at the corrected p-value. The results were the same when testing for mentioning privacy and security separately.

Finally, we tested the relationship between the IUIPC constructs and stating a privacy or security consideration with a logistic regression model using the three IUIPC dimensions as predictors. For both the motivator and blocker question, the IUIPC Collection dimension was a significant predictor ($p < .05$, exp(estimate) = 1.06 for blocker, 1.11 for motivator), indicating that people who were more concerned about data collection in general (based on the IUIPC) were more likely to be associated with mentioning privacy or security considerations in our study. In a model comparison via the



Figure 2: Ranked importance of considerations. The X axis represents the absolute average value of each consideration. If the non-absolute average value is positive, color is green, or red otherwise. While convenience and cost-saving are the top motivators, privacy, price, security are the top blockers.

Likelihood Ratio Test, both the motivator and blocker models containing the IUIPC predictors resulted in a statistically significant difference against the null model ($p < .001$).

These results suggest that privacy and security considerations may be preventing actual adoption, may not be associated with particular demographics, and may have a relationship with how participants felt about online data collection.

4.2 Relative Importance of Motivators and Blockers

For all respondents, we calculated the frequency of each motivator and blocker. The top five motivators were convenience, mentioned by 41.2% of participants as a motivator, ease of use (28.37%), price (26.94%), cost-saving (20.76%), and need (10.94%). The top five blockers were price, mentioned by 41.36% of participants, followed by privacy (36.61%), security (23.61%), ease of use (19.65%), and reliability (16.8%). Figure 1 shows the percentage of participants who mentioned the motivators and blockers, and Table 1 (Appendix) shows all factors with examples. 234 participants (37%) mentioned at least one factor both as a motivator and a blocker, with the most frequent being price, with 18.2%, then 11.6% for ease of use, 5.1% for security, 3.8% for need, 3% for reliability, 2.5% for privacy, 2.5% for interoperability, then six more factors mentioned as both motivators and factors by fewer than 2% of participants, with the remaining 15 factors being mutually exclusive, meaning they were either mentioned only as a motivator or as a blocker.

When combining motivators and blockers via their average values across all participants, it is possible to determine whether a factor was mostly a motivator or a blocker. Figure 2 shows the distribution of the factors ordered by their absolute average value. The top motivator is convenience, followed by privacy, price, and security as top blockers. This suggests that most consumers might consider the three top blockers after convenience, then whether the device will save money in the long-term, then the risk of owning the device, etc, according to the ranking in Figure 2.

4.3 Clustering Consumers

One of our research goals was to examine whether participants could be clustered with regard to their purchasing considerations. To do this, we used the wide data set with each possible consideration as a column, resulting in 29 columns of value 1 if it was a motivator, -1 if it was a blocker, and 0 otherwise. We identified the optimal number of clusters via a dendrogram generated from Agglomerative Clustering, a bottom-up hierarchical clustering approach. The dendrogram analysis (Figure 3) revealed three major clusters, as indicated by the number of vertical lines crossed by the horizontal black line placed at the end of the longest vertical line. Agglomerative clustering starts by assigning each data point to its own cluster, then moves up, grouping instances based on the smallest distance, such as the Euclidean distance, eventually making all data points belong to a single cluster. The optimal number of clusters is chosen by crossing a horizontal line over the longest vertical line and verifying how many vertical lines

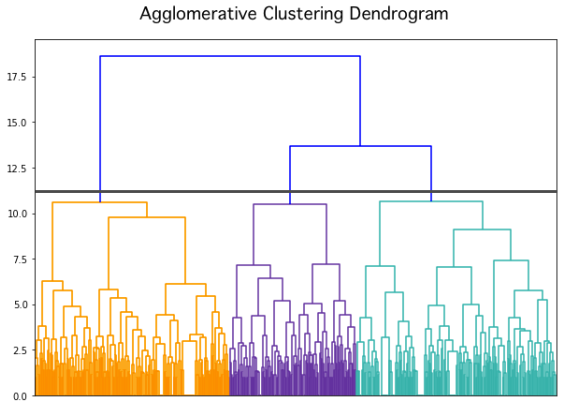


Figure 3: Hierarchical clustering dendrogram. The graph reveals three clusters, as indicated by the number of horizontal lines crossing the longest vertical line.

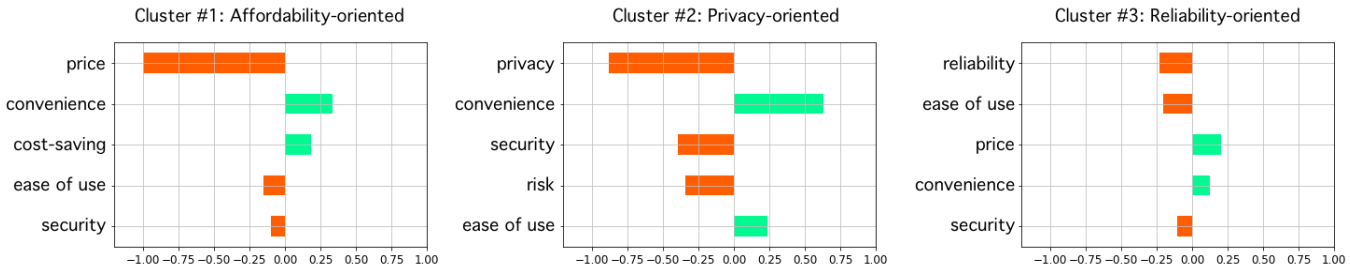


Figure 4: Top five considerations in each of the clusters. Negative and red means the factor is a blocker, positive and green a motivator. Clusters were largely defined by price, privacy, and reliability being blockers.

cross this horizontal line. Clusters are indicated by colors.

Knowing that the optimal number of clusters was three, we clustered participants using k-means with $k = 3$. The clustering resulted in 28% of participants being assigned to the cluster named reliability-oriented, 34% to the cluster named privacy-oriented, and 38% being assigned to the affordability-oriented cluster. We named the clusters based on the sorted absolute average value of factors within each cluster in order to represent the importance of the motivator or blocker. Figure 4 shows the top five considerations for each cluster. The fact that the clusters are largely defined by price, privacy, and reliability blockers suggests that consumers may be segmented with regard to reservations that they may have on these factors. While the average value for the top consideration in the reliability cluster is smaller than the other clusters, a triangulation analysis of quantitative and qualitative data further reinforces the segmentation: reliability is the main and differentiating factor of this cluster (e.g., ease of use is a common factor across all clusters). Also, reliability is not pronounced in the other two clusters, with averages around -0.1, and open-ended responses from this cluster show a recurring theme of reliability, e,g. “fear of malfunction.”

Participants assigned to cluster #1 would see price as a major blocker. For example, if the device is not affordable or too expensive. Participants in cluster #2 would see privacy risks as a major blocker. Finally, participants in cluster #3 might not purchase a device if it is not reliable. For instance, consumers in this cluster would care about how dependable or high-quality the device is, and what happens when Internet connection is lost. These clusters revealed that consumers may approach their decision-making process with different priorities, and that consumers who value price and reliability may not particularly consider privacy as a major factor.

Knowing that the identified clusters were related to privacy, affordability, and reliability being mentioned as blockers, we conducted additional statistical tests in order to understand whether any of the demographics would be associated with participants mentioning price or reliability as blockers. Using the corrected p-value of .00625 (.05/8 demographics), none of the tests came out significant. In other words, we did not find significant relationships between participants mentioning price or reliability and their reported demographics. We also conducted individual multinomial logistic regression analyses where the dependent variable was the cluster and demographics the independent variables, using separate models for each demographic. The results were the same: no statistically significant relationships found between the assigned consumer cluster and people’s demographics.

The percentage of participants in the privacy-oriented cluster who reported not owning a smart home device was 56%, whereas this percentage was 49% for the other two clusters. A Likelihood Ratio Test of a multinomial logistic regression model with the cluster as the dependent variable and whether participants reported owning a device did not produce a statistically significant result.

4.4 Cluster Classification and Decision Tree

We created a decision tree classifier to predict the cluster of each participant and elucidate/reconstruct the decision-making process of participants in each cluster. The goal of this classifier is to be able to segment consumers based on the considerations they might have. For example, one could use our classifier by asking users to select among the factors we identified in our study which ones they consider as motivators or blockers. Separating motivators from blockers in this analysis is important given how consumers may have different considerations in their purchase decisions. For example, while a product not being environmentally-friendly may not be a blocker, being environmentally-friendly may become a motivator. In other words, separating motivators and blockers can uncover more nuanced decisions. Then, based on the selections, a cluster can be assigned to a consumer which will help understand the consumer’s priorities. Does the consumer prioritize price, privacy, or reliability more? The interpretation of this decision tree classifier can uncover how considerations of motivators and blockers can segment consumers.

We initially evaluated a classifier using all motivators and blockers, without specifying a maximum tree depth, with 10-fold cross-validation. This classifier achieved F-1 scores

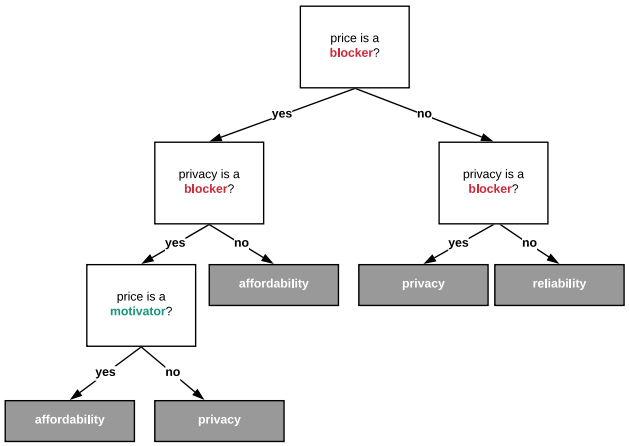


Figure 5: Simplest Decision Tree Classifier to assign clusters based on top considerations. The decision tree suggests that even if privacy is a blocker, consumers may focus on affordability if price becomes a motivator.

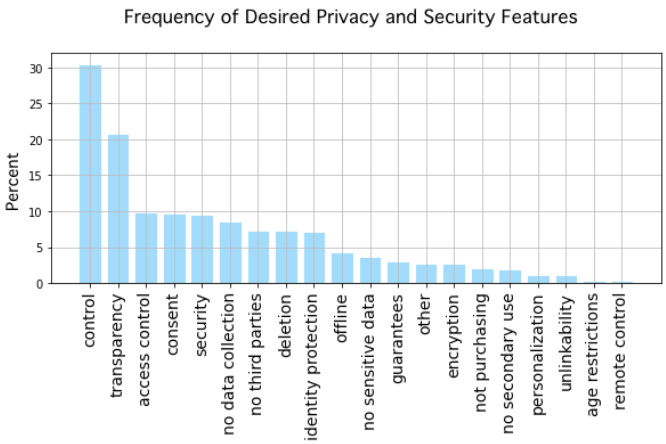


Figure 6: Frequency (%) of desired privacy and security features. The top features are control, transparency, access control, consent, security, no data collection, no third parties, deletion, and identity protection.

from 97% to 98% for all clusters, with a resulting tree depth of 7. The resulting tree included considerations about price, privacy, time-saving, convenience, interoperability, need, ease of use, remote control, and safety. While this classifier helped elucidate the decision-making process with regard to purchase considerations, it was overly complex to interpret.

In order to arrive at a more practical solution – one with good performance and interpretability – we empirically tested different numbers of factors and tree depths using 10-fold cross-validation. The best classifier used only the privacy and price motivators and blockers and a tree depth of 3. F-1 scores for each cluster were 98% for privacy-oriented (Precision=99%, Recall=98%), 100% for reliability-oriented (Precision=100%, Recall=99%), and 99% for affordability-oriented (Precision=98%, Recall=99%). The error rate from 10-fold cross-validation was 1.6%. We then trained this classifier with all of our data and generated the decision tree shown in Figure 5. Based on the process outlined in the decision tree, participants would be affordability-oriented if price is a blocker and privacy is not a blocker or if price is a blocker and affordability is a motivator. Participants would be classified as privacy-oriented if price and privacy are blockers and price is not a motivator, or if price is not a blocker but privacy is. Finally, participants would be classified as reliability-oriented if neither price or privacy are blockers. The fact that the decision tree classifier was able to be trained with only two features and achieve good performance shows that it is likely the decision-making of consumers might rely heavily on price and privacy assessments. For example, the decision tree shows that even if privacy is a blocker, consumers may still be influenced by price if it becomes a motivator.

4.5 Desired Privacy Features

Survey respondents provided 19 unique privacy features they wish were available. The top desired privacy features in the survey were: control, mentioned by 30.43% of participants, followed by transparency (20.75%), access control (9.83%), consent (9.67%), security (9.51%), no data collection (8.56%), no third parties (7.29%), deletion (7.29%), identity protection (7.13%), offline operation (4.12%), no sensitive data (3.65%) and guarantees (3.01%). Figure 6 shows frequencies of all features as percentages of participants who mentioned them, and Table 2 (Appendix) shows all codes, along with examples.

Control Three types of control were mentioned by participants: physical control such as shutting off the devices, controlling what data are collected, and data use opt-out. For example, P99 noted *"Being able to choose exactly which data is being collected and how it is used. Have complete control."*

Transparency Transparency features focused on having the manufacturer/developer show users what data are being collected, whom they are shared with, for what purposes, and whether their system was vulnerable or breached. Participants also mentioned wanting to have such information periodically such as weekly or monthly, and receiving notifications and/or seeing physical visual indicators about data activities. For instance, P183 responded *"I would like full reports on where my data is going from smart home devices sent daily or weekly."*

Access control Participants whose answers hinted at access control features want to be able to have strong passwords, two-factor authentication, and biometrics, as well as to prevent access to their system by anyone else besides themselves. For example P151 said *"to turn on only when activated and had a voice recognition devices so if someone asked for my information it would not display it."*

Consent Consent means that participants wish to be able to know before certain data collection or sharing occurs and be able to allow or deny it. For instance, P175 was concerned about tracking of search habits and noted *"Only allowing tracking of my search history with my consent."*

Strong security Participants who wanted strong security emphasized that they wanted to be able to secure the data in storage and in transit, and make sure that their network was secure. As an example, P220 suggested using firewalls and other software, saying *"I would like to have any device that is used online to have a firewall and virus protection program installed with it."*

No data collection Participants who explicitly mentioned not wanting the device to collect or share any of their data were representative of this category of features. For instance, P227 noted *"Well I don't know at the moment. As long as I am not being tracked and none of my data is being collected, then I would be fine with whatever security or privacy protection tools available out there to keep me safe. Especially in my own home."*

No third parties Participants whose comments fell into this category were explicit about not wanting their data to be shared with any third parties, such as marketing firms or the government. As an example, P270 expressed *"If I were to use one of these devices, I don't want them linked to my identity and I don't want the information shared with third parties. I can see where a utility company might need some of this information in order to bill me properly. But beyond that, I would want a user-friendly interface that allows me to shut off access to anyone else."*

Deletion Participants who emphasized deletion either wanted data to be deleted automatically after a certain period, or have the ability to "go in" and delete any data, either via a user interface or a physical reset button. For example, P228 suggested a feature where she could access and delete information at any point in time, saying *"I'd like to be able to delete things regularly. Permanent deletion."*

Identity protection Participants who wish this feature were explicit about not wanting to have any of the data associated with their identity, and that they wanted specific protections from it. For instance, P217 suggested *"anything to protect my identity, so only the device and I know it."*

Offline mode Offline mode means that participants would want either the device to be offline at all times, or for it not to be online at all times, or for them to be able to control when devices go online or offline. For example, P295 noted *"All smart home devices need to be able to operate offline, without requiring a web app or account with a company. There is no need to gather data and send it to the Internet to operate these devices, they just want to. You should be able to set them up on a local network and control them yourself without them being tied to a brand or company. This helps when Google buys your thermostat company then bricks them."*

No sensitive data Participants in this category would not want any sensitive data to be collected or stored. For example, P85 noted *"Anything that would block personal data from being shared. If I have to enter anything personal to use the equipment, I would want to be able to lock it and that it be never stored. And that I have the say so of what data about me I consider private and personal."*

Guarantees of privacy and security Participants who mentioned this feature wanted to be given guarantees either by the manufacturer or the applicable laws that their privacy and security would be protected and that there would be penalties otherwise. For instance, P117 expressed *"Auto-deletion of pertinent data and guarantee, with legal repercussions, that data will not be shared."*

While some of these features have been uncovered in user-centric studies before, such as control, transparency, strong access control, an offline mode [14, 39], our results show the frequency in which such privacy features were mentioned, which can help developers and regulators navigate what is most important for their users with regard to users' privacy and security considerations.

Finally, we verified whether the most desired privacy features would be different for people in each consumer segment according to our clustering analysis, and the top two features wished for are the same for the three segments: control and transparency, accounting for more than 30% of the responses in all three segments.

5 Discussion

5.1 Summary of Findings

Our findings suggest convenience, ease of use, price, cost-saving, need, remote control, and interoperability are the top motivators for consumers to adopt smart home devices. The top blockers are price, privacy, security, ease of use, reliability, risk, and lack of need. Our analyses showed that participants who see privacy or security as a blocker in purchasing decisions were less likely to own smart home devices at the time of the survey, and that considerations of privacy and security were not associated with demographic traits such as gender identity, age, education or income. Our clustering analysis uncovered three consumer segments with regard to their purchase considerations: affordability-oriented, privacy-oriented, and reliability-oriented. The most desired privacy protections for smart home devices from the survey are: control, transparency, strong access controls, consent, security, no data collection, no third parties, deletion, identity protection, offline operation, no sensitive data, and guarantees.

5.2 Paradox or (Bounded) Rationality?

Our results indicate that consumers heavily weigh privacy and security as blockers, and that these may be preventing them from adopting smart home devices. This finding suggests

that a paradox may not be the explanation for the mixed adoption signals – perhaps privacy-oriented consumers are really not buying smart home devices. On the other hand, our decision tree analysis showed that even if privacy is a blocker, consumers may still value affordability over privacy if the price becomes a motivator. This becomes a potential problem that could lead to undesired outcomes when such devices are sometimes given for free as part of promotions from big tech companies, such as Spotify giving the Google Home Mini to premium subscribers [33], or Google giving the Google Home Mini to Pixel 2 phone buyers [27] and even randomly [17]. In such cases, bounded rationality may lead consumers to overlook privacy considerations, which may become a concern only after a device has been installed [14].

The fact that our decision tree was effective with only the privacy and price feature suggests what consumers may be ultimately considering is whether a low price point for these devices is worth the expected, associated privacy risks. Such price–privacy interplay has been noted in prior work which suggests users would pay a premium for privacy [5, 12, 36], but that even in the context of the smart home, they would pay less than they would be willing to take in exchange for it [5]. Considering our findings in light of prior work, we posit that: (1) privacy-oriented consumers may expect privacy by default or else many may not adopt such technologies, (2) making privacy and security a motivator attached a higher premium may not work well, and (3) if privacy concerns are only a consumer afterthought due to bounded rationality or if a device is extremely affordable, then the market puts consumers in a "privacy-not-included" scenario and thus the value of privacy may be inadvertently or unintentionally diminished.

Paradoxically, one could argue that the knowledge revealed through our analyses can give opportunistic developers more tactical information to manipulate consumers in ways that motivators are highlighted such that privacy considerations remain obscure. This could reinforce bounded rationality and lead to decisions against the consumer's best interest [28]. The lack of visibility of privacy and security-related information on smart home devices is an existing problem, and legislators on this topic have proposed adding concise and accessible labels (e.g., [16, 37]), but little guidance has been provided on how they should be presented. Recent research efforts have focused on how to implement such labels (e.g. [13, 18]) in order to educate and equip consumers and prevent decisions driven by bounded rationality. We endorse such efforts, and based on our segmentation analysis, we further posit that more transparency is needed in order to equip consumers to make informed, rational decisions that suit their specific needs and expectations with regard to the interplay between price, privacy, and reliability. Accordingly, we present several practical recommendations for device developers and policymakers based on our findings which could lead to more informed consumer decisions and meaningful device comparisons.

5.3 Recommendations

5.3.1 Device Developers

Our quantification enables developers to market products based on whether something is more of a motivator or a blocker. For example, marketing strategies may focus on showing empathy about users perceiving adoption coming with a risk of privacy and then offering certain guarantees that their data are not to be shared or used for secondary purposes. Other examples may include highlighting a device's reliability when the Internet goes down, or estimating long-term savings, as this is important for affordability-oriented consumers.

Consumer segmentation based on considerations has further implications for targeting of the products. For example, the reliability-oriented segment may care more about what happens when the Internet goes down, whereas the privacy-oriented segment may care more about whether their data will be shared with third parties or they will be monitored without consent. For this reason, device developers could clarify how devices consider the segment-differentiating factors (i.e., price, privacy, reliability) in order to help consumers find the right device for them among the options available.

The properties that segmented the participants may also be inherently at conflict. For example, if a smart camera is to have onboard object recognition capabilities so that it can work offline both for privacy and reliability reasons (when the Internet goes down), it may end up costing more. Surfacing and communicating such trade-offs to consumers may be a promising strategy and prevent instances of bounded rationality. For instance, our proposed decision tree model could be used by retailers and developers to ask consumers four questions on whether privacy or price are motivators or blockers to them, then determine what their consumer segment is in order to better inform their purchasing.

Some of the privacy features uncovered in our analysis have been identified via qualitative studies before. For example, Yao *et al.*'s co-design study [39] uncovered control, transparency, offline and private modes, and Emami-Naeini *et al.*'s study [14] uncovered strong access controls. We contribute further with a comprehensive list of features presented in our paper, which can guide developers in prioritizing and implementing tools and features that may enable them to appeal to consumers in the privacy-oriented segment. For example, a developer can draw from this list to implement increased control, transparency, strong access controls, consent, security, no data collection, no third parties, deletion, identity protection, offline operation, no sensitive data, and guarantees.

5.3.2 Policymakers

Consumers in the affordability-oriented segment may not have privacy as a strong consideration due to not being educated about Internet business models, which is associated with the aforementioned bounded-rationality issue of IoT devices. The

segmentation we presented could serve as a framework for policymakers to approach the design of privacy regulation in ways that meet consumers where they stand. Ideally, a device would be affordable, privacy-preserving, and reliable. However, special consideration must be given to instances where such desirable properties may be at conflict. For instance, it may be ineffective to implement a privacy feature that poses a trade-off with reliability if a device's target audience is largely reliability-oriented, such as making a smart camera work offline for privacy reasons, yet doing so might limit its reliability such as being unable to recognize objects when offline. Another strategy that might prove ineffective is to offer extra privacy features for an additional premium to affordability-oriented consumers. Notwithstanding, more effort should be placed on regulating the communication of risks involved in owning a smart device (e.g., privacy and reliability), especially when the smart version is cheaper or given away for free as part of promotions. Arguably, more clarity about such risks becomes progressively important as consumers are offered increasingly fewer non-smart device alternatives in the future.

It might also be beneficial for policymakers to introduce requirements for developers to practice tailored consumer education about device privacy/security risks according to individuals' corresponding cluster derived from our decision tree. For example, individuals belonging in the privacy-oriented segment could be given privacy-related device details while seeing only summaries on the other cluster-defining aspects.

5.4 Limitations

Our survey was conducted in the US. This means that the results from our data analyses may not necessarily represent the state of smart home purchase considerations elsewhere, nor it is representative of the universe of considerations that may exist. As it applies to survey data more generally, the data collected may also be subject to the availability heuristic. In addition, privacy considerations, expectations, and awareness are known to be diverse across different geographies and cultures. We do note, however, that smart home growth has been observed heavily in the US [34].

The responses provided by participants to the question about privacy behaviors, tools, and features may carry a bias from the survey design. This is because participants were exposed to different data collection scenarios throughout the survey where they were asked to express their preferences. Nonetheless, we do not find this bias to compromise the quality of our data since every participant was assigned such data collection scenarios randomly, and thus had comparable experiences. We note, however, that had the question been asked prior to the data collection scenarios, perhaps participants would be less aware about certain features, but would have been primed for privacy when responding to the scenarios analyzed in [5]. The scenario questions may also have influenced the IUIPC questionnaire responses.

While our participant sample was diverse with regard to demographics and socio-economical status, it may not be representative of the US population, as participants on AMT may skew towards people with non-traditional forms of employment and people with heightened technical expertise.

5.5 Future Work

We examined consumer considerations about smart home devices in general, but there could be differences in the decision-making across specific devices [14]. For instance, reliability may be more important for a smart lock, whereas privacy may be more important for a camera. Future works could conduct similar analyses considering different devices.

Follow-up experimental studies could be conducted to validate our findings. For example, a potential study may incorporate our decision tree questions to predict which segment consumers belong in (i.e., price, privacy, or reliability) and verify whether the predictions match consumer priorities.

Future studies could explore how interventions could "move" a consumer from one segment to another. In other words, how stable would a person belong to anyone cluster? Would privacy-oriented consumers be more stable than affordability-oriented consumers?

6 Conclusion

Smart home device adoption continues to grow steadily, yet privacy and security concerns reportedly remain a roadblock to mass adoption. User-centric qualitative studies have revealed that consumers often consider price, features, and privacy risks when making smart home device purchases, but no studies have attempted to quantify these considerations in a systematic way. Previous studies have also found that many consumers do not consider privacy at all before purchasing a device. We conducted a mixed-method analysis using online survey data collected from 631 participants based in the US. Our analyses show that privacy and security are considered blockers for half of the participants, but more so for participants who reported not owning a device. We found that convenience, ease of use, price, cost-saving, and need are top motivators and price, privacy, security, ease of use, and reliability are top blockers. We conducted a customer segmentation analysis which revealed three clusters: affordability-oriented, privacy-oriented, and reliability-oriented. A decision tree classifier to predict customer segments revealed that even privacy-oriented consumers may be influenced by a device's price if it becomes a motivator. Finally, we present a comprehensive list of desired privacy behaviors, tools, and features reported by our survey participants. From our findings, we define and discuss implications for the targeting, legislation, and privacy design of smart home devices.

7 Acknowledgments

We thank our survey participants for their insightful and articulated responses. We also thank the anonymous reviewers for their thoughtful comments and suggestions. This work was in part supported by the National Science Foundation (NSF) grant number CNS-1652497.

References

[1] Alessandro Acquisti and Jens Grossklags. Privacy and rationality in individual decision making. *IEEE security & privacy*, 3(1):26–33, 2005.

[2] Alessandro Acquisti, Leslie K John, and George Loewenstein. What is privacy worth? *The Journal of Legal Studies*, 42(2):249–274, 2013.

[3] Idris Adjerid, Eyal Peer, and Alessandro Acquisti. Beyond the privacy paradox: Objective versus relative risk in privacy decision making. *MIS Q.*, 42(2):465–488, June 2018.

[4] Nor Hazlin Nor Asshidin, Nurazariah Abidin, and Hafizzah Bashira Borhan. Perceived quality and emotional value that influence consumer's purchase intention towards american and local products. *Procedia Economics and Finance*, 35(3):639–643, 2016.

[5] Nata M Barbosa, Joon S Park, Yaxing Yao, and Yang Wang. "what if?" predicting individual users' smart home privacy preferences and their changes. *Proceedings on Privacy Enhancing Technologies*, 2019(4):211–231, 2019.

[6] A.J. Bernheim Brush, Bongshin Lee, Ratul Mahajan, Sharad Agarwal, Stefan Saroiu, and Colin Dixon. Home automation in the wild: Challenges and opportunities. In *Proceedings of the 2011 SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*, 2011.

[7] Karen L Courtney. Privacy and senior willingness to adopt smart home information technology in residential care facilities. *Methods of information in medicine*, 47(01):76–81, 2008.

[8] Karen L Courtney, George Demeris, Marilyn Rantz, and Marjorie Skubic. Needing smart home technologies: the perspectives of older adults in continuing care retirement communities. *Informatics in Primary Care*, 16(3), 2008.

[9] George Danezis, Stephen Lewis, and Ross J Anderson. How much is location privacy worth? In *4th Annual Workshop on the Economics of Information Security (WEIS'05)*, 2005.

[10] George Demiris, Brian K Hensel, Marjorie Skubic, and Marilyn Rantz. Senior residents' perceived need of and preferences for "smart home" sensor technologies. *International journal of technology assessment in health care*, 24(1):120–124, 2008.

[11] Serge Egelman, Adrienne Porter Felt, and David Wagner. Choice architecture and smartphone privacy: There's a price for that. In *11th Annual Workshop on the Economics of Information Security (WEIS'12)*, 2012.

[12] Serge Egelman, Janice Tsai, Lorrie Faith Cranor, and Alessandro Acquisti. Timing is everything? the effects of timing and placement of online privacy indicators. In *Proceedings of the 2009 SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*, 2009.

[13] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. Ask the experts: What should be on an iot privacy and security label? In *41st IEEE Symposium on Security and Privacy (S&P'20)*, 2020.

[14] Pardis Emami-Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Faith Cranor. Exploring how privacy and security factor into iot device purchase behavior. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*, 2019.

[15] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[16] Ftc comment to the national telecommunications information administration on "communicating iot device security update capability to improve transparency for consumers", 2017. https://www.ftc.gov/policy/advocacy/advocacy-filings/2017/06/ftc-comment-national-telecommunications-information.

[17] Google is randomly giving away even more free google home mini speakers, 2019. https://www.forbes.com/sites/johanmoreno/2019/09/29/google-is-randomly-giving-away-even-more-free-google-home-mini-speakers/.

[18] Shakthidhar Reddy Gopavaram, Jayati Dev, Sanchari Das, and Jean Camp. Iotmarketplace: Informing purchase decisions with risk communication. 2019.

[19] Jens Grossklags and Alessandro Acquisti. When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information. In *6th Annual Workshop on the Economics of Information Security (WEIS'06)*, 2007.

[20] The trust opportunity: Exploring consumer attitudes to the internet of things - internet society, 2019. https://www.internetsociety.org/resources/doc/2019/trust-opportunity-exploring-consumer-attitudes-to-iot/.

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

[22] Naresh K Malhotra, Sung S Kim, and James Agarwal. Internet users' information privacy concerns (iuipc): The construct, the scale, and a causal model. *Information systems research*, 15(4):336–355, 2004.

[23] Nathan Malkin, Julia Bernd, Maritza Johnson, and Serge Egelman. "what can't data be used for?" privacy expectations about smart tvs in the us. In *European Workshop on Usable Security (Euro USEC'18)*, 2018.

[24] Shrirang Mare, Franziska Roesner, and Tadayoshi Kohno. Smart devices in airbnbs: Considering privacy and security for both guests and hosts. *Proceedings on Privacy Enhancing Technologies*, 2:436–458, 2020.

[25] Emily McReynolds, Sarah Hubbard, Timothy Lau, Aditya Saraf, Maya Cakmak, and Franziska Roesner. Toys that listen: A study of parents, children, and internet-connected toys. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI'17)*.

[26] Patricia A Norberg, Daniel R Horne, and David A Horne. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of consumer affairs*, 41(1):100–126, 2007.

[27] Pixel 2 comes with a free google home mini, 2019. https://www.cnet.com/news/pixel-2-comes-with-a-free-google-home-mini/.

[28] Stefanie Pötzsch. Privacy awareness: A means to solve the privacy paradox? In *IFIP Summer School on the Future of Identity in the Information Society*.

[29] Yu Pu and Jens Grossklags. Valuating friends' privacy: Does anonymity of sharing personal data matter? In *13th Symposium on Usable Privacy and Security (SOUPS'17)*.

[30] Yu Pu and Jens Grossklags. Using conjoint analysis to investigate the value of interdependent privacy in social app adoption scenarios. In *36th International Conference on Information Systems (ICIS'15)*, 2015.

[31] Yu Pu and Jens Grossklags. Towards a model on the factors influencing social app users' valuation of interdependent privacy. *Proceedings on privacy enhancing technologies*, 2016(2):61–81, 2016.

[32] Daniel J Solove. The myth of the privacy paradox. *George Washington Law Review*, 89, 2020.

[33] Spotify premium subscribers can get a free google home mini, 2019. https://9to5google.com/2019/12/11/spotify-free-google-home-mini/.

[34] Smart home - united states - statista, 2020. https://www.statista.com/outlook/279/109/smart-home/united-states.

[35] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford. " i don't own the data": End user perceptions of smart home device data practices and risks. In *15th Symposium on Usable Privacy and Security (SOUPS'19)*, 2019.

[36] Janice Y Tsai, Serge Egelman, Lorrie Cranor, and Alessandro Acquisti. The effect of online privacy information on purchasing behavior: An experimental study. *Information systems research*, 22(2):254–268, 2011.

[37] Plans announced to introduce new laws for internet connected devices, 2019. https://www.gov.uk/government/news/plans-announced-to-introduce-new-laws-for-internet-connected-devices.

[38] Peter Worthy, Ben Matthews, and Stephen Viller. Trust me: doubts and concerns living with the internet of things. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS'16)*.

[39] Yaxing Yao, Justin Reed Basdeo, Oriana Rosata Mcdonough, and Yang Wang. Privacy perceptions and designs of bystanders in smart homes. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[40] Eric Zeng, Shrirang Mare, and Franziska Roesner. End user security and privacy concerns with smart homes. In *13th Symposium on Usable Privacy and Security (SOUPS'17)*, pages 65–80, 2017.

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

[42] Verena Zimmermann, Merve Bennighof, Miriam Edel, Oliver Hofmann, Judith Jung, and Melina von Wick. 'home, smart home'–exploring end users' mental models of smart homes. In *Mensch und Computer 2018-Workshopband*, 2018.

A Appendix

A.1 Survey Protocol

[CONSENT FORM]

The Smart Home

The concept of the smart home involves the control and automation of lighting, heating (such as smart thermostats), ventilation, air conditioning (HVAC), and security, as well as home appliances such as washer/dryers, ovens or refrigerators/freezers.

Wi-Fi is often used for remote monitoring and control. Home devices, when remotely monitored and controlled via the Internet, are an important constituent of the Internet of Things.

Modern systems generally consist of switches and sensors connected to a central hub sometimes called a "gateway" from which the system is controlled with a user interface that is interacted either with a wall-mounted terminal, mobile phone software, tablet computer or a web interface, often but not always via Internet cloud services.

- **Please select 3 images containing a smart home device [QUALIFICATION]**
  o Camera
  o Desk Lamp
  o Blender
  o Smart Thermostat
  o Voice Assistant
  o Smart Bulbs

- What factors do you consider when making decisions about adopting smart home devices? Please answer below.
- **Q1 For example, what are factors that could motivate you to purchase smart home products?** [text entry]
- **Q2 Similarly, what are factors that could keep you from adopting the technology?** [text entry]

The following questions are related to your preferences about collection of certain data by smart home devices. Please proceed when you are ready.

[SCENARIOS 1-4] [repeated four times, half of times purpose was omitted, two of the scenarios included a random device]

- The manufacturer/developer of your ["smart home device" or random device for scenarios 3 and 4] is accessing or inferring [random attribute], for example, [attribute explanation].

  They are using this information for [random purpose], for example, [purpose explanation]

- How do you feel about the data collection in the scenario described above if you were given no additional information about the scenario?
  o Very uncomfortable
  o Somewhat uncomfortable
  o Neither uncomfortable
  nor comfortable
  o Somewhat comfortable
  o Very comfortable

- If you had the choice, would you allow or deny this data collection?
  o Allow
  o Deny

- If you had the choice, when would you like to be notified about this data collection?
  o Never
  o Only the first time
  o Once in a while
  o Every time

- The manufacturer is sharing the data described in the scenario above with third parties (e.g., advertising companies, business affiliates). How do you feel about this?
  o Very uncomfortable
  o Somewhat uncomfortable
  o Neither uncomfortable nor comfortable
  o Somewhat comfortable
  o Very comfortable

- Finally, given the scenario described above, how do you feel about the government having access to this information?
  o Very uncomfortable
  o Somewhat uncomfortable
  o Neither uncomfortable
  nor comfortable
  o Somewhat comfortable
  o Very comfortable

- Please explain the rationale behind your answers [text entry]
- Was there anything unclear in this scenario? Is there a way we can improve the presentation of this scenario? (optional) [text entry]
- **[REVIEW SCENARIOS 1-4]**
- This was the scenario described earlier:

  The manufacturer/developer of your smart home device is accessing or inferring [attribute from respective scenario], for example, [attribute example].

  They are using this information for [purpose from respective scenario], for example, [purpose example]

  You indicated being [comfortable or uncomfortable] with this scenario.

- From the list below, please select the circumstances that could make you [more or less] comfortable. Please select up to three.
  o If the manufacturer was [well known or unknown]
  o If I [gave or did not give] consent to collect data
  o If information was collected [less or more] frequently o If the information involved was [not] sensitive
  o If I could [not] benefit from it (e.g., discounts, serendipitous opportunities)
  o If the information was stored for a [short or longer] period of time, ["then" or "or never"] deleted
  o If the information was [only used for or used beyond] the intended purpose
  o If I was [not] aware of how the data were being used
  o If the data collection was [not] useful for personal and home

safety
o If the data were [not] used for improving products and services
o If the data were [not] used for the common good (e.g., benefit the society at large)
o If I could [not] control the data (e.g., access, copy, and delete)
o If data [not] were handled and secured properly
o Other (please specify) [text entry]

- Please explain why you selected the circumstances above. [text entry]
- Was there anything unclear in this review stage? Is there a way we can improve the presentation of this review stage? (optional) [text entry]
- Please indicate your level of comfort in case your Identity (i.e., who you are) is included along with the data in this scenario. Your original level of comfort was [original comfort level for manufacturer]
  o Very uncomfortable
  o Somewhat uncomfortable
  o Neither uncomfortable nor comfortable
  o Somewhat comfortable
  o Very comfortable

- Please explain the rationale behind the answer above [text entry]
- **Q3 What privacy behaviors you would like to be able to adopt in the context of smart home devices? For example, would there be any privacy-protecting tools, configurations, and techniques you would like to use?** [text entry]

[ECONOMICS-RELATED QUESTION][random assigment to one of four conditions]

Voice assistants take voice commands from users, enabling them to perform various tasks such as listen to music, control video/photo playbacks, and receive news updates. Voice assistants can also enable home automation, allowing users to control smart home appliances through voice commands.

Below is a picture of a popular voice assistant. [voice assistant photo]

In the next step, you will be given a scenario about voice assistants. In this scenario, "personal information" may involve data about your identity, lifestyle, habits, and personal background.

- Consider a scenario where you [are looking to purchase a voice assistant that costs OR had a voice assistant for which you paid] $49. The voice assistant [has OR has little to no] privacy controls and protections against collection and sharing of your personal information
- How much would you be willing to take as a discount off the price tag in exchange for allowing the manufacturer to collect and share personal information in the future? Please specify the amount in dollars [number entry] OR
- How much would you be willing to take as a refund in order to allow the manufacturer to collect and share your personal information? Please specify the amount in dollars [number entry]. OR
- How much would you be willing to pay as a one-time additional fee to add such privacy controls and protections? Please specify the amount in dollars [number entry]. OR
- How much would you be willing to pay extra in order to have more privacy controls and protections such as limited collection and sharing of your personal information? Please specify the amount in dollars (number entry). [number entry]
- Please explain why you would chose this amount. [text entry]

[IUIPC QUESTIONNAIRE (7-POINT LIKERT GRID)]

[DEMOGRAPHICS] We are almost done! Please answer the following questions regarding your demographics.

- What is your gender?
  o Male
  o Female
  o Or specify [text entry]

- What is your age?
  o 18-25
  o 26-35
  o 36-45
  o 46-55
  o 56-65
  o >65

- Do you live in the US?
  Yes
  No

- How many hours do you spend using the Internet every week? [slider entry from 0 to 168]

- Do you currently own a smart home device?
  o Yes
  o No

- How many smart home devices do you currently own? [if answer is yes to previous answer]
  o 1
  o 2
  o 3
  o 4
  o 5
  o 6
  o 7
  o 8
  o 9
  o 10 or more

- What types of smart home devices do you currently own? Please check all that apply [if answer is yes to owning device]
  o Security camera
  o Doorbell camera
  o Baby monitor
  o Pet technology
  o Motion sensor

o Smoke detector
o Leak sensor (water consumption)
o Smart lock
o Door/window alarm
o Garage door
o Smart lighting
o Switch/Plug
o Voice assistant
o Audio/speakers
o Thermostat
o Smart/automation hub
o Other (please specify)

- What is your occupation? [text entry]
- What is the highest level of school you have completed or the highest degree you have received?
  o Less than high school degree
  o High school graduate (high school diploma or equivalent including GED)
  o Some college but no degree
  o Associate degree in college (2-year)
  o Bachelor's degree in college (4-year)
  o Master's degree
  o Doctoral degree
  o Professional degree (JD, MD)

- Information about income is very important to understand. Would you please give your best guess? Please indicate the answer that includes your entire household income in (previous year) before taxes.
  o Less than $10,000
  o $10,000 to $19,999
  o $20,000 to $29,999
  o $30,000 to $39,999
  o $40,000 to $49,999
  o $50,000 to $59,999
  o $60,000 to $69,999
  o $70,000 to $79,999
  o $80,000 to $89,999
  o $90,000 to $99,999
  o $100,000 to $149,999
  o $150,000 or more

- What is the size of your household?
  o 1
  o 2
  o 3
  o 4
  o 5
  o 6
  o 7 or more

- Do you have children?
  o Yes
  o No

- Are you now married, widowed, divorced, separated or never married?
  o Married
  o Widowed
  o Divorced
  o Separated
  o Single

- Do you have any comments or suggestions for this survey? Thanks! [text entry]

A.2 List of Codes

Code	Motivator Example	%	Blocker Example	%
convenience	If I have a need for something to maximize my work output or it makes my life more convenient, I could look into smart devices. If a smart device can help me function more I would use it.	41.20	If the smart device is something that won't drastically improve my life then I would not buy it.	2.55
privacy	If it is non-intrusive	4.12	Mostly, companies obtaining information on my personal life. Just because consumers buy from a company doesn't mean the company can own them.	36.61
price	good prices	26.94	Probably the costs if they're far too high as well as complicated technology.	41.36
security	The security of the system and how protected it is from outside tampering.	8.56	The security of the item, could it be hacked? Could it have a camera that could turn on and be hacked? Would my personal information be safe?	23.61
cost-saving	The money I would save over time on utility bills	20.76	If it is not going to be saving in any large chunk of money in the long run.	3.01
risk	-	-	Safety of personal data and lack of assurance devices are not video taping and recording audio of my every move like I am on reality TV.	13.47
reliability	I want to make sure that there is a "dumb" fallback in case the cloud fails, and I want security.		Single use products. Inexperienced brands. Behavior if they lose connection.	16.80
control	Freedom, meaning that I want these devices to control everything and give me the satisfaction that I want to have.	3.33	-	-
remote control	I'm most motivated by smart home products that can inform me of something in my home while I am away – those products that monitor my home	10.30	-	-
fun	Making life simpler. They're cool and cutting edge. They are fun.	3.01	-	-
time-saving	If it will save me a lot of time. If it can do things for me that I would have never imagined possible. If it reduces the amount of things I have to remember to do on my own.	6.34	Something that doesn't integrate well or reall doesn't save me either time or money	0.32
functionality	Some factors are their ability. I would like to have them be able to generally listen and complete my commands.	7.61	If the technology has a too many unnecessary features that make it an annoyance.	2.54
ease of use	Convenience and ease of use are the definite priorities in getting a smart home device. I want to make sure that it's easy for me to use and that it will save me time and money in my everyday life.	28.37	I would likely not purchase technology if it was exceedingly complicated or difficult to use. I also wouldn't purchase anything too expensive.	19.65
need	If they will fulfill a need that I am currently in need of. If it allows me to be more efficient with my time and money.	10.94	Doesn't do anything I need it to do, is too expensive, is not secure.	9.67
novelty	The novelty factor might motivate me to purchase a smart home device.	2.69	-	-
nothing	-	-	None	2.54
online	-	-	I do not like products that have no need to connect to a network...lights, Thermostats..etc.	2.06
environ.-friendly	For example the biggest factor would be conservation, trying to be efficient and making the best of resources by limiting my use as much as I can.	2.06	-	-
reputation	The company that maintains the data must be trustworthy. The mustn't have a poor reputation for cyber security.	3.01	The reviews being bad	3.80
personalization	Being able to program them to operate when I wanted them to, especially thermostatic products and appliances.	1.58	-	-
safety	Safety. I like the one that works for the oven. sometimes people forget to turn off the oven when going on a trip or think they forgot so it gives you the alleviation of knowing your house is safe.	5.23	Whether their use makes me more vulnerable to home invasion, unfavorable cost-to-savings ratio, whether their use compromises the security of my personal information...	3.01
fear of tech	-	-	Artificial intelligence becoming too smart, if power goes out or internet fails, technology dependant, paranoia or knowing people could hack me	1.27
interoperability	It has to be compatible with my smartphone	9.51	If it's not compatible with other devices. For example if I purchased an Apple product and I currently have devices that aren't compatible with Apple. Everything would need to work together.	5.71
offline	The product would need be able to be used locally and not rely on an internet connection. It would need to keep working if the company went out of business.	0.79	-	-
aesthetics	I would like the product to have a sleek design so that it looks nice in my home. I don't want the product to be an eyesore.	2.85	Lack of physical appeal, don't match my style or decor and too many steps to use	2.22
durability	The durability of the equipment. I would want it to last as long as possible.	1.27	Cheaply made, too complicated to use, not on the market long enough	1.27
other	Better broadband availability in my area. I can't do it before that happens.	2.69	How busy my life is going	1.58

Table 1: List of codes along with examples of when they were considered a motivator or a blocker, ordered by absolute average values of each factor for all of the data when each factor is encoded as 1 for motivator, -1 for blocker, and 0 otherwise. "Other" was assigned to answers where a participant made a statement which we could not assign to any of the codes.

Code	Example	%
control	I would like the choice to control what gets shared and why. And I don't want it to be underhanded.	30.43
transparency	I would like full reports on where my data is going from smart home devices sent daily or weekly	20.76
access control	More password, 2 factor configuration, anything that can make security better	9.83
consent	Letting the customer know ahead of time and asking for the customers consent and just keeping the customer involved as much as possible. i know for me it would make me feel better. Trustworthy companies, in my opinion will communicate with the customer on a regular basis and that in turn would make me feel the safest. i really cannot think of any of behaviors	9.67
security	I would most likely buy a firewall or some type of security to ensure my privacy and safety	9.51
no data collection	A legitimate way to block any data collection, though I doubt that would ever occur.	8.56
no third parties	I would like the smart home devices to work in my home without giving info to third parties. However this can be done should be done.	7.29
deletion	I'd like to be able to delete things regularly. Permanent deletion.	7.29
identity protection	If there was a way to de-identify the information by changing our voices or not attaching a location to the information, I would feel safer using these devices	7.13
offline	I would like for smart home tools to not transmit any user data out of the device. I would like them to be able to download new software but not to upload any of their own collected data.	4.12
no sensitive data	Anything that would block personal data from being shared. If I have to enter anything personal to use the equipment, I would want to be able to lock it and that it be never stored. And that I have the say so of what data about me I considers private and personal.	3.65
guarantees	Opt-outs, control of data, contractual obligations to not use my data	3.01
encryption	Yes, I'd like to have all of my data encrypted, scrambled and rendered useless by any third parties, and the data that is being used should only be used for my benefit and no one else's.	2.69
not purchasing	I wouldn't use smart home devices at all.	2.06
no secondary use	I would like to be sure I could deny or allow any collection or use of 3rd party data collection. No one needs to know that info. For any reason. Other than to enhance my smart home experience.	1.9
personalization	I would like to be able to run my own server for the devices to communicate with – if not all the time, at least to be able to do so as a backup in case the central server gets shut down.	1.11
unlinkability	Most of all, I do not want my smart devices linked to an already existing account, i.e., I want the smart devices to have a separate account from which I can control my devices, and those devices will not have access to personal or sensitive information.	1.11
age restrictions	Yes. There would be privacy locks for children, privacy feautures that only adults can access, and privacy features having to do with keeping the home more safe, and keeping our data secured.	0.32
remote control	I would like to be able to remotely check on my house without anyone having access to those data.	0.32
other	Smart home devices can make my life easier. They do not malfunction and can last a long time without being replaced in the future. As for configuration of the device, I would just make it easier for the user to have and practice with.	2.69

Table 2: Codes ordered by frequency from privacy tools/behaviors question with percentage of respondents who mentioned them. "Other" was assigned to answers where a participant made a statement which we could not assign to any of the codes.

Enhancing Privacy through the Visual Design of Privacy Notices: Exploring the Interplay of Curiosity, Control and Affect

Agnieszka Kitkowska
Karlstad University

Mark Warner
Northumbria University

Yefim Shulman
Tel Aviv University

Erik Wästlund
Karlstad University

Leonardo A. Martucci
Karlstad University

Abstract

When people sign-up to new online services, privacy notices are the initial means by which data handling practices are communicated. Yet, their design seldom ensures users' privacy comprehension or provides people with privacy choices, resulting in negative feelings associated with the sign-up process. In this paper, we investigate how to improve privacy notice design to enhance privacy comprehension and control, while inducing more positive feelings towards these notices. In an online experiment ($N = 620$), we examine the factors of curiosity, privacy concerns, trust, and time. We study how these factors and visual designs of notices (framing and control) influence privacy comprehension, intention to disclose, and affect (negative-positive). Our results show that, depending on an individual's level of curiosity, control can influence privacy comprehension, disclosure, and affect. We demonstrate that affect moderates the relationship between privacy concerns and disclosure. We elaborate on our results, highlighting how privacy notices that activate curiosity and provide control, could enhance usability and strengthen privacy-conscious behaviors.

1 Introduction

Privacy issues are on the rise since people's daily activities have become increasingly reliant on internet-connected applications. Such accelerating technological dependency may increase personal information disclosure and data collection; furthermore, it can put people's privacy at risks resulting in harms to individuals [61]. Privacy breaches have been often reported in the media (e.g., [22, 29]), and people became concerned about their online information and how they lack control over its use [17].

Regardless of their concerns people are often required to sign-up to new online applications in order to gain access to services. During the application sign-up process people have to make one of the first decisions about their online privacy, seldom provided with a choice to restrict their disclosures. Moreover, privacy is a complex and context-dependent notion [49], and people may disregard it. Without sufficient understanding and control around personal information during the application sign-up process people's decisions around information disclosure may not be informed or rational. Moreover, they may result in behaviors which contradict the privacy beliefs of individuals.

Policymakers try to improve this by enacting data protection regulations, e.g., the EU General Data Protection Regulation (GDPR) [16], or the California Consumer Privacy Act (CCPA) [7]. These require companies to provide users with adequate information to promote informed consent. Yet, little change has been applied in the visual display of privacy notices communicating data handling practices.

To address this, the research presented in this paper focuses on the visual display of privacy notices. We investigate ways of encouraging users to make more informed privacy decisions which are aligned with their beliefs. We examine the role of framing and control within the design of privacy notices as they have been previously shown to influence privacy decision-making [2, 5, 23]. We also draw on prior work [13], to understand how affective state (negative-positive valence), as well as stable factors (curiosity, privacy concerns, trust) influence privacy interactions. We explore how these factors affect privacy comprehension and intention to disclose.

This work has two main contributions. First, our findings demonstrate that providing users with control can lead to more privacy-aware information disclosures. However, control by itself may be insufficient, as curiosity influences the relationship between control and disclosure. Control may also have an effect on users' affective state (valence). Therefore, we

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

propose incorporating methods to enhance both control and curiosity in the design of privacy notices. Such designs have the potential to improve privacy and disclosure decisions during the application sign-up process, as well as lead to better usability through elevated levels of user satisfaction.

Second, our research contributes to the body of knowledge on privacy decisions. We demonstrate that affective state (valence) can moderate the relationship between trust and privacy concerns and, as a result, indirectly affect information disclosure. Such knowledge can be used in future experimental designs and studies modeling privacy decisions.

2 Background

To clarify our motivations and introduce our research questions, in this section we present past theoretical and empirical research fundamental for our work.

One of the frameworks explaining the relationships between different factors influencing privacy-related decisions is the APCO model (Antecedents→Privacy Concerns→Outcomes) [13]. We utilize this framework as it is comprehensive, and draws upon previous multidisciplinary research on privacy; the proposed model conceptualizes factors that influence outcomes of privacy decisions. Among the elements incorporated in the APCO model are antecedents of privacy concerns, such as individual characteristics; next, in the centre of the framework is the relationship between trust and privacy concerns; the central part of the model relates directly to behavioral outcomes (e.g., disclosure). The recent revision of the APCO model broadened its scope, and incorporated the level of effort that may be influenced by mental shortcuts and heuristics (e.g., affect). The level of effort relates to dual-process theories, wherein cognition contains two types of processing [18]. Type 1 is low-effort, fast, automatic, and relies on pre-existing mental models and experiences. Type 2 requires high levels of cognitive effort, is less automatic and therefore, a slower form of cognitive processing.

The current research investigates the relationships between the factors mentioned above, which relate to the low-effort, Type 1 cognitive processing. We examine them in the context of two outcome variables: privacy comprehension and intention to disclose. Further, we study the effect that external factors—framing and control—might have on these outcome variables.

2.1 APCO factors

In the APCO model, information disclosure is one of the behavioral outcomes of privacy decision-making. As demonstrated in the meta-review by Gerber et al. [21] privacy concerns and trust can be predictors of intention to disclose. The APCO framework also proposes that affect might have a moderating role in the relationship between attitudes and behavior, and that individual characteristics might have an indirect effect on behavioral outcomes.

2.1.1 Privacy Concerns

Privacy concerns are considered an attitudinal factor influencing decision-making, and they were investigated in many studies (e.g., [13, 60]). Some of the studies focusing on privacy concerns addressed the *privacy paradox*, meaning the phenomenon when people may express high levels of privacy concerns whilst also tending to over-disclose their personal information [6,51]. However, the findings of the *privacy paradox* research are inconclusive. In one study privacy-concerned people were found to disclose less [14], whilst in another study, this finding existed only under certain conditions, e.g., when perceived damage and enjoyment might have altered the relationship between concerns and disclosure [10]. On the other hand, Taddicken [65], in the context of the social web, found privacy concerns having little to no effect on self-disclosure.

2.1.2 Trust

Past research has shown that people use trust beliefs in the decision-making process around information disclosure [39]. Trust has primarily been found influential when the decision is made under uncertainty, as is frequently the case when people make decisions around online privacy [53]. Trust may also influence "rationally" calculated privacy decisions, e.g., users involve their trust beliefs in the context of sensitive information disclosure [10]. Visual cues might alter trust, e.g., in their study, Zhang et al. [71] showed that cues displaying "instant gratification" (financial reward for registration) decreased trust towards a website. On the other hand, visual cues granting control over the information, combined with salient information about how data might be used for advertisement, were found to increase trust towards the application provider [69]. Consequently, such cues seemed to positively impact the willingness to install applications, which could result in increased information disclosure.

2.1.3 Affective state

Decisions around privacy have also been investigated through the lens of biases and heuristics that may take over the rational, in an economic sense, decision-making. One of the approaches explaining the "irrational" decisions is the affect heuristic related to information processing. There is not much of a consensus about the definition of affect, and the current work follows the description from Lerner: "the superordinate umbrella of constructs that involves emotion, mood, and emotion-related traits" [42, p. 801]. Further, we recognize the circumplex components of affect: valence (positive-negative) and arousal (high-low) [55].

According to the affect heuristic, people add either positive or negative value to their decision outcome [20]. The affect-as-information hypothesis postulates that emotions are felt, and this feeling has a significant impact on cognitive processing, providing conscious information from unconscious appraisal situations [9]. These feelings can guide immediate actions. Similarly, the feelings-as-information theory proposes that positive affect indicates if a given situation is safe [58]. Negative affect signifies that a situation is unsafe, and more cognitive processing is needed. Therefore, positive affect may serve as an incentive to rely on internal thoughts and inclinations, whereas negative affect should direct attention to new, external information. The affect may be elicited by an external stimulus, such as the way information is presented or semantic context, in which the situation takes place [58].

In the context of privacy, affect has been shown to shape risk perceptions [38]. It has a lasting consequence on privacy beliefs, e.g., in an e-commerce environment [43]. Further, negative valence may increase privacy attitude and decrease sharing, while positive valence may increase sharing attitude and decrease privacy attitude [11]. In the current work, we want to further investigate the affect by asking the following research questions:

RQ1 Does the visual design of privacy notices (framing and control) influence affective state?

RQ2 What is the role of the affective state (if any) in the relationship between attitudinal factors and intention to disclose?

2.1.4 Individual characteristics: curiosity

In psychology and behavioral research, curiosity is regarded as one of the stable personality characteristics that drive how people perceive the world, and how they make judgements and decisions [44]. To the best of our knowledge, not much of attention has been given to curiosity in privacy research. Curiosity is closely related to learning and knowledge acquisition. Information-gap-theory proposes that curiosity is "arising when attention becomes focused on a gap in one's knowledge" [44, p. 87]. In consequence, it makes an individual curious and motivates them to seek more information. Hence, curiosity may play the role of a marker, the reference point that encourages an individual to obtain more information. Curiosity might be stimulated by external factors and reduce uncertainty about current circumstances [24, 44].

Considering scarce research about the interplay of curiosity and privacy interactions, we raise the following research question:

RQ3 Does curiosity influence privacy comprehension?

2.2 External factors

Past work investigated privacy comprehension in many contexts, e.g., mobile permission warnings, data visualizations, end-user licence agreements [19, 36, 68]. The results showed that visual representation might impact comprehension. For instance, supplementary information may lead to a higher understanding of data collection practices [14]. Further, the visual cues with salient privacy information can not only improve understanding and increase privacy awareness, but also enhance management of privacy permissions and influence information disclosure [38].

2.2.1 Framing

One of the approaches applied to investigate privacy interactions is framing, meaning that the frame of a decision is designed in a way that constrains how the problem is presented to the decision-maker [47]. Such framing is expected to influence the decision outcome. The framing was used to improve risk communication, and help with pro-privacy decisions (e.g., choice of application, protective attitudes and behaviors) [2, 54]. Furthermore, positive framing successfully nudged users towards less privacy-invasive actions [8]. Emotion eliciting images were shown to influence decisions: the more affective the images, the more weight was placed on impression formation and decision-making [59].

2.2.2 Elements of visual design

Studies demonstrated that visual stimuli might influence memory, when they incorporate animations, anthropomorphic designs, clear layouts, such as division into columns [63, 66]. In the context of privacy, research revealed that the end-user agreements presented in abbreviated style, divided into short sections, elicited positive attitudes, increasing comprehension and time of exposure [68].

Past work suggests that text insufficiently communicates privacy information, and other approaches are required to enhance usability [3]. Nevertheless, visual design needs to be carefully crafted to avoid the effects of cluttered or over-symbolic representations. Anthropomorphic designs were shown to increase personal information disclosure [4, 48]. Moreover, comic strips were found to enhance users' attention [64]. Comics may trigger emotions, enabling a greater understanding of the displayed issues [50].

2.2.3 Control over information disclosure

Prior research suggests that people want to have control over their personal information [5, 38]. Therefore, some researchers provided participants with control and investigated whether it influenced disclosure. The results revealed that control may not necessarily lead to a decrease in information disclosure [5].

People appeared to alter their willingness to disclose in response to non-normative factors (control over publishing their data), but failed to change their behavior in response to the normative factors (e.g., personal identification). As mentioned above, control embedded in visual design, supported by salient information may result in a decrease in disclosing behavior, depending on the context [69].

Considering the effects of visual design on privacy comprehension and information disclosure, in the current work, we want to investigate such a relationship further. Mainly, we aim to examine the role of visual design that incorporates framing and control, in the context of interaction with privacy notices. Therefore, we propose the following research questions:

RQ4 Does visual design of privacy notices (framing and control) affect comprehension?

RQ5 Does visual design of privacy notices (framing and control) affect intention to disclose information?

3 Method

To answer our research questions we designed an online experiment. The experiment aimed to elicit affective states through framing and control applied in the visual design of privacy notices. The experiment contained four phases: entry questionnaires, interactive application sign-up task, measurement of outcome variables, and exit questionnaires (Figure 1).

Our experiment was designed drawing on findings from prior exploratory studies, which we discuss first.

3.1 Exploratory studies

Two exploratory studies preceded the current research, and were used to inform the current study design. Both exploratory studies used freely available platforms to recruit participants, such as the Reddit r/SampleSize.

In the first exploratory study (ST1) [37], we examined why people agree or disagree with privacy notices, and whether the framing of the notice's design elicits changes in affective state. To frame these privacy notices, we used positive and negative anthropomorphic or human-like illustrations. Based on responses from 88 participants, we found that people lack control when acknowledging notices, feeling that they have no other choice but to agree. Drawing on this finding, in the current study, we investigate the effect of control on privacy interactions.

Further, the results of ST1 showed that notice's design alters affect. However, we identified that the framing effects differ, depending on the use of anthropomorphic or human-like illustrations. To clarify the framing effects around these illustrations, we ran the second exploratory study. The study had 36 participants, and each participant was shown 16 images: eight positively and eight negatively framed. Framing groups contained four anthropomorphic and four human-like illustrations, each. The images were arranged in a randomized sequence, and participants were asked to state what feelings they associated with each illustration. We assessed feelings through an instrument similar to the 2D EmojiGrid [67], scoring from 1 (negative) to 5 (positive). For the current study, we chose the positive and negative anthropomorphic representations, because they had the highest and lowest means, respectively (Table 1).

Table 1: Means and standard deviations for positive and negative framing from the second exploratory study (sorted by means, descending).

Illustration, and framing type	M	SD
Anthropomorphic, positive	2.69	0.85
Human, positive	2.52	0.79
Human, negative	2.35	0.69
Anthropomorphic, negative	2.28	0.77

3.2 Main study

In the main study, we applied a 2×2 between-group design. The between-group variables were (1) framing and (2) control. Further, we measured the constructs presented in Figure 1.

3.2.1 Variables

Framing (positive vs negative), and control (present vs absent) were our independent variables, manipulated between groups. Reported curiosity level of participants (high, low) was also included as an independent variable, but was not manipulated. Outcome variables were post-stimulus measurements of affect, intention to disclose, and privacy comprehension.

We included covariates (privacy concerns, trust, approximate time spent on the notice page, and pre-stimulus affective state) to control for their influence on the dependent variables.

3.2.2 Ethical review

The experimental design underwent an ethical review from the Karlstad University Ethical Review Board. The review board determined that this work would not expose participants to any undue risk. To comply with the legal requirements, the researchers made an effort to minimize data collection and reduce the probability of identifying an individual. No personal information was requested from the participants. However, where participants identified themselves (e.g., sent an email), their data was anonymized after the data collection had been completed.

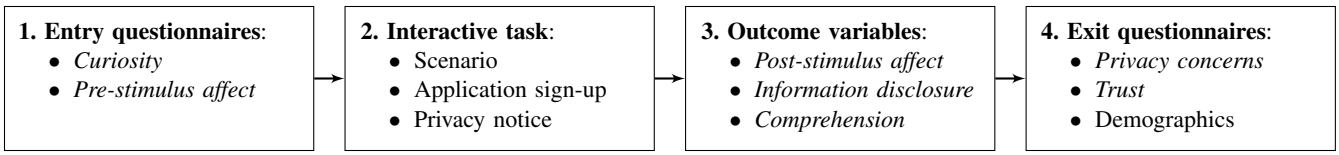


Figure 1: The four phases of the study. The psychological constructs examined in the experiment are italicized.

3.2.3 Study order

Before the start of the experiment, each participant was presented with an informed consent form containing the details of the study, and explaining data handling practices. After the acknowledgement of the consent, participants were redirected to the experiment.

1. Entry questionnaires. First we measured curiosity using a previously validated scale obtained from Kashdan et al. [34] (Appendix A). The scale measures two dimensions of curiosity: *stretching* and *embracing*. Stretching is having the motivation to seek new experiences and knowledge, whilst embracing is the general intention to embrace new, unexpected and unpredictable aspects of everyday life. The construct was measured on a five points Likert scale, and the participants were asked to state to what extent the presented sentences reflected the way they would behave or feel (1 - slightly, 5 - extremely; Appendix A).

Next, we measured affect before displaying privacy notices to ensure that the stimuli elicited its expected affective state. We utilized the validated Affective Self Report (ASR) scale to estimate levels of valence and arousal [31] (Appendix B). We chose this scale, because it is time-efficient to use, and past research shows its performance is comparable with some physiological measures (e.g., thermography and electroencephalography) [31]. ASR consists of 10 semantic-differential items (five for valence and five for arousal), for example, *Unpleasant – Pleasant* (Appendix B).

2. Interactive task. Next, participants were asked to complete an interactive task. We asked them to imagine they were signing-up to use a well-being application (improving mental and physical health) that contained social features. Next, we asked participants to complete a sign-up form requesting their name, username, and password. We did not collect this information; it was only visible to the participants in their browser to increase the ecological validity. In the task instructions, we informed participants that such personal information would not be collected.

After completing the sign-up form, participants were presented with a notification prompting them to review the application's privacy policy. At this point, each participant was randomly assigned to one of the four study conditions, each presenting a different privacy notice design (for details see the Appendix 6).

The designs were as follow: control with positive framing; control with negative framing; no control with positive framing; no control with negative framing (Figure 2). For participants who were provided control, we used toggle like switches which were defaulted to "Enabled". Participants were able to adjust some of the privacy settings using these switches (e.g., decide whether they wanted to share data, connect data with social networks). In the framing conditions we used anthropomorphic images accompanied with text. The Gunning's Fog Index of readability for the text of privacy policy was 10.75 for control absent, and 10.88 for control present groups, indicating that it should be readable by high school sophomores [70].

3. Measurement of outcome variables. We took a second measurement of affect, and asked participants whether their affective state had changed. Particularly, whether they felt more or less positive or negative, using the same ASR instrument.

Next, we wanted to learn how much of the information participants would be willing to disclose. To do this, we built an instrument based on the scale created by Joinson et al. [32]. The questionnaire contained 14 items, and asked participants whether they would share different types of personal information (Appendix C).

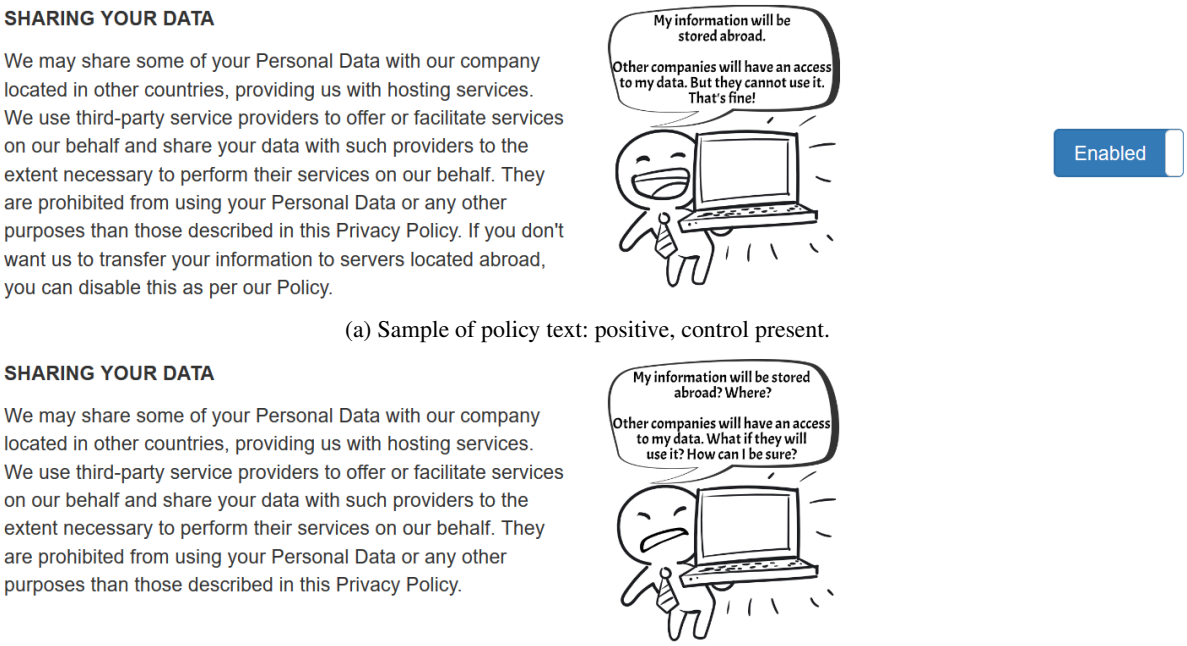
For measuring the participants' comprehension, we created a quiz-like instrument with statements describing information included in the privacy notice (Appendix D). They were presented with ten sentences and asked to declare whether the statement was "True", "False", or "I don't remember / I don't know."

4. Exit questionnaires. We measured privacy-related beliefs, using previously validated instruments. Specifically, we obtained privacy concerns and trust beliefs scales from Malhotra et al. [45] (Appendix F and E).

Last, we asked participants about their demographic characteristics: age, gender, nationality, and education.

3.3 Participants

We used an online crowd-sourcing platform, Prolific, to gather participants. The platform enabled us to compensate participants for their work (£9.82/hr). We wanted to gather participants from English speaking countries, and Prolific's participants pool contains mostly respondents from the UK. To participate in the study, each respondent had to read and agree



(a) Sample of policy text: positive, control present.

(b) Sample of policy text: negative, control absent.

Figure 2: Excerpts of privacy notifications. The notifications' texts were identical between groups, with an exception in the group provided with control (instructional text and the last sentence in the paragraph containing the toggles; Appendix G). The facial expressions of anthropomorphic figures and the accompanying texts were used to elicit positive and negative framing (Appendix G.2).

with the informed consent. Only participants 18 years old or more were allowed to participate in the study.

In total, we received 650 responses. After cleaning the data and removing univariate and multivariate outliers (Mahalanobis distance), the final data-set included 620 cases. The participants were predominantly female (59%); mostly from the UK (74.2%). The respondents were mostly educated (36.1% with Bachelor's degree) and predominantly young (39.5% between 25–34 years old). The detailed demographic characteristics are presented in Table 2.

4 Results

In this section, we first discuss the validity and reliability of instruments applied in the current work. Next, we present the main results concerning the research questions.

4.1 Validity and reliability of scales

To increase the validity and reliability of our study, when possible, we utilized validated instruments acquired from past research. We checked reliability with statistical tests (factor analysis and scales' reliability estimated with Cronbach's α measurements of internal consistency).

Intention to disclose. We asked participants to what extent they would be willing to disclose different types of information. In total, there were 14 types of information, e.g., name, health-related data, or personal economic situation (Appendix C). To score, participants could choose one of two options: "I would disclose" (1) or "I would prefer not to say" (0). Internal consistency of the scale was acceptable ($\alpha = 0.90$). To compute the variable, we summed the scores.

Privacy comprehension. Privacy comprehension was measured as the awareness of information displayed in the privacy notice. The scale consisted of 10 statements associated with the information included in the privacy policy, focusing on information highlighted in framing messages (Appendix D). Participants were asked to state whether each statement was "True", "False", or select "I do not remember / I do not know". Correct answers scored 1, while incorrect, and cases where participants selected the latter option, scored 0. The latter option was presented to the participants to reduce the potential effects of guessing. Because the instrument aimed to measure knowledge, not a latent construct, we could not check Cronbach's reliability. The variable was computed as the sum of correct answers.

Privacy concerns and trust. Both traits were assessed with instruments obtained from Malhotra et al. [45]. The trust beliefs scale contained five items that aimed to measure general attitude towards online companies (Appendix E). Similarly,

Table 2: Detailed demographic characteristics of the sample.

Demographic		N	%
Gender	Female	366	59
	Male	243	39.3
	Other/Self identify	4	0.6
	Prefer not to say	7	1.1
Age	18-24	103	16.6
	25-34	245	39.5
	35-44	135	21.8
	45-54	74	11.9
	55+	63	10.2
Nationality	UK	460	74.2
	USA	139	22.4
	Ireland	11	1.8
	Other	10	1.6
Education	No school/School, no diploma	21	3.4
	High school	101	16.3
	College credit, no degree	92	14.8
	Professional/associate degree	76	12.3
	Bachelor's degree	224	36.1
	Master's degree	96	15.5
	Doctorate degree	10	1.6

the privacy concerns scale measured the general approach to online privacy (Appendix F). The two instruments contained seven-point scoring answers, anchored from "Strongly disagree" to "Strongly agree". Both scales underwent the same procedures during which we ran principal component analysis (PCA), and checked Cronbach's reliability. Privacy concerns did not load strongly into one factor, and after revision, two items were deleted. Both scales had good internal consistency, privacy concerns $\alpha = 0.82$, and trust $\alpha = 0.91$.

Affective state. We measured affective state with ASR to distinguish two dimensions: valence and arousal [31] (Appendix B). To confirm whether the instrument measurements were correct, we first ran PCA. In the case of both pre-, and post-stimulus data, the PCA did not load correctly. The valence loaded strongly into one factor, and its reliability scores were acceptable. Both pre-, and post-stimulus scores of internal consistency were acceptable, $\alpha = 0.91$. We created a pre-, and post-stimulus valence variables by computing the mean score for each scale. These new variables were employed in further analysis. The inappropriate loadings of the items measuring arousal undermined the scale's validity and reliability. Hence, we excluded the arousal scale from further analysis.

Personality characteristic: curiosity. Few curiosity related constructs might be measured. One of them is the Need for Cognition (NFC) — "a dispositional variable reflecting the tendency toward thoughtful analysis and reflective thinking" [62]. In our work, we did not use NFC as it relates to complex, effortful decisions [15, 33], while our research was focused on Type 1 processing. We measured a related, but broader construct of curiosity with an instrument comprising of ten items, acquired from Kashdan et al. [34] (Appendix A). The original scale intended to measure two dimensions of curiosity: *stretching* and *embracing*. We ran PCA to confirm whether the items load correctly. Unfortunately, they did not. Instead, the *stretching* facet loaded strongly to one dimension, while *embracing* loaded to both. Both scales had good internal consistency (*stretching* $\alpha = 0.81$; *embracing* $\alpha = 0.84$). Because of unreliable loadings, in further analysis, we used only the curiosity stretching dimension. We used means to compute the curiosity variable. To apply it as an independent variable, based on the median value, we divided curiosity into a two-level categorical variable (low vs high).

4.2 Main results

Our data analysis is structured around the outcome variables. Hence, first, we present tests' assumptions and correlations to explain our statistical choices. Next, we present models explaining privacy comprehension (**RQ3**, **RQ4**), intention to disclose (**RQ1**, **RQ5**), and the role of affect (**RQ2**).

We checked parametric assumptions which were good, with slight violations of normality, acceptable in large samples. To establish whether the variables included in the experiment are related, we ran the Pearson correlation analysis. We considered an additional variable in the correlation analysis: time spent on the page displaying privacy notice. The test results revealed mostly small to moderate correlations between some of the variables. Table 3 presents the correlations' details.

In order to check changes in valence, we compared pre-, and post-stimulus scores. We used pairwise t-test to investigate changes. There was a significant difference in the scores for pre-stimulus ($M = 3.63$, $SD = 0.85$) and post-stimulus ($M = 2.71$, $SD = 0.65$) valence; $t(619) = 22.62$, $p < 0.001$, $d = 0.91$. Therefore, we presumed that either framing or control had influenced shifts in valence.

To investigate the research questions, we applied different statistical methods: univariate and multivariate analyses of covariance. We checked tests' assumptions, such as univariate and multivariate normality, outliers (Mahalanobis distance), linearity, absence of multicollinearity (correlation tests), homogeneity (Box's M and Levene's tests), homoscedasticity (scatterplots).

In further analysis, we compared between- and within-group effects based on the independent variables. The group sizes differed as presented in Table 4.

4.2.1 Effects on comprehension

To select an appropriate test (univariate or multivariate) we examined correlations. Low correlations ($r < 0.20$) imply that variables should be investigated separately, while moderate correlations (r between 0.20 and 0.50) imply that variables should be analyzed together [12] . Correlations between comprehension and disclosure, as well as between comprehension

Table 3: Correlations between variables: curiosity stretch (CUR), valence pre-stimuls (VAL_PR), valence post-stimulus (VAL_PO), privacy comprehension (COMP), intention to disclose (DIS), privacy concerns (PCS), trust and time spent on policy page. ** significant at 0.01 level; * significant at 0.05 level.

	CUR	VAL_PR	VAL_PO	COMP	DIS	PCS	TRUST	TIME
CUR	1	0.23**	0.04	0.09*	0.08*	0.08*	0.00	-0.00
VAL_PR		1	0.16**	0.10*	0.07	0.11**	0.13**	0.10**
VAL_PO			1	-0.14**	0.24**	-0.06	0.31**	-0.17**
COMP				1	-0.08*	-0.05	-0.03	0.50**
DISC					1	-0.25**	0.30**	-0.14**
PCS						1	-0.19**	0.19
TRUST							1	-0.11**

Table 4: Number of participants per independent variables.

	Presence of control		Framing		Curiosity		Control over information	
	Present	Absent	Positive	Negative	Low	High	Adjusted	Not adjusted
Frequency	318	302	310	310	317	303	190	128
Percent	51.3	48.7	50	50	51.1	48.9	59.7	40.3
Total	620		620		620		318	

and post-stimulus valence were small (Table 3). Hence, to study comprehension, we used univariate analysis of covariance (**RQ3**, **RQ4**).

The Levene's test was good, $p > 0.05$. The model included three independent variables: framing, control, and curiosity; and four covariates: time spent on the policy page, post-stimulus valence, privacy concerns, and trust. We found a significant small between-subject effect of curiosity on comprehension, $F(1,608) = 8.47, p = 0.004, \eta p^2 = 0.01$. The results show that comprehension was significantly higher among the participants with high curiosity ($M = 5.82, SD = 0.10$) than among the participants with low curiosity ($M = 5.37, SD = 0.10$). Further, the time spent on the page with privacy policy had a significant effect on comprehension ($p < 0.001, \eta p^2 = 0.25$).

To further investigate the effects of control on comprehension, we repeated the univariate test only on the data from the participants who were provided with control. For this purpose, we have created a new categorical variable, splitting participants into two groups: the participants that adjusted settings, and the participants that did not adjust them. As a result, the total sample size decreased to 318 participants.

We ran the test with the same parameters. There were significant small effects of curiosity, $F(1,306) = 7.87, p = 0.005, \eta p^2 = 0.02$, and of control, $F(1,306) = 11.11, p = 0.001, \eta p^2 = 0.03$, on comprehension. Again, the participants scoring high on curiosity scored significantly higher on comprehension ($M = 5.79, SD = 0.15$) than those with lower curiosity ($M = 5.21, SD = 0.14$). Similarly, comprehension was significantly higher among the respondents that changed their settings ($M = 5.90, SD = 0.14$) than among the participants who did not use controls ($M = 5.10, SD = 0.17$). Additionally, two covariates had a significant effects on comprehension: time spent on the policy page ($p < 0.001, \eta p^2 = 0.19$) and privacy concerns ($p < 0.05, \eta p^2 = 0.01$).

4.2.2 Effects on affect and intention to disclose

To investigate affect (valence) and intention to disclose (**RQ1**, **RQ5**), we ran a multivariate analysis of covariance. The independent variables were framing, control, and curiosity; covariates were pre-stimulus valence, time spent on the policy page, privacy concerns, and trust. The Box's test was good, significant, but at the level $p > 0.01$, which is acceptable for larger samples. The Levene's test for both outcome variables was good, $p > 0.05$.

There was a small but significant main effect of control on combined dependent variables, $F(2,607) = 2.89, p = 0.05, \eta p^2 = 0.009$, Wilks' $\lambda = 0.99$. The between-subject test confirmed that post-stimulus valence significantly differed among the control groups, $F(1,608) = 5.78, p = 0.01, \eta p^2 = 0.009$. Valence scores were significantly higher for the participants provided with control ($M = 2.78, SD = 0.03$) than for those who did not have a control ($M = 2.65, SD = 0.03$). Further, the model resulted in interaction effect between control and curiosity on the combined dependent variables, $F(2,607) = 3.60, p = 0.02, \eta p^2 = 0.01$, Wilks' $\lambda = 0.98$. The univariate analysis identified the interaction effect for post-stimulus valence, $F(1,608) = 7.19, p = 0.008, \eta p^2 = 0.01$) (Figure 3). The mean scores for valence were higher among the participants provided with control who scored higher on curiosity ($M = 2.84$) than among the participants who scored lower on curiosity ($M = 2.72$). However, the participants with high curiosity not given control scored

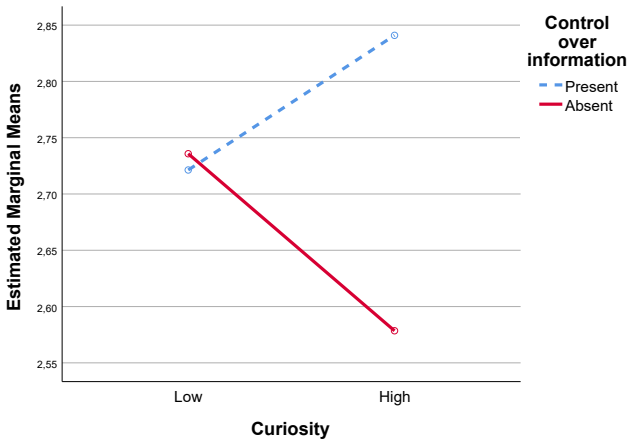


Figure 3: Interaction effect on post-stimulus valence. Covariates appearing in the model are evaluated at the following values: pre-stimulus valence= 3.63, time on policy page= 84.55, privacy concerns= 4.78, trust= 3.10.

lower in valence ($M = 2.73$) than those with lower curiosity ($M = 2.57$).

The covariates had significant effects, $p < 0.001$. Particularly, time spent on the policy page and trust affected outcome variables; pre-stimulus valence influenced post-stimulus valence; privacy concerns significantly affected intention to disclose.

Consistent with the tests of comprehension, we re-ran the analysis on the smaller sample, considering only participants provided with control. Both Box's and Levene's tests were insignificant, $p > 0.05$.

The multivariate test results indicated small significant effect of curiosity, $F(2,305) = 2.99, p = 0.05, \eta p^2 = 0.01$, Wilks' $\lambda = 0.98$. However, the univariate tests results did not confirm it. Further, the multivariate test revealed a small effect of adjusted settings, $F(2,305) = 7.55, p = 0.001, \eta p^2 = 0.04$, Wilks' $\lambda = 0.95$. The univariate test confirmed that the groups differed in intention to disclose, $F(1,306) = 12.68, p < 0.001, \eta p^2 = 0.04$, which was significantly higher among the participants who did not adjust settings ($M = 9.66, SD = 0.37$), than among those who adjusted them ($M = 7.84, SD = 0.29$).

4.2.3 The role of affect

The above statistical models revealed that the stable factors influenced intention to disclose. Following the conceptual framework proposed by Dinev et al. [13], and the current results, we sought to investigate further the relationship between these factors and behavioral outcomes (**RQ2**).

First, we ran bootstrapped mediation analysis [27] to identify whether privacy concerns mediated the influence of trust on the intention to disclose.

Table 5: Results of mediation analysis: trust→privacy concerns→intention to disclose.

Predictor	Coeff.	*SE*	*t*	*p*
	Privacy concerns			
TRUST	-0.17	0.03	-5.03	<0.001
	$R^2 = 0.04$			
	$F(1,618) = 25.36, p < 0.001$			
	Intention to disclose			
TRUST	0.85	0.12	7.03	<0.001
PRIVACY CONCERNS	-0.70	0.13	-5.20	<0.001
	$R^2 = 0.09$			
	$F(1,617) = 47.42, p < 0.001$			
	Intention to disclose (total effect)			
TRUST	0.977	0.12	8.06	<0.001
	$R^2 = 0.09$			
	$F(1,618) = 65.05, p < 0.001$			

The results of simple mediation, demonstrated that trust indirectly influenced the intention to disclose through its effect on privacy concerns. The analysis showed that trust was a significant predictor of privacy concerns, $b = -0.17, t(618) = -5.03, p < 0.001$. Privacy concerns were significantly predicting intention to disclose, $b = -0.70, t(617) = -5.12, p < 0.001$. There was a significant effect of trust predicting disclosure, mediated by privacy concerns, $b = 0.97, t(618) = 8.06, p < 0.001$. Lastly, the direct effect of trust on intention to disclose was also significant, $b = 0.85, t(618) = 7.03, p < 0.001$. The analysis of direct and indirect effects showed that the indirect effect $= 0.12, SE = 0.03$ was significant with bias-corrected bootstrap CI 95%$[0.06, 0.20]$. Thus, the presence of mediation was confirmed.

After establishing the mediation effect, we wanted to examine the role of valence. We used the index of moderated mediation to evaluate whether moderated mediation was present. When bootstrapped confidence intervals of the index of moderated mediation do not include zero, it is assumed that the relationship between the indirect effect and the moderator is not zero, indicating presence of moderated mediation [28]. Additionally, an index not including zero indicates that "any two indirect effects conditioned on different values of [moderator] are statistically different from each other" [28, p. 14]. Hence, there is no need to probe the moderator via further statistical tests.

We examined whether a different level of valence influenced the indirect relationship between trust and intention to disclose. We looked for an interaction effect, either at the first or the second stage of the path model. Figure 4 shows paths in the model, and Table 6 presents the model's results.

There was an interaction effect (Figure 5) at the first stage

Table 6: Moderated mediation: trust→privacy concerns→intention to disclose; moderator: valence.

		Consequent						
		M (Privacy Concerns)				Y (Intention to disclose)		
Antecedent		Coeff.	*SE*	*p*		Coeff.	*SE*	*p*
X (Trust)	a_1	-0.63	0.13	<0.001	c'_1	0.80	0.44	0.07
M (Privacy concerns)					b	-0.70	0.13	<0.001
W (Valence)	a_2	-0.52	0.15	<0.001	c'_2	1.05	0.51	0.04
X × W	a_3	0.16	0.04	<0.001	c'_3	-0.03	0.14	0.81
Constant	i_M	6.76	0.42	<0.001	i_Y	6.77	1.68	<0.001
		$R^2 = 0.06$				$R^2 = 0.15$		
		$F(3,616) = 13.17, p < 0.001$				$F(4,615) = 28.41, p < 0.001$		

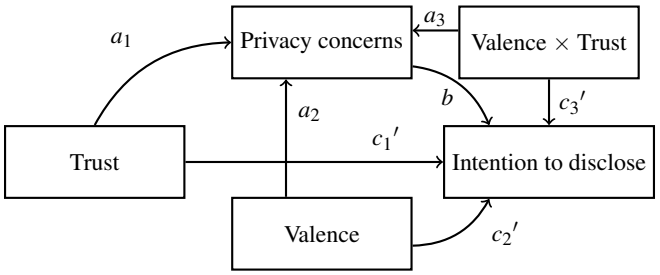


Figure 4: Paths in the model of moderated mediation.

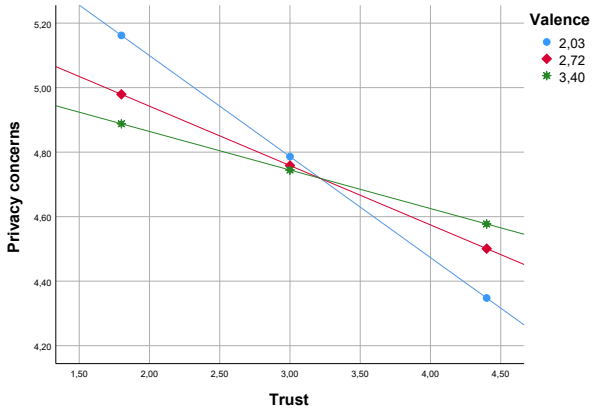


Figure 5: Interaction effect: moderated mediation.

of the model (a_2 in Figure 4). The relationship between trust and privacy concerns was moderated by valence. The analysis shows that among the participants with low trust and low valence, scores for privacy concerns were higher than among the participants with low trust and high valence. However, this effect is reversed among the participants with higher trust levels. Among those, the participants with low valence scored lower in privacy concerns than those with a high level of valence.

The bootstrapped index of moderated mediation was significant, confirming that there is an indirect effect of trust on the intention to disclose when controlling for privacy concerns, moderated by valence. The analysis of the conditional effect of focal predictor at values of the moderator showed that at scores of valence smaller than 2.71, trust and privacy concerns were significantly related, $b = -0.19, t(3,616) = -5.26, p < 0.001$, CI 95%$[-0.2, -0.1]$. With the decrease of valence, the relationship between trust and concerns becomes more negative, with the lowest score on valence 2.03, $b = -0.30, t(3,616) = -0.6, p < 0.001$, CI 95%$[-0.4, -0.2]$.

5 Discussion

In this section, we discuss our findings according to our two stated contributions. We present both practical implications and research insights for designers and researchers, respectively. Limitations and directions for future work conclude our discussion.

5.1 Practical implications

Affect. Our first research question asked *"does visual design of privacy notices influence affective states?"* (**RQ1**). Our findings show that designs providing control combined with curiosity may lead to an increase in valence. People might feel more positive (e.g., happy, pleased, satisfied) when provided with control, but only when holding high levels of curiosity. Our findings suggest that control increased general satisfaction with the design. Such results strengthen our prior exploratory findings [37], which showed that participants expressed a desire for control and choice when consenting to privacy notices.

As aforementioned, increased valence may influence satisfaction, which is one of the key elements contributing to usability (next to efficiency and effectiveness of use) [30]. Considering usability, Habib et al. [25, 26] found that current choices and controls implemented in privacy notices frequently lacked usability as they are inconsistent in their design, and challenging to understand from users perspective. Such designs often require users to go through a lengthy process (e.g., a few clicks, links redirecting users to different pages) before reaching the UI containing privacy controls. This has been further explored by Nouwens et al. [52],

who found that participants would consent to a privacy policy without seeking additional controls, if the controls were not placed on the first page of the consent screen. Further, Schaub et al. [57] discussed layered designs of privacy notices (i.e., interactive privacy notice instead of a single block of text), considering them as a way to overcome the status quo of notice and choice. Taking into account past research and our findings, it appears that control given during the sign-up process can carry the potential to increase the usability of privacy-related interactions.

Comprehension. Our fourth research question asked *"does visual design of privacy notices affects comprehension?"* (**RQ4**). The results did not confirm such a relationship, which suggests that differently framed notices have had no influence on privacy comprehension. However, in addressing our third research question, *"does curiosity influence privacy comprehension?"* (**RQ3**), we determined that curiosity positively affected comprehension. Whilst this may sound self-evident, to the best of our knowledge curiosity has not been considered in the design of privacy policies. Oddly, as it is a robust motivational trait that drives human cognitive development [35].

As postulated in the information-gap theory, people seek knowledge to fill the gap in their current understanding [44]. While some people may possess a basic understanding of online privacy, this theory suggests a desire may exist in some users to seek further information. The visual appearance of privacy notices could promote such desires by purposefully designing interfaces that stimulate users' curiosity. For instance, designers could implement methods acquired from game design to encourage user participation and engagement. The intuitive and immediate interactions could be triggered when prompting users to modify settings, instead of idly reading the text of privacy notice. Moreover, privacy designers could follow some of the guidelines for UIs enhancing curiosity, such as those proposed by Malone [46]. Malone postulated evoking curiosity through the *optimal level of informational complexity* — through environments that are "neither too complicated nor too simple with respect to the user's existing knowledge" [46, p. 67]. The optimal level might be achieved with the application of *novelty* and *surprise*, *randomness* and *humor* that could elicit positive experiences, increasing satisfaction. However, such methods would have to be thoroughly tested, as they may bring an inverse effect, and result in dissatisfaction or incomprehensibility.

Perhaps privacy policies arranged into "gradually discoverable" elements of information may be able to activate curiosity and enhance engagement. Such an approach was previously investigated in the context of crowd workers' performance, where it was found to be successful [41]. The progressive revelation of the information improved noticeability, and kept people curious to unravel more information to fill gaps in their knowledge. However, while in the context of privacy such an approach might influence engagement and prominence of information, applied to real-life situations, it could result in habituation effects or be disruptive during the application sign-up process. Taking into consideration previous findings, and recognizing the small effect sizes of our results, we believe that "gradually discoverable" privacy notices need to be thoroughly researched to establish their efficacy and efficiency during an application sign-up.

Control. In our fifth research question, we asked, *"does visual design of privacy notices affect intention to disclose information?"* (**RQ5**). Our results indicate that among people provided with a choice to adjust their settings, those holding higher levels of curiosity disclosed less information. This finding suggest the need to design privacy notices that provide users with options to set individual preferences, which allow them to adjust the amount and type of information disclosed or shared.

Consistent with our findings, past research found that control influences the value that people attach to personal information, and impacts disclosure [1]. Hence, our results supported by prior findings, indicate that future designs of privacy notices affording people with greater levels of usable control may be able to improve information disclosures. Still, we recognize that such solutions might be untimely and difficult to implement, as they require severe modifications in the system design or business models. Further, such solutions might restrict access to the complete functionality of an application, leading to the potential loss of a customer or disparity in service levels between users.

On the other hand, providing users with control may have a positive effect on companies. Improved usability may encourage users to utilize an application, perceiving it as more attractive due to enhanced transparency around information disclosures. Additionally, providing control may increase legal compliance, e.g., the fulfillment of the GDPR's requirements.

Importance of time. We found that the time spent on the notice page had a significant influence on comprehension, having an effect size larger than other variables. This finding indicates that people contributing their time to gather information are more aware of the privacy notice. We interpret this result as a call for designs of notices that engage users to spend more time on the notice's page. Perhaps the methods discussed above, such as the incremental revelation of information or layered designs of privacy notices might have such an effect.

Concerns, trust, and valence. Our findings suggest that valence moderates the relationship between trust and privacy concerns. Such a result might be applied to privacy UI designs with malicious intentions in mind. For instance, as the *dark pattern* that might influence valence, and lead to manipulating users' attitude towards lesser concerns. Effectively, this might "trick" users into disclosing more. On the other hand, we believe that such a result could be implemented in UI to increase privacy concerns. For example, by highlighting

risks and harms to privacy through interface design, which may lower valence, and result in a more negative relationship between trust and concerns, indirectly reducing information disclosure.

5.2 Research insights

Concerns, trust, and valence. Our second research question asked, *"what is the role of the affective state in the relationship between attitudinal factors and intention to disclose?"* (**RQ2**). According to our findings, privacy concerns mediate the relationship between trust and disclosure. More importantly, our results show that valence is moderating such a relationship. It seems that lower trust leads to greater privacy concerns; however, the affective state may alter the direction of this relationship. Our findings suggest that an increase in valence diminishes the effects of trust on concerns, which might be interpreted as the possibility to alter privacy concerns via elicitation of different emotional states.

To the best of our knowledge, little attention has been given to the role of affect in privacy-related decision making. Our results call for future privacy studies, which address this gap, focusing on the role of affective states in different contexts. Studying affect may contribute to a greater understanding of cognitive processes active during day-to-day privacy-related actions, and consequently may have practical implications.

Additionally, our findings confront the *privacy paradox*, and demonstrate a significant relationship between trust, privacy concerns, and intention to disclose. Perhaps, the phenomenon is not present in the context of a well-being application sign-up process, in which people's behaviors are aligned with their beliefs.

Cognitive processing. The moderating effect of valence suggests that intention to disclose might be an effect of the Type 1 processing, which utilizes mental shortcuts and simple solutions as the means for information processing. Further, our results indicate that the affect-as-information and feeling-as-information theories may be applied in the privacy research [9,58]. Our findings demonstrate that concerns increase with a decrease in valence. Perhaps because of the negative valence, people perceive the situation as unsafe, using affect as an indicator / marker of safety.

5.3 Limitations and future work

We used a crowdsourcing platform and gathered participants mostly from English speaking countries, which make our findings less generalizable to a wider population. We examined only one context of privacy interaction, a sign-up process for the well-being application. Recognizing the contextuality of privacy, the sign-up process for another type of application could lead to different conclusions.

Additionally, we might have primed participants by exposing them to privacy controls, and as a result, affect their intention to disclose. However, if this was the case, we might perceive such priming as a relevant outcome, demonstrating that solely the presence of controls has the potential to make people more privacy-conscious.

Our statistical analysis resulted in small effect sizes, which is anticipated in exploratory studies. However, our results are significant, and the effects found provide motivation for future replication studies to determine the magnitude of our findings. Additionally, it is essential to mention that the interpretation of the effect sizes for models with covariates is challenging, as they are difficult to compare to standard benchmarks defined by Cohen, which were based on unrestricted populations [40]. Cohen was cautiously discussing these benchmarks, acknowledging that they might not apply to all research areas [56]. More recent findings revealed that he might have been correct, as there are significant differences in effect sizes between different sub-disciplines of psychology, with the actual effect sizes being lower than the commonly used criteria [56].

Considering the limitations mentioned above, we propose that future work should include comparable experiments, but place them in diverse settings, targeting different population and privacy contexts. Further, we call for research incorporating methods different than self-reported measures, particularly in the case of affect. Both observational data, as well as physiological measurements (e.g., EEG, fMRI), could be applied in future inquiries to assess levels of affect more accurately.

6 Conclusion

We conducted an empirical analysis of privacy interactions during the application sign-up process. To gather the necessary data, we ran an online experiment with 620 English speaking participants. Our results show that people driven by curiosity utilize control over their information. We examined how this affects their intention to disclose, privacy comprehension, and affective state (positive–negative valence). Further, we investigated the role of valence in the relationship between trust and privacy concerns. Our research indicates that the visual design of privacy notices may have a beneficial influence on personal information disclosures. However, other factors should be taken into consideration to ensure improvement in individuals' privacy practices. We discuss our findings in the context of their applicability to the design of privacy notices as well as future research directions, calling for a change in both practical and theoretical approach to privacy research.

Acknowledgment

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 675730 (Privacy&Us Innovative Training Programme).

References

[1] A. Acquisti, M. Sleeper, Y. Wang, S. Wilson, I. Adjerid, R. Balebako, L. F. Brandimarte, L.and Cranor, S. Komanduri, P. G. Leon, N. Sadeh, and F. Schaub. Nudges for Privacy and Security. *ACM Computing Surveys*, 50(3):1–41, 2017.

[2] I. Adjerid, A. Acquisti, L. Brandimarte, and G. Loewenstein. Sleights of Privacy : Framing , Disclosures , and the Limits of Transparency. *Symposium on Usable Privacy and Security (SOUPS)*, page 17, 2013.

[3] J. Angulo, S. Fischer-Hübner, E. Wästlund, and T. Pulls. Towards usable privacy policy display and management. *Information Management & Computer Security*, 20(1):4–17, 2012.

[4] G. Bente, T. Dratsch, S. Rehbach, M. Reyl, and B. Lushaj. Do you trust my avatar? Effects of photo-realistic seller avatars and reputation scores on trust in online transactions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8527 LNCS:461–470, 2014.

[5] L. Brandimarte, A. Acquisti, and G. Loewenstein. Misplaced Confidences: Privacy and the Control Paradox. *Social Psychological and Personality Science*, 4(3):340–347, 2013.

[6] B. Brown. Studying the Internet experience. *HP Laboratories Technical Report HPL*, 2001.

[7] California State Legislature. SB-1121 California Consumer Privacy Act of 2018, 2018. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1121.

[8] E. K. Choe, J. Jung, B. Lee, and K. Fisher. Nudging People Away from Privacy-Invasive Mobile Apps through Visual Framing. In *IFIP Conference on Human-Computer Interaction*, pages 74–91. Springer, 2013.

[9] G. L. Clore, K. Gasper, and E. Garvin. Affect as information. *Handbook of affect and social cognition*, pages 121–144, 2001.

[10] G. Conti and E. Sobiesk. An Honest Man Has Nothing to Fear: User Perceptions on Web-based Information Disclosure. In *Proceedings of the 3rd Symposium on Usable Privacy and Security (SOUPS)*, page 112, 2007.

[11] K. P. L. Coopamootoo and T. Groß. Why Privacy Is All But Forgotten. *Proceedings on Privacy Enhancing Technologies*, 2017(4):97–118, 2017.

[12] P. Dattalo. *Analysis of multiple dependent variables*. Oxford University Press, 2013.

[13] T. Dinev, A. R. Mcconnell, and H. J. Smith. Informing Privacy Research Through Information Systems, Psychology, and Behavioral Economics: Thinking Outside the "APCO" Box. *Information Systems Research*, 26(4):639–655, 2015.

[14] S. Egelman. My profile is my password, verify me!: the privacy/convenience tradeoff of facebook connect. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2369–2378, 2013.

[15] S. Epstein, R. Pacini, V. Denes-Raj, and H. Heier. Individual Differences in Intuitive-Experiential and Analytics-Rational Thinking Styles. *Journal of Personality and Social Psychology*, 71(2):390–405, 1996.

[16] European Commission. EU general data protection regulation (GDPR) 2016/679, 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679.

[17] European Commission. Special Eurobarometer 487a: The General Data Protection Regulation. Technical Report March, EC, 2019.

[18] J. B. T. Evans and K. E. Stanovich. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, 8(3):223–241, 2013.

[19] A. Felt, E. Ha, S. Egelman, and A. Haney. Android permissions: User attention, comprehension, and behavior. In *Symposium on Usable Privacy and Security*, SOUPS '12, pages 1–16, 2012.

[20] M. L. Finucane, A. Alhakami, P. Slovic, and S. M. Johnson. The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, 13(1):1, 2000.

[21] N. Gerber, P. Gerber, and M. Volkamer. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers and Security*, 77:226–261, 2018.

[22] A. Ghorayshi and S. Ray. Buzzfeed - grindr is letting other companies see user hiv status and location data, 2018. https://www.buzzfeed.com/azeenghorayshi/grindr-hiv-status-privacy?utm{_}term=.xorXWLx2{#}.mdk3Q5ml.

[23] J. Gluck, F. Schaub, A. Friedman, H. Habib, N. Sadeh, L. F. Cranor, and Y. Agarwal. How Short is Too Short? Implications of Length and Framing on the Effectiveness of Privacy Notices. In *Proceedings of the Twelfth*

USENIX Conference on Usable Privacy and Security, SOUPS '16, page 321–340, USA, 2016. USENIX Association.

[24] J. Gottlieb, P. Y. Oudeyer, M. Lopes, and A. Baranes. Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11):585–593, 2013.

[25] H. Habib, S. Pearman, J. Wang, Y. Zou, A. Acquisti, L. F. Cranor, N. Sadeh, and F. Schaub. "It's a scavenger hunt": Usability of Websites' Opt-Out and Data Deletion Choices. In *Conference on Computer Human Interaction (CHI)*, 2020.

[26] H. Habib, Y. Zou, A. Jannu, N. Sridhar, C. Swoopes, A. Acquisti, L. F. Cranor, N. Sadeh, and F. Schaub. An Empirical Analysis of Data Deletion and Opt-Out Choices on 150 Websites. *Proceedings of the 15th Symposium on Usable Privacy and Security, SOUPS 2019*, pages 387–406, 2019.

[27] A. F. Hayes. Process: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling, 2012.

[28] A. F. Hayes. An Index and Test of Linear Moderated Mediation. *Multivariate Behavioral Research*, 50(1), 2015.

[29] A. Hern. Fitness tracking app Strava gives away location of secret US army bases, 2018. https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases.

[30] International Organization for Standardization. Ergonomics of human-system interaction–part 11: Usability: Definitions and concepts, 2018.

[31] S. Jenkins, R. Brown, and N. Rutterford. Comparing Thermographic, EEG, and Subjective Measures of Affective Experience During Simulated Product Interactions. *International Journal of Design*, 3(2):53–65, 2009.

[32] A. N. Joinson, C. Paine, T. Buchanan, and U. D. Reips. Measuring self-disclosure online: Blurring and non-response to sensitive items in web-based surveys. *Computers in Human Behavior*, 24(5):2158–2171, 2008.

[33] D. Kahneman. A Perspective on Judgment and Choice. *American Psychologist*, 3(4):7–18, 2003.

[34] T. B. Kashdan, M. F. Steger, B. P. Winterstein, W. E. Breen, P. J. Silvia, M. W. Gallagher, and D. Terhar. The curiosity and exploration inventory-II: Development, factor structure, and psychometrics. *Journal of Research in Personality*, 43(6):987–998, 2009.

[35] C. Kidd and B. Y. Hayden. The Psychology and Neuroscience of Curiosity. *Neuron*, 88(3):449–460, 2015.

[36] Y.-S. Kim, K. Reinecke, and J. Hullman. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1375–1386, 2017.

[37] A. Kitkowska, Y. Shulman, L. A. Martucci, and E. Wästlund. Facilitating privacy attitudes behaviors with affective visual design. In *To appear in: 35th International Conference on ICT Systems Security and Privacy Protection – IFIP SEC 2020*, 2020.

[38] A. Kitkowska, Y. Shulman, L. A. Martucci, and E. Wästlund. Psychological effects and their role in online privacy interactions: A review. *IEEE Access*, 8:21236–21260, 2020.

[39] B. P. Knijnenburg and A. Kobsa. Making Decisions about Privacy: Information Disclosure in Context-Aware Recommender Systems. *ACM Transactions on Interactive Intelligent Systems*, 3(23), 2013.

[40] D. Lakens. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4(NOV):1–12, 2013.

[41] E. Law, M. Yin, J. Goh, K. Chen, M. A. Terry, and K. Z. Gajos. Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4098–4110, New York, NY, USA, 2016. Association for Computing Machinery.

[42] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam. Emotion and decision making. *Annual Review of Psychology*, 66:799–823, 2015.

[43] H. Li, R. Sarathy, and H. Xu. The role of affect and cognition on online consumers' decision to disclose personal information to unfamiliar online vendors. *Decision Support Systems*, 51(3):434–445, 2011.

[44] G. Loewenstein. The Psychology of Curiosity: A Review and Reinterpretation. *Psychological Bulletin*, 116(1), 1994.

[45] N. K. Malhotra, S. S. Kim, and J. Agarwal. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, 15(4):336–355, 2004.

[46] T. W. Malone. Heuristics for designing enjoyable user interfaces: Lessons from computer games. In *Proceedings of the 1982 Conference on Human Factors in Computing Systems*, CHI '82, page 63–68, New York, NY, USA, 1982. Association for Computing Machinery.

[47] T. Mirsch, C. Lehrer, and R. Jung. Digital Nudging: Altering User Behavior in Digital Environments. *Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017)*, pages 634–648, 2017.

[48] S. Monteleone, R. Van Bavel, N. Rodríguez-Priego, and G. Esposito. Nudges to privacy behaviour: Exploring an alternative approach to privacy notices. *JRC Science and Policy Report. Luxembourg, Luxembourg: Publications Office of the European Union*, 2015.

[49] H. Nissenbaum. A contextual approach to privacy online. *Daedalus*, 140(4):32–48, 2011.

[50] E. Noll Webb, G. Balasubramanian, U. OBroin, and J. M. Webb. WHAM! POW! Comics as User Assistance. *Journal of Usability Studies*, 7(3):105–117, 2012.

[51] P. A. Norberg, D. R. Horne, and D. A Horne. The Privacy Paradox : Personal Information Disclosure Intentions versus Behaviors. *The Journal of Consumer Affairs*, 41(1):100–126, 2007.

[52] M. Nouwens, I. Liccardi, M. Veale, D. Karger, and L. Kagal. Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence. In *Conference on Human Factors in Computing Systems(CHI)*, 2020.

[53] C. Phelan, C. Lampe, and P. Resnick. It's Creepy, But It Doesn't Bother Me. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5240–5251, San Jose, 2016.

[54] P. Rajivan and J. Camp. Influence of privacy attitude and privacy cue framing on android app choices. In *Twelfth Symposium on Usable Privacy and Security*, 2016.

[55] J. A. Russell and L. F. Barrett. Core Affect, Prototypical Emotional Episodes, and Other Things Called Emotion: Dissecting the Elephant. *Journal of Personality and Social Psychology*, 76(5), 1999.

[56] T. Schäfer and M. A. Schwarz. The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, 10(APR):1–13, 2019.

[57] F. Schaub, R. Balebako, A. L. Durity, and L. F. Cranor. A Design Space for Effective Privacy Notices. In *Symposium on Usable Privacy and Security (SOUPS) 2015*, pages 1–17, Ottawa, Canada, 2015.

[58] N. Schwarz. Feelings-as-Information Theory. *Handbook of Theories of Social Psychology: Volume 1*, pages 289–308, January 2012.

[59] F. Slovic. The Affect Heuristic. In *Heuristics and Biases; The Psychology of Intuitive Judgement*, pages 397–420. Cambridge University Press, 2002.

[60] J. H. Smith, T. Dinev, and H. Xu. Information Privacy Research: An Interdisciplinary Review. *MIS Quarterly*, 35(December):989–1015, 2011.

[61] D. Solove. A Taxonomy of Privacy. *University of Pennsylvania Law Review*, 154(3):477–560, 2006.

[62] K. E. Stanovich and R. F. West. Individual differences in reasoning: implications for the rationality debate? *The Behavioral and brain sciences*, 23(5):645–665; discussion 665–726, 2000.

[63] A. Sutcliffe and A. Namoune. Getting the Message across: Visual Attention, Aesthetic Design and What Users Remember. In *Proceedings of the 7th ACM Conference on Designing Interactive Systems*, DIS '08, page 11–20, New York, NY, USA, 2008. Association for Computing Machinery.

[64] M. Tabassum, A. Alqhatani, M. Aldossari, and H. Richter Lipford. Increasing User Attention with a Comic-Based Policy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, New York, NY, USA, 2018. Association for Computing Machinery.

[65] M. Taddicken. The 'Privacy Paradox' in the Social Web: The Impact of Privacy Concerns, Individual Characteristics, and the Perceived Social Relevance on Different Forms of Self-Disclosure. *Journal of Computer-Mediated Communication*, 19(2):248–273, 2014.

[66] D. Tasse, A. Ankolekar, and J. Hailpern. Getting Users' Attention in Web Apps in Likable, Minimally Annoying Ways. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, pages 3324–3334, 2016.

[67] A. Toet, D. Kaneko, S. Ushiama, S.e Hoving, I. de Kruijf, A. M. Brouwer, V. Kallen, and J. B.F. van Erp. EmojiGrid: A 2D Pictorial Scale for the Assessment of Food Elicited Emotions. *Frontiers in Psychology*, 9(NOV):1–21, 2018.

[68] T. F. Waddell, J. R. Auriemma, and S. S. Sundar. Make it Simple, or Force Users to Read? Paraphrased Design Improves Comprehension of End User License Agreements. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5252–5256, 2016.

[69] N. Wang, B. Zhang, B. Liu, and H. Jin. Investigating effects of control and ads awareness on android users' privacy behaviors and perceptions. In *Proceedings of the 17th international conference on human-computer interaction with mobile devices and services*, pages 373–382, 2015.

[70] M. Zamanian and P. Heydari. Readability of Texts: State of the Art. *Theory & Practice in Language Studies*, 2(1), 2012.

[71] B. Zhang, M. Wu, H. Kang, E. Go, and S. S. Sundar. Effects of security warnings and instant gratification cues on attitudes toward mobile websites. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, pages 111–114, 2014.

Appendix

A Curiosity

Participant instructions:

Rate the statements below for how accurately they reflect the way you generally feel and behave. Do not rate what you think you should do, or wish you do, or things you no longer do. Please be as honest as possible.

1. I actively seek as much information as I can in new situations.
2. I am the type of person who really enjoys the uncertainty of everyday life.
3. I am at my best when doing something that is complex or challenging.
4. Everywhere I go, I am out looking for new things or experiences.
5. I view challenging situations as an opportunity to grow and learn.
6. I like to do things that are a little frightening.
7. I am always looking for experiences that challenge how I think about myself and the world.
8. I prefer jobs that are excitingly unpredictable.
9. I frequently seek out opportunities to challenge myself and grow as a person.
10. I am the kind of person who embraces unfamiliar people, events, and places.

Scoring:

Items 1, 3, 5, 7, and 9 reflect curiosity *stretching*. Items 2, 4, 6, 8, and 10 reflect curiosity *embracing*. Items were anchored on the scale: 1 – very slightly or not at all, 2 – a little, 3 – moderately, 4 – quite a bit, 5 – extremely.

B Affective Self Report

Participant instructions:

1st measurement instructions: Thinking about yourself, to what extent do you currently feel:

2nd measurement instructions: Earlier in the study, we asked you how did you feel. Thinking back about the sign-up process, would you say that now you feel different or the same in comparison to when we previously asked you?

1. Annoyed – Pleased
2. Tired – Energetic
3. Unpleasant – Pleasant
4. Patient – Anxious
5. Irritated – Content
6. Unhappy – Happy
7. Calm – Restless
8. Disappointed – Satisfied
9. Relaxed – Tense
10. Indifferent – Curious

Scoring:

Items scored on 7 points Likert scale. The second measurement was labelled with the word *more*, e.g., More Annoyed – More Pleased.

C Intention to disclose information

Participant instructions:

Thinking back about the sign-up process and considering the previously presented scenario, if you were to sign up for this application, would you be willing to share any of the following information with this application provider?

1. Your age
2. Your weight
3. Your height
4. Gender
5. Ethnicity
6. Your sexual orientation
7. Your marital status
8. Number of children
9. Chronic conditions
10. Overall number of sexual partners, since you became sexually active
11. Religious beliefs
12. Employment status
13. Political beliefs
14. Monthly income

Answers:

"I would disclose" or "I would prefer not to say."

D Privacy comprehension

Participant instructions:

Thinking back about the sign-up process, could you please tell us which of the following statements you believe are true considering the privacy policy that you have been asked to read.

1. Personal information is any information about you that is collected by an online service provider.

2. Information about you collected through any forms, including sign-up form is used to personalize services.
3. The service provider will collect your health information.
4. You are contractually obliged to provide your contact information.
5. You have full control over your personal information if you sign up for forums and create a public profile on this application, and you control how this information is being shared with others.
6. There are third parties that collect data about you and this service's policy applies to the processing of your information by such third parties.
7. If you are logged in to your social media and use the application at the same time, information about your activities will be tracked and recorded by social media providers.
8. This application transfers personal data to companies located abroad. These services can freely process your personal information for their purposes.
9. The service provider is legally obliged to share your personal information, and it does not need to inform you about it.
10. The service provider can process your personal data without your consent, for any purpose that was not explained in its privacy policy.

Answers:

"True", "False", "I don't remember / I don't know."

E Trust

Participant instructions:

Please read the statements below and indicate to what extent you disagree or agree with each of the statements.

1. Online companies would be trustworthy in handling the information.
2. Online companies would tell the truth and fulfil promises related to the information provided by me.
3. I trust that online companies would keep my best interests in mind when dealing with the information.
4. Online companies are in general predictable and consistent regarding the usage of the information.
5. Online companies are always honest with customers when it comes to using the information that I would provide.

Scoring:

Items scoring on 7 points Likert scale, anchored "Strongly disagree" – "Strongly agree".

F Privacy concerns

Participant instructions:

Please state, to what extent do you agree with the following sentences.

1. All things considered, the Internet may cause serious privacy problems.
2. Compared to others, I am more sensitive about the way online companies handle my personal information.
3. To me, it is the most important thing to keep my privacy intact from online companies.
4. I believe other people are too much concerned with online privacy issues.
5. Compared with other subjects on my mind, personal privacy is very important.
6. I am concerned about threats to my personal privacy today.

Scoring:

Items scoring on 7 points Likert scale, anchored "Strongly disagree" – "Strongly agree."

G Text of privacy policies

OUR POLICY

On this page, you can find an overview of our privacy policy. If you think the information here is insufficient, you can check the full text of Privacy Policy.

WHAT DATA WE COLLECT AND USE WHEN YOU VISIT OUR SITE

We collect Non-personal and Personal information when you visit our website. Personal Data is information that identifies you or could be used to identify you, e.g., name, address, email. Some of our services require the processing of your health-related data. We collect information that you provide directly to us when you choose to use our Services. We also collect data that you submit through responses to any forms such as sign up or profile creation forms, questionnaires, etc. We use this data to personalize our services and to optimize your experience.

WHEN WE COLLECT YOUR DATA AND WHY

We collect information, e.g., Personal Data, when you browse our website or use our service. Among the information collected are your IP address, browser type, operating system, error logs, and the like. Such aggregated information does not identify you and is used by us to analyze trends, to administer and monitor our site, its use, and to gather general information about the use of our website.

HOW WE DISCLOSE YOUR DATA

There are a few instances when we are obliged to disclose your information. E.g., to pursue our legitimate interest in applying or enforcing the terms and conditions, or to respond to any claims. We may disclose your data to protect our rights or the rights of a third party; to protect the safety of any person or to prevent any illegal activities. If legally required to do so, we will collect your prior consent before sharing your Personal Data with other companies.

HOW WE USE YOUR DATA

We use your data to send you service announcements and updates regarding our Website. You are contractually required to provide us with such Personal Data as, without it, we will not be able to send you service-related communication.

PROCESSING FOR OTHER PURPOSES
If your Personal Data are processed for purposes not mentioned in this policy, we will provide you with information on that other purposes and any additional relevant information as referred to in this Privacy Policy.

SHARING YOUR DATA
We may share some of your Personal Data with our company located in other countries, providing us with hosting services. We use third-party service providers to offer or facilitate services on our behalf and share your data with such providers to the extent necessary to perform their services on our behalf. They are prohibited from using your Personal Data or any other purposes than those described in this Privacy Policy.

SOCIAL FEATURES
We feature public forums such as message boards, bulletin boards or activities where you and other users can communicate with one another. The Public Profile feature permits you to share information about yourself (including, if you elect, Personal Data) with others. If you use Social Features, we cannot control how other users might use your data. We also cannot prevent you from receiving unwanted messages from others. You are not legally required to provide us with your Personal Data, but without it, we cannot offer you to use our Social Features.

SOCIAL PLUGINS
Our Website contains links to or features from other sites. This Policy does not cover the privacy practices of third-party websites or features. We use social networks plugins of Facebook, Twitter and YouTube. If you visit our Website while signed in to your social media account, results in the transfer of information about you to the social network. Such information can be linked with your social network account. This data transfer is triggered already when you visit our Website, irrespective whether you interact with the plugin. To prevent this, you must log out of your social network account before visiting our Website.

CONTACT
If you have any questions about our Privacy Policy or feel that we are not abiding by the terms of our posted Privacy Policy or the applicable data protection laws, please contact our data protection officer at legal@abc.com.

G.1 Amended text of policy for groups given control

OPT-OUT FROM INFORMATION PROCESSING
We do not want to collect all of the information about you. However, the more information we have, the more accurate and personalized services we can offer. To ensure your control over the information, we offer you options to opt-out from particular data collection and processing. If you wish to limit the collection of your information, change the switches to Disabled mode.

SHARING YOUR DATA
We may share some of your Personal Data with our company located in other countries, providing us with hosting services. We use third-party service providers to offer or facilitate services on our behalf and share your data with such providers to the extent necessary to perform their services on our behalf. They are prohibited from using your Personal Data or any other purposes than those described in this Privacy Policy. If you don't want us to transfer your information to servers located abroad, you can disable this as per our Policy.

SOCIAL FEATURES
We feature public forums such as message boards, bulletin boards or activities where you and other users can communicate with one another. The Public Profile feature permits you to share information about yourself (including, if you elect, Personal Data) with others. If you use Social Features, we cannot control how other users might use your data. We also cannot prevent you from receiving unwanted messages from others. You are not legally required to provide us with your Personal Data, but without it, we cannot offer you to use our Social Features. If you do not want to have Social Features, you can disable this functionality, and we will not provide you with such services.

SOCIAL PLUGINS
Our Website contains links to or features from other sites. This Policy does not cover the privacy practices of third-party websites or features. We use social networks plugins of Facebook, Twitter and YouTube. If you visit our Website while signed in to your social media account, results in the transfer of information about you to the social network. Such information can be linked with your social network account. This data transfer is triggered already when you visit our Website, irrespective whether you interact with the plugin. To prevent this, you must log out of your social network account before visiting our Website. Alternatively, you can disable the social media plugins as offered in our Policy.

G.2 Images applied in the policy display

Each section of the text in the privacy policy contained framing image, as presented in figs. 6 to 13 (A.- negative, B.- positive).



A. B.

Figure 6: Images displayed next to the policy section "WHAT DATA WE COLLECT AND USE WHEN YOU VISIT OUR SITE."



A. B.

Figure 7: Images displayed next to the policy section "WHEN WE COLLECT YOUR DATA AND WHY."



A. B.

Figure 8: Images displayed next to the policy section "HOW WE DISCLOSE YOUR DATA."



A. B.

Figure 9: Images displayed next to the policy section "HOW WE USE YOUR DATA."



A. B.

Figure 10: Images displayed next to the policy section "PROCESSING FOR OTHER PURPOSES."



A. B.

Figure 11: Images displayed next to the policy section "SHARING YOUR DATA."



A. B.

Figure 12: Images displayed next to the policy section "SOCIAL FEATURES."



A. B.

Figure 13: Images displayed next to the policy section "SOCIAL PLUGINS."

"I Have a Narrow Thought Process": Constraints on Explanations Connecting Inferences and Self-Perceptions

Emilee Rader
Michigan State University
emilee@msu.edu

Samantha Hautea
Michigan State University
hauteasa@msu.edu

Anjali Munasinghe
Michigan State University
amunasinghe1620@gmail.com

Abstract

Most people are unfamiliar with the kinds of inferences that platforms like Facebook and Google can automatically associate with them, despite the existence of interfaces designed to provide transparency to end users. We conducted a study to investigate people's reactions upon being exposed to these inferences, to learn if and how they perceived the inferences to be connected to themselves. Through qualitative analysis, we found that the evidence participants used to relate the inferences with their self-perceptions was bounded by what they remembered about their own past behaviors in connection with the platform. Inferences that participants felt were implausible given their own behavior were rationalized as being related to family members, outdated, or could fit anyone with similar demographic characteristics. Participants also identified some inferences they believed had no connection with themselves whatsoever. We discuss implications for how participants' reasoning might lead to expectations about what kinds of inferences are possible, and what this means for people's ability to make informed privacy decisions regarding consent and disclosure.

1 Introduction

The current model for governing digital data collection and use, *notice and choice*, entails providing access to complex terms of use or privacy policies. These documents describe how platforms—systems consisting of networked hardware and software that connect people with information and with each other—will collect, use, and share the data they collect about people. The notice and choice model assumes that if proper transparency is provided, then people will choose not to use platforms that have data collection and use practices they don't like [28]. This model reinforces the idea that digital privacy—control over how data about oneself is collected and used—can be effectively managed by individuals who make informed choices that are aligned with their preferences.

But, in reality, people struggle to manage their privacy. Widespread data collection and use of machine learning technologies combined with the rise of data brokers makes it possible for platforms to generate inferences about users, which consist of new information that is automatically derived from data the platforms collect or obtain [2]. For example, one person's browsing history aggregated with data collected from millions of other people can be used to derive inferences about anyone who uses the platform, revealing sensitive personal characteristics people might prefer not to disclose [4, 7, 14, 17]. Most people don't have a very good understanding of what kinds of inferences can be derived in this way [16, 23, 24, 32] or how those inferences can be used to direct their attention and influence their choices and opportunities.

New data privacy laws like the E.U. General Data Protection Regulation (GDPR)[1] and the U.S. California Consumer Privacy Act (CCPA)[2] have mandated better transparency about data collection and use. Partially in response to this, platforms have begun showing more information to users about what they are doing with the data they collect about them. Both Facebook and Google provide a web page where people with an account on the platform may view their "Ad Settings factors" (Google)[3] or "Ad Preferences categories" (Facebook)[4]. Despite language that connotes some level of user ownership and control—the inferences are described as "your ads" by Google, and "your information" by Facebook—these pages display information that has been automatically generated by the platform, not specified by the user. Interfaces

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

[1] https://gdpr-info.eu
[2] https://oag.ca.gov/privacy/ccpa
[3] https://adssettings.google.com/
[4] https://www.facebook.com/ads/preferences/

like these provide some visibility into the kinds of information it is possible for platforms to infer about users, and as such are potentially an important way for people to develop an understanding of their inference-generating capabilities.

We conducted a survey and interview study to investigate participants' reactions to the inferences Facebook and Google have made about them, and what their explanations about how and why the inferences were assigned reveal about their understanding of the platforms' inferencing capabilities and motivations. The way participants explain the connection (or lack of connection) they perceive between themselves and the inferences can reveal their guesses, knowledge, and insights about where the inferences come from.

We found that participants' understanding of the inferences was based primarily on their own memories about their past online behavior, their self-perceptions of their interests, and their perceptions of the interests of family members. Inferences that could not be explained as related to these things were assumed to be mistakes. These findings show that the framing people use for understanding inferences is limited to things they are directly knowledgeable about. This framing does not support envisioning inferences from the platforms' perspective. Therefore, inferences derived from data people cannot anticipate or for purposes they are unfamiliar with would be extremely difficult for them to truly consent to. This paper contributes novel findings to the privacy literature about the constraints on people's knowledge and understanding of what platforms can infer about them, which have implications for their ability to make informed informed privacy decisions regarding consent and disclosure.

2 Related Work

2.1 User Awareness of Online Tracking and Inferences

Tracking of user browsing behavior is a phenomenon that has become increasingly complex. At least one study has empirically shown that the scope of information tracked and the variety of trackers on websites have greatly increased over the past two decades, though studies on such tracking behavior came much later [21]. Media controversies surrounding large-scale tracking and the sharing of data have made some of the tensions between users and companies more public, but may not necessarily equip users to conceptualize more broadly about these systems [12].

One type of user tracking that has been studied extensively is online behavioral advertising (OBA), which presents users with advertisements targeted to them based on their tracked behavior (see [3] and [30] for a more comprehensive overview). User studies of reactions to targeted advertising show that users perceive useful and beneficial properties of OBA, even as there are aspects of it that they find uncomfortable. The delivery of more personalized, relevant content is considered useful, while discomfort can stem from perceptions of excessive intrusiveness, sometimes described as 'scary' or 'creepy' [5, 29]. Users who 'feel watched' in this way may be intuitively expecting systems to conform with social norms of not conducting unsolicited observation of others [22]. What these user perceptions have in common is an understanding of algorithmic inferences as driven by social entities and human logic, which may not be true for complex systems that make inferences across patterns beyond the intuitive capacity of the human mind to connect. Indeed, though websites may provide various disclosures about information collection (whether legally required or industry self-regulated), these do not appear to be particularly effective in providing users with meaningful notice that it is taking place [18, 34].

2.2 Awareness and Inference Literacy

The increasing complexity of online tracking, and the implications this tracking holds for privacy decision-making, has led some scholars to call for the need to recognize 'inference literacy,' or the "beliefs and misconceptions users hold about companies inferencing methods and capabilities" [31]. Users perceive that their online activities are being tracked or followed, though industry-standard icons, taglines, and explicit disclosures are often misinterpreted [19, 29]. Uncertainty about the usage of behavioral data is a subject of concern for users [25], and individuals articulate clear preferences for the type of information and categories they are comfortable with having associated with them based on what they believe others can do with that data [11, 20].

Feedback from information collection systems can greatly influence what users are able to conceptualize them doing [8, 10, 24]. Some companies have provided their users with access to their online behavioral profile. Experiments have been used to gain a more technical understanding of how some of the infrastructures underlying these data collection and inference activities respond to user behavior [6], but studies have also explored user reactions to these systems. A study that exposed users to their site-generated behavioral profile reported that participants found these difficult to comprehend, and focused solely on identifying user concerns with data collection [25]. Several tools have been created to increase user awareness about tracking and inferences, some of which employ a soft paternalism or 'nudging' approach to guide users to different decision-making processes by using underlying behavioral biases [1]. Other approaches have focused on creating novel ways of visualizing behavioral data, exemplified in one study where browsing information was collected from a plugin in order to show users what can be inferred from certain types of data [32]. A key theme that emerges from these studies is that user awareness is greater when they are able to recognize themselves in the data collected about them.

2.3 Research Goals

This study focuses on users' reactions to and explanations of real-world inferences assigned to them, in order to investigate how they make sense of the inferences and relate them

to their lives and self-perceptions. This differs from previous research on user understanding of underlying tracking mechanisms [29, 35], and platforms' reasons for tracking and tailoring ads and how that relates to privacy concern regarding tracking [9]. It also differs from Weinshel et al. [32] which generated its own inferences rather than investigating reactions to platforms' actual inferences. While Eslami et al. [11] found that that users "justify" algorithmic decisions by looking for connections between themselves and inferences, we take this idea further by identifying patterns in the explanations and evidence present the reactions our study elicited from participants to the inferences that had been assigned to them. These reactions, and the connections users make between the inferences and their self-perceptions, can help researchers and designers understand what makes inferences seem plausible and therefore what kinds of inferences people might expect when consenting to a platform.

3 Method

3.1 Approach

Our study focused on eliciting participants' reactions to the actual inferences made about them by Facebook and Google. We focused on these platforms because earlier research about awareness of inferences showed that participants' perceptions of the likelihood of different types of inferences varied by platform. For example, more participants who were shown a hypothetical scenario involving clicking on links in the Facebook News Feed believed inferences about their social relationships were likely than participants who viewed a scenario involving clicking on links in Google search results [23].

Many studies have used hypothetical scenarios to investigate inference-related awareness, concern, and privacy preferences; however, people can react differently to real evidence of platforms' data practices contextualized in their own lives and experiences, than to general descriptions or hypotheticals [22]. We began by conducting a survey in which we asked respondents to answer questions about the specific inferences made about them by either Facebook or Google. We then invited a subset of the survey respondents to participate in a follow-up interview where we showed them a report that presented the inferences they had answered questions about and a summary of their responses to the questions. We used the reports as a form of breaching experiment [13] that provided visibility into the platforms' inferencing capabilities, which most of our participants were previously not aware of. The interview focused on participants' overall reactions to the inferences, and their explanations about why they thought those inferences had been assigned to them and how they had been generated.

3.2 Survey

The survey took place in late March and early April 2019. Respondents were recruited using a subject pool consisting of members of the community surrounding Michigan State University. The first page of the survey informed potential respondents before they consented that they would be asked to log in to either Facebook or Google and navigate to a web page which they would be asked to download and then submit to the survey so that the survey could generate questions for them based on the content on that page. Eligible participants were at least 18 years old (*Min*=22, *Mean*=38, *Max*=71), and regular users of either Google or Facebook. Students at the university were not eligible to participate. People who consented to the survey and passed the screening questions were randomly assigned to answer questions about either their Facebook or Google inferences.

	Facebook	*Google*
Man	12	16
Woman	31	35
Other	1	0
Total	44	51
No college	1	4
Some college	7	3
Four-year college degree	15	19
Some postgraduate	1	4
Postgraduate degree	20	21
Income less than $34,999	15	13
Income $35,000 to $74,999	13	19
Income $75,000 to $149,999	11	17
Income $150,000 or more	5	2
Aware of inferences page: Yes	11	5
Aware of inferences page: Unsure	4	5
Aware of inferences page: No.	29	41
Num. inferences: Mean (SD)	12 (9.86)	67.6 (35.1)
Num. of inferences: Median	4	66
Num. of inferences: Max	33	131

Table 1: Characteristics of the survey respondents.

The survey asked questions about respondents' use of Facebook or Google and their demographics, and then it provided instructions for navigating to their inferences web page, saving the web page as a file, and uploading it to our server via the survey interface. The files submitted by respondents were parsed to identify the inference categories and then deleted. The survey then generated three questions about each inference, which respondents answered on a 7-point Likert scale (Strongly disagree–Strongly agree). "[inference]" in each question was replaced by the text parsed out of the file each respondent submitted via the survey:

- *sensible*: It makes sense that [inference] is associated with me.
- *relevant*: [inference] is relevant to who I am as a person.
- *accurate*: [inference] is an accurate description of my everyday activities.

	ID	Gender	Age	Aware?	Number of Inferences	Mean Accuracy
Facebook Participants	P01	Woman	62	No	20	5.05
	P02	Woman	28	No	26	5.44
	P10	Man	44	Unsure	25	4.48
	P11	Man	63	No	24	5.62
	P13	Woman	28	No	33	4.39
	P15	Woman	44	Yes	14	4.00
	P16	Woman	34	No	25	4.84
	P20	Man	38	Unsure	26	4.42
Google Participants	P03	Man	38	No	59	4.05
	P04	Woman	32	No	98	4.76
	P05	Man	29	Yes	63	2.32
	P07	Man	71	No	62	4.44
	P08	Woman	59	Unsure	83	4.49
	P09	Woman	63	No	40	4.78
	P12	Man	31	No	48	3.77
	P14	Man	30	No	94	4.68
	P17	Woman	48	No	65	3.54
	P18	Woman	30	No	104	3.87
	P19	Man	38	No	44	3.05
	P21	Woman	40	No	46	4.00
	P22	Woman	43	No	109	3.07

Table 2: Characteristics of the interview participants. 'Aware' indicates whether the participant had previously seen the platform's inferences web page. 'Mean Accuracy' is the average of each participant's responses to the *accuracy* survey question about each inference. Higher means greater perceived accuracy on a scale of 1–7.

For respondents with a very large number of inferences assigned to them, our code randomly selected 85 to ask about, because our pilot testing indicated that any more than this would result in an excessively long survey. The last question in the survey asked if respondents would be interested in participating in a follow-up interview, and 57% said yes and provided their contact information. On average, it took 32.5 minutes (*SD*=91) to complete the survey. Ninety-five respondents completed the survey (44 Facebook, 51 Google; 28 men, 66 women, 1 did not disclose gender). Eighty-seven percent of respondents were white, 7% were Asian, and the remaining respondents reported multiple ethnicities. Further demographics are presented in Table 1. Respondents who completed the survey received a $5 USD Amazon gift card. The survey questions are included in Appendix A.

3.3 Interviews

The interviews took place in April 2019, shortly after the survey data collection ended. From the 57 survey respondents who said they were interested in being interviewed, we randomly chose participants to invite, prioritizing gender and platform balance. After the first five interviews had been conducted, we realized that the nature of the Facebook inferences made them more difficult for participants to react to (see more about this in the Findings section). At that point we decided to conduct more of the remaining interviews with Google participants than Facebook participants. We conducted 22 interviews, but excluded P06 before analysis due to a high proportion of low quality, off-topic answers to questions. Of the 21 remaining interviews, 9 were with men and 12 with women, and 8 focused on Facebook inferences while 13 focused on Google inferences. Interview participants ranged in age from 28 to 71 (*Mean*=43). Most participants were currently or formerly employed as knowledge workers (e.g., staff or instructors at the university; teachers; working in local government; retired). See Table 2 for further information about the interview participants.

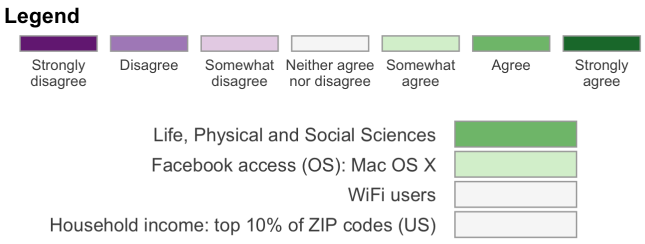


Figure 1: Example excerpt from a Facebook inferences report generated from data about one of the authors.

Each interview began by showing the participant a report that we generated which included the inferences they had answered questions about in the survey, and the average of the ratings of the three questions asked about each inference, rounded to the nearest whole number and then color-coded according to the position of the average on the 7-point disagree-agree Likert scale. The report was designed to be a simple artifact that enabled us to elicit participants' reactions to the inferences, independent of the jargon and branding on the platforms' own pages. The last page of the report also showed inferences that were unique to each participant among the survey respondents who had been assigned the same platform, and inferences that all survey respondents asked about that platform had been assigned. See Figure 1 for an excerpt of a report generated using the inferences of one of the authors. The interview protocol and example Facebook and Google inference reports are included in Appendix C.

The semi-structured interview protocol included questions designed to investigate participants' beliefs about why and how the inferences had been assigned to them, and about what the inferences are used for by Facebook and Google. Care was taken to elicit participants' thoughts and reactions in their own words and to ask non-leading follow-up questions, so that the interviewers did not impart a sense that technical correctness or incorrectness was an important framework for thinking about the inferences or prime their answers in other ways. We also avoided using the word "privacy" in the interviews unless the participant mentioned it first, to avoid influencing their reactions to the inferences. Interviews lasted 30-40 minutes,

and participants received a $25 USD Amazon gift card as a thank you for participating.

3.4 Analysis

The interviews were digitally recorded, transcribed, and identifying information was removed from the transcripts. We conducted iterative qualitative analysis of the transcripts that progressed over several rounds [26]. In the first round, two coders conducted inductive open coding that identified descriptive themes in the data. These themes included beliefs about participants' interests, characteristics, past activities, goals, etc. that were related to any connection they perceived between themselves and the inference they were talking about, and varying levels of understanding about how the inferences might have been selected. After the first round, the coders summarized all of the codes by participant, and the research team engaged in a several-day immersive interpretation session with the goal of identifying higher level themes. Afterward, one member of the research team engaged in a second round of coding focusing on characterizing types of reactions to the inferences (e.g., plausible, implausible, no connection; see the Findings section for details), and then a second member of the team focused on the reactions in order to group them by the types of evidence and examples participants provided in their explanations about how and why those inferences had been assigned to them. Our findings primarily focus on the higher-level themes and evidence identified in the second round of coding.

3.5 Limitations and Ethical Considerations

We used a convenience sample that was more highly educated, white, and generally had a higher income level than the general population of the United States. A sample with a different demographic composition would likely have been assigned different kinds of inferences. In addition, people who saw sensitive or uncomfortable inferences when they viewed their inference web page before saving it may have chosen not to complete the survey, and so people with characteristics that may make them more likely to be assigned sensitive inferences may not have participated. Our study was also U.S.-focused, and people in other countries and from other cultures might have different inferences, and different reactions to the inferences, than we observed in our sample.

In the interviews, we asked participants to react to and think about inferences that most were not aware of before the study. This means that participants' reactions were what first came to mind as they processed and thought about the inferences, which most of them had not spent much time doing before. These initial reactions are based on their existing mental models—the knowledge, beliefs and understandings of cause and effect related to data collection and use—that were present before the interviews. This means that our method was able to identify the kinds of thinking and reactions that might occur when people encounter information for the first time about the kinds of inferences a platform makes about users, at the time when they are choosing whether to consent to its terms or not. But, we cannot know from these interviews how participants' thinking about the inferences could evolve over time, or how they might have reacted if they had been assigned different inferences or if the platform had revealed more/less about the inferences it makes. Also, our data do not allow us to conclusively enumerate all of the different types of inferences, data, and mechanisms for deriving inferences that participants were aware of and knowledgeable about, since we just asked them about the specific inferences assigned to them by a single platform.

This research was approved by our institution's IRB. Consent was obtained separately for the survey and follow-up interviews. The inferences files were parsed and only the labels for the inferences were stored in our database, associated with a random participant identifier string. No other information was gathered from the files. Previous research about Facebook's inferences has found that only a very small proportion (1.66%) are potentially sensitive, although the same study provided evidence that 73% of EU Facebook users in their sample had been assigned at least one sensitive inference [4]. Through piloting, we determined that participants would be very unlikely to feel that any of the inferences in the files were something they would be uncomfortable discussing in an interview. Inferences about survey respondents who did not complete the survey were deleted from our database. Also, both platforms provide controls that allow the user to "turn off" (Google) or "remove" (Facebook) inferences from the visible list. However, our testing showed that these deactivated inferences were present in participants' files. We did not store deactivated inferences in our database, nor did we ask questions about them in the survey.

4 Findings

4.1 Inferences Differed Across Platforms

The Facebook and Google inferences differed in number and in character for our survey respondents. Facebook had assigned fewer inferences to each respondent in our survey sample (*Mean*=12, *Max*=33) than Google had (*Mean*=66, *Max*=131). In addition, there were more unique inferences assigned by Google across all of our survey respondents than by Facebook (Google: 561; Facebook: 110), indicating that Google's inferences might appear to be more highly personalized to end users based on the information that is provided in the interface. However, Facebook survey respondents reported that they felt the inferences were more accurate, on average, than Google respondents (Facebook: *Mean*=4.64, *SD*=1.88; Google: *Mean*=3.84, *SD*=2.05).

The most common inferences assigned to Facebook respondents in our sample had to do with how respondents accessed Facebook. For example, "WiFi users" and "Mobile network or

device users" were two of the most common Facebook inferences in our data, both assigned to 42 of 44 Facebook survey respondents. All Facebook interview participants found it hard to react to inferences like these, because they seemed to be about obvious facts that the participants easily recognized as being related to themselves and therefore did not dispute. (Where participants' reactions were about a specific inference, we have included the inference in italics immediately before the quote.)

> *[Facebook access (browser): Chrome; Birthday in August]* I use Chrome. That's when I was born. –P01, Woman, Facebook

> *[WiFi users; Gmail users]* I use wifi, I use Gmail. –P10, Man, Facebook

In contrast, the most common inference among Google survey respondents was "Parenting" (46 out of 51 Google respondents) followed by "Home & Garden" (43) and "Shopping" (41). Interview participants asked about their Google inferences typically had a lot more to say about them, because the Google inferences seemed more descriptive and personally relevant. For example:

> *[Vehicle Shopping; Autos & Vehicles]* Yeah, I've been doing that recently for sure. I'm kind of in the market for a different vehicle. I mean, yeah I guess autos and vehicles probably went along with vehicle shopping... actually I was just, before I came here I was Googling how to replace a headlight in one of our cars. So yeah, I use that a lot. –P12, Man, Google

Google interview participants reacted in a similar manner as the Facebook respondents did to the inferences that seemed to be facts about themselves, e.g., "I am male, I'm over 65. No problems with that." (P07, Man, Google) and "Mobile phones, I'm sure that's because of Sprint." (P09, Woman, Google). However, the majority of Google inferences felt more personal to interview participants, and potentially descriptive of their interests. There were many Google inferences for genres of music, travel destinations, hobbies, sports, types of news (e.g., "American Football", 33 participants; "Gourmet & Specialty Foods", 31; "TV Talk Shows", 30). See Appendix B for examples of the most and least common inferences among our survey respondents.

Sixty-six percent of Facebook survey respondents and 80% of Google respondents had not seen their inferences page before. Only 9% of Facebook respondents and 10% of Google respondents said they had seen the page; the remaining respondents from both platforms were unsure. In addition, only 2 of the 21 interview participants were among those who indicated they'd seen their inferences page before the survey.

4.2 Inference Reactions and Explanations

Our analysis of the interview transcripts identified characteristics of participants' reactions to the inferences assigned to them that revealed three broad themes. The themes differed in terms of the types of evidence participants described to support their thinking about how the inferences connected with their self-perceptions, and in terms of the degree of connection they believed was present. We focus below on describing these themes, and characterizing participants' explanations of how they thought the inferences were connected to themselves. These explanations are important, because they illustrate participants' understanding of how the platforms' inferences relate to them and their lives. In general, generating explanations helps people to develop and evolve their understanding of how and why things work the way they do [15, 33]. As such, characterizing participants' reactions and explanations is important for identifying how people might anticipate what kinds of inferences are possible in these and other platforms and systems. See Table 3 for a summary of definitions of important concepts used in this section.

Inference:	Information that a platform has automatically assigned to a person that is derived from data the platform has collected or obtained.
Reaction:	A person's sense of the extent to which an inference seems to apply to them, focused on the relationship between the inference and themselves.
Explanation:	Interpretations that involve causal relations, about why and how an inference was assigned.
Plausible:	The inference was believable, and participants provided specific evidence or explanations supporting why it made sense to them.
Implausible:	The inference was initially not believable, but participants subsequently provided an explanation justifying why it might be related to them in some way.
No Connection:	The inference did not make sense, and participants were certain that it did not apply to them.

Table 3: Summary of the conceptual framework used to understand participants' reactions to the inferences.

4.2.1 'Plausible' Reactions Were Grounded

The most common reaction to the inferences from both Facebook and Google participants was certainty that the inference was related to themselves and their interests. We refer to this type of reaction as a *plausible* reaction, because the inference was believable to them. The example below illustrates what a plausible reaction looked like:

> *[Parenting]* I think just because it's a main part of my life. It makes sense... it's I guess obvious that it would be on there. I feel like a good majority of stuff that I'm doing online outside of work is stuff for the kids. Like Googling summer camps and trying to figure out activities and I think of all this stuff that I'm randomly looking up online as kid related... –P21, Woman, Google

Participants had this type of reaction when inferences seemed to them to be grounded in what they could remember about their past actions and perceptions of their present activities, interests, and demographics. They felt like these inferences made sense to them, and they could produce specific evidence or explanations for why this was, without hesitation. Fifty-five percent of the reactions to inferences that were discussed by participants in the interviews (176 of 321) were *plausible*, and every participant had at least one plausible reaction to an inference; the average number of plausible reactions per participant was 8[5].

In a large proportion of the plausible reactions, the participant gave evidence of their past activities on the platform to support their feeling that the inference made sense for them. This evidence looked mostly like general recollections about the kinds of topics they search for (Google) or about typical content of posts they had made or articles they had read (Facebook). Nineteen participants had plausible reactions similar to these:

> *[TV Comedies]* How can I explain... I do view, they're my favorite genre of television. And [this inference] describes, I guess, my everyday activities probably more so than other things. I guess probably, I do cast searches on occasion, things like that. –P05, Man, Google

> *[Parents (All)]* Because I have three kids, well, a lot of my posts are about my kids. –P13, Woman, Facebook

Other plausible reactions, described by 11 participants, seemed related to memories of specific recent past searches. For example, in reaction to the inferences "Flooring" and "Lamps & Lighting", one participant described how she and her family had just been shopping for lamps online and had their floors replaced as part of remodeling their home (P22, Woman, Google). Another participant talked about how she was in the process of selling her house and was doing some searching online about appraisals ("Property Inspections & Appraisals"; P08, Woman, Google).

A third very common form of plausible reaction, mentioned by 14 participants, talked about the participant's general interests or how they spent a lot of their time as an explanation for why the inference made sense to them, but without mentioning any specific online activities. For example, P17 related one of her inferences to her volunteer activities, but did not mention online searching or web browsing:

> *[Dogs]* Because ninety-nine point nine percent of anything I do online outside of work is probably dog related. I volunteer at an animal shelter and I volunteer for a rescue. And I foster for them. –P17, Woman, Google

[5]Note that some participants commented on multiple inferences at once, and these were all coded together as one reaction. All participants only discussed a subset of the inferences present in their reports. The counts, averages and percentages reported in this section are only intended to convey relative prevalence within our sample.

There was little difference in the proportion and type of plausible reactions between Facebook and Google participants, with one exception: Facebook participants were more likely to mention inferences as related to their Facebook Friends or engagement with content on Facebook, whereas Google participants were more likely to mention search history. The two examples below illustrate this difference:

> *[Close Friends of Women with a Birthday in 7-30 days]* That's based on my friends list in Facebook –P13, Woman, Facebook

> *[Cooking & Recipes]* Yeah, all of my recipes come from the Internet, so same thing. I'm googling food a lot. –P22, Woman, Google

Another pattern we noticed in the data was that participants were more clear in their plausible reactions about describing the types of online activity they believed the inferences were based on, and therefore the types of data they understood the platforms to have access to, than any specifics about technical mechanisms that gave the platforms access to these data. When specifically probed about how they thought the platforms were able to assign specific inferences to them, participants' explanations were vague and high level, conveying notions of access and visibility but not detailed speculation about transmission or collection. Below are a few examples of this.

- *P13, Facebook:* Based on the content that I engage with and the pages that I follow
- *P16, Facebook:* Facebook has some way of knowing
- *P03, Google:* It'll look at everything from your email history to your search history
- *P22, Google:* Google sees it in my searches

As these findings show, participants had a *plausible* reaction to an inference when they were certain that it was relevant to them, because they could explain its' relevance using evidence they felt was true about themselves. Explanations about what caused plausible-seeming inferences to be assigned were constrained by participants' ability to identify potential reasons for this, related to the person's self-perceptions and platform use.

4.2.2 'Implausible' Reactions Were Rationalized

The second theme that emerged from our coding of participants' reactions to the inferences was uncertainty about why the inferences had been assigned, followed quickly by an explanation rationalizing a possible source for the connection between the inference and themselves. We refer to this type of reaction as an *implausible* reaction, because these inferences initially were difficult for participants to believe or relate to. But, after further consideration, the participant identified a reason that the inference may have been assigned. For example, below is a participant's reaction to an inference assigned to her by Facebook:

> *[Lived in United States (Formerly Expats - United States)]* Well, I still currently live here.

> So those were pretty inaccurate to me... I'm sure they're collecting information on where I'm at. I do travel a lot for work so I can see how they would get that. –P02, Woman, Facebook

This participant initially rejected the idea that the inference was related to her, but then backtracked, explaining that perhaps her recent location history had something to do with why this inference had been assigned. It was as though, for these inferences, no immediate reason it would have been assigned to the participant came to mind. But, then a realization occurred that something about their past history, behavior online, or relationship with someone might be interpreted as supporting the inference, even if it seemed like a stretch to them. The explanations participants generated as part of implausible reactions to inferences were more speculative, and less convincing to themselves, than the explanations they had given for the inferences for which they'd had plausible reactions. Twenty-eight percent of the reactions we coded (89 of 321) were like this, and 19 out of 21 participants had an implausible reaction to at least one of the inferences that had been assigned to them. On average, participants had 5 implausible reactions to inferences.

The implausible reactions for which clear exceptions did not come to mind for participants formed a spectrum of perceived separation between the participant's self-perception and the inference. The most related inferences were those participants felt used to be true about them, but were not true any longer, and six participants reacted in this way:

> *[Groupon]* Yeah, when I first got into that, I used it for golf stuff, but I haven't done that in quite some time. It's like these things never go away. –P07, Man, Google

The next level of separation, slightly more distant than an outdated inference, had to do with inferences that the participant felt were related to their close family members or friends, but not themselves. They believed these inferences had to do with the specific interests of those people, or with searching they had done for things they themselves were not actually interested in, but people they were close to were. Seven participants had a similar reaction to an inference:

> *[Coffee & Tea]* I don't drink coffee. My husband does though. So again maybe, you know? –P18, Woman, Google

> *[Fitness]* I don't do a lot with fitness online, so I'm not sure why they came up. Although I do go to soccer.com a lot and do stuff a lot related to soccer because of my children, maybe that's why that's there? –P22, Woman, Google

The third and most distant characterization of the relatedness of inferences that seemed implausible is a little bit more far-fetched, in that the participants imagined that the inference might just be assigned to everyone in a certain demographic. P03 provided an example of this, in reaction to the inference "Parenting" (8 participants reacted similarly):

> *[Parenting]* I don't have any kids and so I was like, why in the world did parenting come up... I mean they can determine your age bracket pretty quickly and if they feel like, okay that should be something that's relevant to majority of people in that age bracket, we're going to add it or something... –P03, Man, Google

Participants' *implausible* reactions occurred when they were unsure about whether an inference applied to them or not, and then justified the presence of a connection with weak, speculative evidence. This evidence was not some recent fact or belief about themselves; rather, it relied on things they were knowledgeable about like their family members' behavior and interests to contextualize the inferences. The constraints on explaining a connection with an implausible-seeming inference involved having an awareness of one's past self, as well as of other people one is close to, or similar to in the case of demographic-based explanations for these inferences.

4.2.3 'No Connection' Reactions Were Unrecognizable

The third theme among the reactions our participants had to the inferences was that they seemed certain some inferences did not apply to them. We call these *no connection* reactions, because participants could not identify a reason why these inferences had been assigned to them. The two examples below, from a Facebook participant and a Google participant, illustrate this type of reaction:

> *[Multicultural Affinity: African American (US)]* Well, I'm not African American... maybe it's something that got put in, maybe it's an unintentional error... or if it is based on an algorithm, then somebody's got some flaws in their system. –P10, Man, Facebook

> *[Rothy's]* I looked at Rothy's because an ad popped up. I think it's shoes. If I remember right it's shoes, and I looked, and I can't buy these. These are way out of my price range. Why are they targeting me? I understand targeting for Walmart. I don't understand targeting me for Rothy's because I am not going to pay 200 dollars for a pair of shoes. –P08, Woman, Google

Participants felt these inferences were unrecognizable, because they had no connection with their own past behavior or self-perceptions, and because they could not imagine other ways to justify why those inferences might have been assigned. Participants could not see themselves in them, and sometimes expressed surprise or disbelief as part of their reaction to them. In addition, participants could not produce evidence or an explanation, as they could in their plausible and implausible reactions, supporting that inference's relationship to themselves. Instead, the evidence and reasons they provided were about why these inferences were *not* connected to them. This type of reaction appeared least often in our data: 56 out of the 321 reactions (17%) that we coded were like

this, and 19 out of 21 participants reacted in this way to an inference at least once (*Mean*=3).

There were several ways in which participants described the lack of connection between themselves and these inferences. Their descriptions illustrate the kinds of evidence and reasoning they were using to understand what the platforms were basing the inferences on. Most often, participants differentiated their activities and interests from their understanding of the meaning of the inference. For example, there were a few inferences for which participants provided a strong refutation in terms that conveyed they never, or hardly ever, do anything like what the particular inference implied about them:

> *[Golf Equipment]* I don't know why that's on there... golf equipment? I don't golf. That's weird. –P04, Woman, Google

> *[Convenience Stores]* Except that I never go [to convenience stores], unless I have to use the restroom. I never go into the store part to buy something. –P09, Woman, Google

Participants also reacted to some of these inferences not by refuting specific activities that the inference implied, but by refuting the idea that they'd be interested in anything related to the inferences at all. Below, a Facebook and a Google participant both describe how their interests do not align with certain inferences:

> *[US politics (moderate)]* I'm just not interested in politics at all, really. Once in a while I'll post something that I find funny that I know a little bit about, but as far as getting down to the truth of politics and stuff, I just don't care. –P16, Woman, Facebook

> *[Apparel]* I mean, I'm not really interested in fashion or clothing all that much, other than as a utilitarian pursuit, something that you need. I'm not particularly concerned about brands or anything like that. –P19, Man, Google

There were some inferences that participants were so confused by that it was difficult to react to them. For example, this was one participant's reaction to such an inference:

> *[Combat Sports]* And combat sports, I have no clue. –P18, Woman, Google

In their *no connection* reactions, participants were unable to establish a causal relation between themselves—their past behavior, interests, relationships, demographics, etc.—and the inferences. Without evidence supporting an interpretation of the inference as plausible, it was difficult for participants to reason about why these inferences were associated with them.

4.3 "Household Income" and Evidence of Aggregation

One particular type of inference made by Facebook provided a unique glimpse into participants' beliefs about how information that might not be readily ascertained from their web use behaviors could be taken into account (or not) when assigning inferences to platform users. Facebook assigned inferences about "Household income: top X% of ZIP codes" to 7 interview participants (and 13 out of 44 Facebook survey respondents). The 'X%' was either 10%, 10%-25%, or 25%-50%. Five interview respondents chose to talk about this inference during the interview. Four participants had *implausible* reactions, and one had a *no connection* reaction. For example, here is the implausible reaction that P13 had to this inference:

> *[Household income: top 10%-25% of ZIP codes (US)]* The household income thing... because that's not something that's public on your Facebook profile. So I'm not really sure where they got that information. That's kind of weird... and we just recently bought a house. So I'm wondering if they somehow got our income information from our mortgage, from our bank. But yeah, that's kind of weird. I mean, I understand why they would want that. But they're digging pretty deep to get that information. It's not readily available on my Facebook. It's like where did you get that? –P13, Woman, Facebook

This participant essentially said that the inference is implausible because Facebook should not be able to figure out her household income from the information that the platform had about her. We (the research team) interpreted this inference type to be indicative not of information about the participants' income specifically, but about the aggregate income level in the zip code in which the participants were located when they connected to Facebook most often. However, in addition to P13, the other four respondents with this inference also seemed to have made the same assumption that it was about their specific household income rather than about incomes in general in their location. Only P11 acknowledged the comparative aspect of this inference, while at the same time relating it to his own income level:

> *[Household income: top 10% of ZIP codes (US)]* I'm not in the top 10% although I live in a community that probably is in the top 10%... I live in [city] so I would consider that would be probably a high income zip, right? So I live there but I don't earn the top 10%, does that make sense? –P11, Man, Facebook

The reactions to the "Household income" inference are an interesting special case among all of the reactions in our data, because of the way this inference was presumably derived. It appears at face value, to an outside observer, to be about the zip code in which the participant lives, and not necessarily about their specific income situation. Our participants did not think this inference applied to them, but this was because they reacted to the inference as if it was about them and not about the community in which they live. The evidence they relied on to establish a connection (or not) between themselves and the inference involved their own beliefs about how wealthy they are compared to others. This interpretation of the inference was likely evoked by "top percent" framing. Therefore, the constraints on explaining this particular inference involve

participants' assumptions that all of the information in their list of inferences is about them personally, as an individual human being.

4.4 Acceptance and Discomfort Coexist

After discussing the specific inferences in their reports, participants were asked about what they thought the purpose of the inferences was. All 21 participants stated that they believed that the purpose of the inferences was primarily for advertising. This is not surprising, since both platforms use the word "ad" in their presentation of the inferences to users. While some participants drew on their personal experiences and knowledge when relating the inferences to the ads they saw online, the majority of participants appeared to be mainly describing a vague awareness gained through hearsay or news coverage. Some participants talked about seeing advertisements that were directly relevant to their search history and online activity as they were browsing, but did not offer overly technical explanations. For example:

> Google's like probably getting our search histories or whatever and then maybe selling them to companies and that's how like the ads pop up on the side of your screen, like that these may interest you. –P18, Woman, Google

Eleven participants explicitly stated their belief that inferences are part of the platforms' business model, and assumed a profit motivation for the companies. For example, P21 (Woman, Google) observed that since they don't pay to use Google's services, it seems reasonable that the platform must have an alternative way of earning revenue to support its operations. And P11 said:

> I'm assuming they're monetizing it in some way. They're probably selling it to companies so that they can target ads to me, the consumer, to everybody else as consumers. – P11, Man, Facebook

This is in line with previous research that has found a general awareness of targeted advertising, even if users are not necessarily able to articulate the mechanism behind it [5, 29, 35].

While the majority of participants (17 out of 21) appeared to accept the idea of information collected about them being used for advertising, 20 participants also expressed feelings about the inferences which indicated a level discomfort with some of them. Acceptance of the business model and discomfort with specific inferences were not mutually exclusive, and reactions of discomfort and acceptance were often even expressed in the same sentence. P07 referred to the "evil empire" and asserted that he did not like being the target for ads, but also did not take take great offense to receiving targeted ads. Some participants explicitly referenced the idea of the "world" (P01) or "most people" (P22) during their interviews, suggesting that data collection is something that's going on at a grand scale over which they felt they had little control. This seemed to play a role in their willingness to accept information being collected about them, but it was also a source of discomfort. For example:

> [It makes me uncomfortable] that they can get so many specific things without me realizing that people have it. But, I kind of know that. I wish they didn't know anything. I would rather be anonymous, but I know that's not our world. –P01, Woman, Facebook

> I think most people feel the same way that I do. I mean, there's a level of acceptance that it happens, but then the more you think about it, it kind of starts to disturb you a little bit more because when you see lists like this, then some of the things that they make connections for, it takes you a while in your head to get to how they got to that. And it's a little bit, I don't know, disconcerting or something. It's just a little bit uncomfortable. –P22, Woman, Google

Highly specific inferences which participants could connect to non-typical daily activities also provoked speculation and reactions of discomfort. Participants' comments about these inferences suggested the discomfort came from being made explicitly aware of the extent and detail to which their actions are being tracked by platforms. This also served as an opportunity for participants to concretely self-reflect on their online activities to an extent they appeared to not have previously done before. Although they knew information was being collected, and sometimes even had experiences seeing targeted ads that allowed them to deduce thair their shopping habits were being tracked in some way, it was still different to see the information aggregated in one place.

> On the one hand, I understand it, but at the same time it kind of makes you think like, wow, they really can actually key in on very specific things... everything else was so broad and then that one [inference] really stuck out because it was the one thing out of everything that was very, very specific to a span of time of search history... I mean, it's not like it's so personal that I'm freaking out about it. It's just that it brought to my attention how much data is floating around out there that Google or Facebook or whoever is able to capture and is aware of about you that you might've even long since forgotten about. –P03, Man, Google

> [It made me feel] a little bit in the spotlight. A little bit like yeah, they know what I'm doing. A little uncomfortable I guess... I'm not surprised, but you just don't think about what you're leaving, the tracks that you're leaving. You don't think about it when you're doing it. –P08, Woman, Google

Some participants appeared to be uneasy about the breadth and variety of inferences associated with them, indicating a platform's ability to know things about them to an extent nobody else does. This feeling appeared to be related to inferences that participants considered to be accurate or relevant to themselves, suggesting that accurate information can feel

uncomfortable when a great amount of it is visibly collected in one place:

> As I went through this I was like, oh my God look at all the information they've gathered about me... Google knows more about me than anybody else. It's scary. –P17, Woman, Google

> No matter like where you... They track your movements or they know what you've searched for or what you're interested in. So I mean probably my husband doesn't even know that, right? So I find it a little creepy. –P15, Woman, Facebook

In contrast to this, some participants expressed little discomfort with the existing inferences. One reason commonly cited was that the categories appeared to be harmless, or sufficiently generic that it could not reflect badly on them or compromise their personal information. For example, P08 (Woman, Google) talked about not caring what information the platforms have collected about her, because "I don't think that it's necessarily done to hurt people."

5 Discussion

Our findings focused on participants' reactions to the inferences Facebook and Google had made about them, and the evidence they used in their explanations for how and why the inferences were related to them. Generating explanations is one way that people increase understanding of phenomena they encounter, by helping them to predict and make sense of future events and situations [15]. Participants' explanations for the inferences they were assigned provided visibility into the knowledge they used to try to make sense of the inferences, and the new understandings that resulted [33].

5.1 Three Ways of Framing Inferences

Our findings show that the understandings participants demonstrated of the inferences assigned to them are constrained by their own awareness of the information the inferences might be generated from, at the time they are considering the inferences. P14 actually had a moment of insight related to this during his interview:

> I have a narrow thought process. For example, when we talked about sports. The idea that I'd be thinking about sports I play and not the sports I might watch or might have watched. –P14, Man, Google

This insight happened when the participant was asked by the interviewer to compare the information in the inferences with the kinds of information that a person would normally enter when filling out an online profile. In response, he described his thinking when he had reacted to the sports that were listed among his inferences (e.g., "Basketball," "Hockey," "Golf"). He indicated that his understanding of the meaning of these inferences was initially narrower—framed by his recollection of sports he plays—than what the inference might actually represent to the platform.

There were three broad framings which participants used to reason about their connection to the inferences assigned to them. The inferences were framed as 1) being related to past and present online behavior and interests, 2) stemming from their relationships with others they are close to, and 3) inaccurate, and therefore not useful. The extent to which participants recollected or speculated about evidence consistent with one of these framings determined whether they had a *plausible*, *implausible*, or *no connection* reaction to an inference.

5.1.1 Past vs. Future

The most prevalent framing, that inferences are related to online behavior and interests, demonstrates an assumption by our participants that inferences are descriptive of their pasts. But, advertisers use inferences to exert control over people's future attention and actions through the ads that are targeted to them [36]. This means that the goal of inferences is more about labeling people so that they can be targeted, rather than creating a representation that agrees with how people would describe themselves. If people assume that inferences are descriptive rather than predictive, that assumption leads to explanations and understandings that fail to anticipate the true purpose of the inferences and could lead to people deeming inferences as inaccurate or inapplicable that could actually be effective for targeting by the platforms and institutions that use them.

One aspect of the platforms' user interfaces that may reinforce this impression, ironically, is the explanations of the inferences provided by the platforms themselves. Both Google's and Facebook's help webpages describing how ads are targeted focus on the data the inferences are based on, e.g. "Adding a product to a shopping cart or making a purchase" (Facebook)[6] or "Previous search activity" (Google)[7]. One way to convey the predictive nature of inferences to end users could be to instead use language that describes the inferences as predictions or guesses. It might also be possible for platforms to provide an accuracy score or indicator for each inference to indicate how well each user's behavior since the inference was assigned aligned with the prediction. This might convey a different framing that emphasizes an orientation toward the future rather than the past.

5.1.2 Individual vs. Aggregate

The second framing, that inferences are related in some way to people's close relationships, shows that our participants were constrained in how they reasoned about about aggregation as it related to how inferences were assigned. Our participants said that they sometimes did things online that reflected others' interests and not their own, and felt that this could result in the interests of people they were close to being be mis-attributed to them. However, this framing is still based on the unit of analysis being the individual person, and thus their belief

[6] https://www.facebook.com/ads/about/?entry_product=ad_preferences
[7] https://support.google.com/accounts/answer/1634057

that the mistake in generating the inference was related to mapping the interests of one person onto another.

There were a few instances where participants felt like interests were mapped by the platform onto broad groups of people according to "demographic" similarities—like "Parenting" being associated with participants that did not have children living at home with them or had no children (P03, P04, P08, P11, P12, P14, P19). But the demographic indicators they mentioned in rationalizing these inferences were the ones that were most salient to them, e.g., age, gender, and income level. This means that the framing of inferences as related to close relationships or similar others according to obvious demographic characteristics underestimates the ability of machine learning models to infer other features that don't fit into this framing. Thus, targeting is likely based on attributes that people cannot envision [2].

The key insight absent from participants' reasoning consistent this framing was that behavioral patterns of complete strangers can be used to infer and assign attributes to them. Facebook and Google both reinforce an understanding of inferences as belonging to individuals rather than being derived from aggregate patterns of characteristics by using language like "your activity" in the platforms' explanations. Providing some information about the proportion of other users who share inferences in common with an individual could help to convey a broader perspective of the similarities across users that the inferences are based on.

5.1.3 Accurate vs. Useful

The third framing, that inferences that are not accurate or explainable are not useful, indicates a presumption that the inferences should be accurate, or that the intention of the inferences is to describe users accurately. There were several examples of this in our data; the reaction to the "Rothy's" inference on page 8 is one. This framing assumes that platforms should not want to associate inaccurate inferences with users. However, it is possible that mis-targeting a few people could still be beneficial for platforms' goals in aggregate. Of course, we cannot know from this study whether the platform or advertisers think the inferences are either accurate or useful, but one can imagine that a particular inference doesn't need to be 100% accurate for all users it is assigned to in order to serve its intended purpose from the platform's perspective. This means that a framing that inferences should be accurate prevents people from speculating about situations in which inaccurate inferences might actually be useful or even intentional. One way that platforms could attempt to convey the notion that even inaccurate inferences might be beneficial from their perspective is to include information about the overall profitability (or other platform success metric) of various inferences that a user has been assigned, along with the accuracy rating mentioned above. This could serve to make the platform's stake in the inferences more transparent.

5.2 Implications for Privacy Decisions about Consent

Users' implicit assumptions and guesses about the data collection and inference generation they are consenting to may vary based on their existing knowledge and understanding of cause and effect related to the consent decision they need to make. For example, people cannot envision inferences that may be related to their potential future interests, or derived via aggregation with data from people they don't know, or that are inaccurate from their perspective but still useful for the platform. And yet, notice and choice demands that users provide one-time, up front consent to incomplete descriptions of platforms' data practices. New privacy legislation mandates improved transparency (while retaining the notice and choice model), but our findings show that the framing by which people explain and understand the inferences does not support them in envisioning the inferences from the platforms' perspective.

Without framing the inferences from the perspective of advertisers and platforms and imagining how they might put the inferences to use, people are unlikely to be able to give informed consent to many of the inferences they are assigned. Knowing the inferences are used for advertising is not enough for consent, because in our study participants were broadly aware of that and still did not achieve that understanding. Hence the discomfort when they began to realize the amount and specificity of the information the platform must have amassed about them. Based on this work, we argue that the understandings people develop about inferences through the 'Ad Settings/Preferences' model of transparency are unlikely to help them realize that what they are really consenting to is allowing the platform to make whatever inferences it wants to about their future, to target them based on the behavior of people they don't even know exist, and to profit from inaccurate assumptions about them.

Our study shows that there are types of inferences which are straightforward for people to understand and anticipate: inferences that seem *plausible* because of their relationship to easily recalled or past actions. Platforms which place a priority on obtaining true consent should therefore restrict the inferences they make about users to those which fit the constraint of being explainable by users, given their normal use of the platform. It may be possible to expand user understanding of inferences through framing them differently in the ways suggested in this paper. Future research is needed to explore the effectiveness of lightweight interventions such as these for expanding the range of inferences that seem plausible to users.

Acknowledgments

We thank Joe Freedman and Rick Wash for their assistance with developing the software and Madison Heise for conducting some of the interviews. We also thank the BITLab @ MSU research group for their feedback. This material is based upon

work supported by the National Science Foundation under Grant No. CNS-1524296.

References

[1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, and et al. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Comput. Surv.*, 50(3), 2017.

[2] Athanasios Andreou, Márcio Silva, Fabrício Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. Measuring the Facebook Advertising Ecosystem. In *Network and Distributed Systems Security Symposium (NDSS)*, 2019.

[3] Sophie C. Boerman, Sanne Kruikemeier, and Frederik J. Zuiderveen Borgesius. Online Behavioral Advertising: A Literature Review and Research Agenda. *Journal of Advertising*, 46(3):363–376, 2017.

[4] José González Cabañas, Ángel Cuevas, and Rubén Cuevas. Unveiling and Quantifying Facebook Exploitation of Sensitive Personal Data for Advertising Purposes. In *Proceedings of the 27th USENIX Security Symposium*, pages 479–495, 2018.

[5] Farah Chanchary and Sonia Chiasson. User Perceptions of Sharing, Advertising, and Tracking. In *Proceedings of the Eleventh Symposium On Usable Privacy and Security (SOUPS)*, pages 53–67, 2015.

[6] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies*, pages 92 –112, 2015.

[7] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting Depression via Social Media. In *Proceedings of the 7th International Conference on Web and Social Media*, 2013.

[8] Michael A. DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. How People Form Folk Theories of Social Media Feeds and What It Means for How We Study Self-Presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[9] Claire Dolin, Ben Weinshel, Shawn Shan, Chang Min Hahn, Euirim Choi, Michelle L. Mazurek, and Blase Ur. Unpacking perceptions of data-driven inferences underlying online targeting and personalization. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[10] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. First I "like" It, Then I Hide It: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2371–2382, 2016.

[11] Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. Communicating Algorithmic Process in Online Behavioral Advertising. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2018.

[12] Casey Fiesler and Blake Hallinan. "We Are the Product": Public Reactions to Online Data Sharing and Privacy Controversies in the Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2018.

[13] Harold Garfinkel. Studies of the Routine Grounds of Everyday Activities. *Social Problems*, 11(3):225–250, 1964.

[14] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. Demographic Prediction Based on User's Browsing Behavior. In *Proceedings of the 2007 World Wide Web Conference*, 2007.

[15] Frank C. Keil. Explanation and understanding. *Annu. Rev. Psychol.*, 57:227–254, 2006.

[16] Predrag Klasnja, Sunny Consolvo, Tanzeem Choudhury, Richard Beckwith, and Jeffrey Hightower. Exploring Privacy Concerns about Personal Sensing. In *Proceedings of the International Conference on Pervasive Computing*, pages 176–183, 2009.

[17] Michal Kosinski, David Stillwell, and Thore Graepel. Private Traits and Attributes are Predictable from Digital Records of Human Behavior. *PNAS*, 110(15):5802–5805, 2013.

[18] Oksana Kulyk, Annika Hilt, Nina Gerber, and Melanie Volkamer. This Website Uses Cookies: Users Perceptions and Reactions to the Cookie Disclaimer. In *Proceedings of the 3rd European Workshop on Usable Security*, 2018.

[19] Pedro Giovanni Leon, Justin Cranshaw, Lorrie Faith Cranor, Jim Graves, Manoj Hastak, Blase Ur, and Guzi Xu. What Do Online Behavioral Advertising Privacy Disclosures Communicate to Users? In *Proceedings of the ACM Workshop on Privacy in the Electronic Society*, pages 19–30, 2012.

[20] Pedro Giovanni Leon, Blase Ur, Yang Wang, Manya Sleeper, Rebecca Balebako, Richard Shay, Lujo Bauer, Mihai Christodorescu, and Lorrie Faith Cranor. What

Matters to Users? Factors That Affect Users' Willingness to Share Information with Online Advertisers. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, 2013.

[21] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. Internet Jones and the Raiders of the Lost Trackers: An Archaeological Study of Web Tracking from 1996 to 2016. In *Proceedings of the 25th USENIX Security Symposium*, 2016.

[22] Chanda Phelan, Cliff Lampe, and Paul Resnick. It's Creepy, But It Doesn't Bother Me. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 5240–5251, 2016.

[23] Emilee Rader. Awareness of Behavioral Tracking and Information Privacy Concern in Facebook and Google. In *Proceedings of the Tenth Symposium On Usable Privacy and Security*, pages 51–67, 2014.

[24] Emilee J Rader and Janine Slaker. The Importance of Visibility for Folk Theories of Sensor Data. In *Proceedings of the Thirteenth Symposium on Usable Privacy and Security*, 2017.

[25] Ashwini Rao, Florian Schaub, and Norman M. Sadeh. What do they know about me? Contents and Concerns of Online Behavioral Profiles. In *Proceedings of the 6th ASE Conference on Privacy, Security, Risk and Trust*, 2014.

[26] Johnny Saldaña. *The Coding Manual for Qualitative Researchers*. Sage, 2015.

[27] H. Jeff Smith, Sandra J. Milberg, and Sandra J. Burke. Information Privacy: Measuring Individuals' Concerns about Organizational Practices. *MIS Quarterly*, 20(2):167–196, 1996.

[28] Daniel J Solove. Privacy Self-Management and the Consent Dilemma. *126 Harvard Law Review*, pages 1880–1903, 2013.

[29] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. Smart, useful, scary, creepy: Perceptions of online behavioral advertising. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, 2012.

[30] Kaan Varnali. Online Behavioral Advertising: An Integrative Review. *Journal of Marketing Communications*, 2019.

[31] Jeffrey Warshaw, Nina Taft, and Allison Woodruff. Intuitions, Analytics, and Killing Ants: Inference Literacy of High School-educated Adults in the US. *Proceedings of the Twelfth Symposium On Usable Privacy and Security*, 2016.

[32] Ben Weinshel, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L. Mazurek, and Blase Ur. Oh, the Places You've Been! User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Inferencing. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 149–166, 2019.

[33] Robert A. Wilson and Frank C. Keil. The Shadows and Shallows of Explanation. In Frank C. Keil and Robert A. Wilson, editors, *Explanation and Congnition*, pages 87–114. MIT Press, 2000.

[34] Yuxi Wu, Panya Gupta, Miranda Wei, Yasemin Acar, Sascha Fahl, and Blase Ur. Your Secrets Are Safe: How Browsers' Explanations Impact Misconceptions About Private Browsing Mode. In *Proceedings of the 2018 World Wide Web Conference*, pages 217–226, 2018.

[35] Yaxing Yao, Davide Lo Re, and Yang Wang. Folk Models of Online Behavioral Advertising. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1957–1969, 2017.

[36] Shoshana Zuboff. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile Books, 2019.

Appendix

A Survey Questions

Participants were provided with a description of the survey and asked for consent before answering screening questions. If eligible, they were directed to the beginning of the survey. Respondents were informed about how long the survey would take, that they would be asked to log in to Facebook or Google and save and then upload a file, that there were attention check questions, and that they were free to withdraw from the study at any time.

A.1 Demographics and Screening

1. What gender do you identify as?
   - Man
   - Woman
   - Other (fill in the blank)
   - Prefer not to answer
2. What is the last grade or class you completed in school?
   - None, or grades 1-8
   - Some high school
   - High school graduate or GED certificate
   - Technical, trade, or vocational school AFTER high school
   - Some college, no 4-year degree
   - 4-year college degree
   - Some postgraduate or professional schooling, no postgraduate degree
   - Postgraduate or professional degree, including master's, doctorate, medical or law degree
3. What was your total household income before taxes during the past 12 months?
   - Less than $25,000
   - $25,000 to $34,999
   - $35,000 to $49,999
   - $50,000 to $74,999
   - $75,000 to $99,999
   - $100,000 to $149,999
   - $150,000 to $199,999
   - $200,000 or more
4. Which categories below best describe you? Select all that apply:
   - White
   - Hispanic, Latino or Spanish
   - Black or African American
   - Asian
   - American Indian or Alaska Native
   - Middle Eastern or North African
   - Native Hawaiian or Other Pacific Islander
   - Some Other Race, Ethnicity or Origin (please specify)
5. What is your current employment status?
   - Employed full time
   - Employed part time
   - Unemployed looking for work)
   - Unemployed not looking for work
   - Retired
   - Student
   - Disabled
6. Have you ever received formal training in computer science, software engineering, IT, computer networks, or a related technical field?
   - Yes
   - No

A.2 Instructions to save Ad Preferences (Facebook) or Ad Settings (Google) page

1. Click here to open [URL to Ad Preferences/Ad Settings] in a new window. Log in to Facebook/Google if you are prompted to do so.
2. Save that page as "HTML Only" (ctrl+S on Windows, Command+S on Mac, OR right click on the page –> Save As/Save Page As...).
3. Make sure you save the file somewhere it will be easy for you to find it. On the next screen you will submit the file you just downloaded to the survey, which will automatically generate questions for you based on the content of the file.
4. After you have saved the file, take a quick look at the information on the web page. Then, please close the browser tab or window you used to save the file and return here to complete the rest of this survey. If you are not using your own computer, don't forget to log out.
5. Have you ever seen your [Facebook Ad Preferences/Google Ad Settings] web page before?
   - Yes, I have seen it before.
   - No, I have not seen it before.
   - I'm not sure.

A.3 Questions about each Inference

After participants submitted their file, it was uploaded to our server where it was parsed to extract only the inference labels for "active" inferences. The inference labels were then sent back to the survey, where three questions were generated for each inference. We limited the total number of inferences asked about in the survey to 85, which were selected randomly from a respondent's inferences if they had more inferences than this. The three responses for each inference were then averaged to produce the reports used for interviews. See Appendix C for example reports.

1. *inference* is relevant to who I am as a person. [7-point Likert, Strongly agree–Strongly disagree]

2. *inference* is an accurate description of my everyday activities. [7-point Likert, Strongly agree–Strongly disagree]

3. It makes sense that *inference* is associated with me. [7-point Likert, Strongly agree–Strongly disagree]

A.4 Attention Check Questions

The survey included up to three attention check questions. Each page of questions about the inferences asked the three questions about five inferences, for a total of 15 questions per page. The attention check questions were inserted among these questions, on the third, eighth, and fifteenth page of questions about the inferences, if the participant had enough inferences to reach that point in the survey. Responses that failed to pass the checks were excluded.

1. To help us monitor the quality of our data, please select "Somewhat agree" from the choices below. [7-point Likert, Strongly agree–Strongly disagree]

2. To help us monitor the quality of our data, please select "Strongly disagree" from the choices below. [7-point Likert, Strongly agree–Strongly disagree]

3. Which website did you save a web page from earlier in this survey? [Select one: Pinterest, Facebook, Reddit, Twitter, Google]

A.5 Questions about Social Media and Online Behaviors

1. When was the last time you viewed the [Facebook Ad Preferences/Google Ad Settings] web page, not including for this survey?
   - Today
   - Yesterday
   - Within the past week
   - Within the past month
   - More than a month ago
   - I have never viewed the Facebook Ad Preferences web page before

2. Do you use an ad blocker when you browse the web?
   - Yes
   - No
   - I Don't Know

3. Have you ever had one of the following experiences? Select all that apply:
   - Fell victim to a phishing email message or other scam email
   - Received a notification from a company that your information was involved in a data breach
   - Had a virus on your computer or mobile device
   - Someone broke in or hacked your computer, mobile device, or account
   - Stranger used your credit card number without your knowledge or permission
   - Identity theft more extensive than use of your credit card number without permission
   - None of the above

4. How much negative information have you heard about [Facebook/Google] in the past several months?
   - A huge amount
   - A lot
   - A moderate amount
   - A little
   - None at all

5. Has how often you've used [Facebook/Google] increased, decreased or stayed about the same recently?
   - Increased a lot
   - Increased a little
   - Stayed about the same
   - Decreased a little
   - Decreased a lot

A.6 Privacy Questions

As part of the survey, we asked questions from an instrument developed by Smith, Milberg and Burke [27] to identify and measure concern about privacy practices. These questions were not analyzed for this paper. These questions were asked after the first question in the previous section.

Here are some statements about personal information. From the standpoint of personal privacy, please indicate the extent to which you, as an individual, agree or disagree with each statement. [7-point Likert, Strongly Agree–Strongly Disagree]

- It usually bothers me when companies ask me for personal information.
- When companies ask me for personal information, I sometimes think twice before providing it.
- It bothers me to give personal information to so many companies.
- I'm concerned that companies are collecting too much personal information about me.
- Companies should not use personal information for any purpose unless it has been authorized by the individuals who provided the information.
- When people give personal information to a company for some reason, the company should never use the information for any other reason.

- Companies should never sell the personal information they have about people to other companies.
- Companies should never share personal information with other companies unless it has been authorized by the individuals who provided the information.
- All the personal information companies have about people should be double-checked for accuracy, no matter how much this costs.
- Companies should take more steps to make sure that the personal information they have about people is accurate.
- Companies should have better procedures to correct errors in personal information.
- Companies should devote more time and effort to verifying the accuracy of the personal information they have about people.

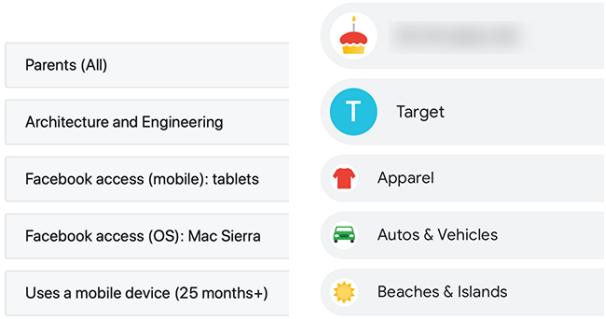
B Descriptives about the Inferences

This appendix provides descriptive information about the Facebook and Google inferences assigned to study participants. We present figures to show how these inferences normally appear on the Facebook Ad Preferences and Google Ad Settings pages before they were parsed, summary statistics for inferences across platforms, and both the most common and most uncommon inferences in our dataset.

B.1 Inferences Web Pages

The figure below shows a screen capture from the Facebook "Ad Preferences" and Google "Ad Settings" of one of the authors. In June 2020, these pages could be found at:

- Google: https://adssettings.google.com/
- Facebook: https://www.facebook.com/ads/preferences/



(a) Facebook inferences (b) Google inferences

B.2 Average Inferences Per Survey Respondent

Summary statistics for inferences across Facebook and Google:

Service	Mean	SD	Min	Median	Max
Facebook	11.97727	9.86063	3	4	33
Google	67.60784	35.08451	1	66	131

B.3 Common and Uncommon Facebook Inferences

There were 110 unique Facebook inferences in the survey data. 41 were assigned to only one respondent. Here are the most common inferences assigned to Facebook survey respondents (n=44), ranked from 1-12 by number of participants.

Facebook Inference	*N*
Mobile network or device users	42
Potential mobile network or device change	42
Recent mobile network or device change	42
WiFi users	42
Facebook access (mobile): all mobile devices	20
Facebook access (mobile): smartphones and tablets	19
Facebook access (network type): 4G	14
Facebook access (network type): WiFi	13
Gmail users	12
Engaged Shoppers	11
Facebook access (mobile): Apple (iOS)devices	10
Likely engagement with US political content (liberal)	10

The table below shows 12 inferences randomly selected from the 41 Facebook inferences that were only assigned to one person in the survey data.

Facebook Inferences	*N*
Owns: Galaxy Tab 4	1
Arts, Entertainment, Sports and Media	1
Management	1
Newly engaged (1 year)	1
Owns: iPhone 7 Plus	1
Owns: ZTE	1
Anniversary within 30 days	1
Lived in UK (Formerly Expats-UK)	1
Business and Finance	1
Lived in United States (Formerly Expats-United States)	1
Business page admins	1
Returned from travels 1 week ago	1

B.4 Common and Uncommon Google Inferences

There were 561 unique Google inferences in the survey data. 294 were assigned to only one respondent. Here are the most common inferences assigned to Google survey respondents (n=51), ranked from 1-12 by number of participants.

Google Inference	*N*
Parenting	46
Home & Garden	43
Shopping	43
Business Services	41
Holidays & Seasonal Events	41
Performing Arts	39
Books & Literature	38
Family & Relationships	38
Visual Art & Design	38
Computers & Electronics	37
Cooking & Recipes	35
American Football	33

The table below shows 12 inferences randomly selected from the 294 Google inferences that were only assigned to one person in the survey data.

Google Inferences	*N*
Tile Games	1
Cincinnati	1
Food	1
Historical Sites & Buildings	1
Interior Design	1
GetYourGuide	1
Oceania	1
Perfumes & Fragrances	1
45+ years old	1
OfficeSupply.com	1
Morgans Hotel Group	1
Refrigerators & Freezers	1

B.5 Inference Ratings

The next table shows the mean and standard deviation of the ratings for each of the three questions (sensible, relevant, and accurate) for both platforms.

Platform	*Mean*	*SD*
Sensible:		
Facebook Inferences	5.00	1.78
Google Inferences	4.29	2.08
Relevant:		
Facebook Inferences	4.68	1.87
Google Inferences	4.16	2.03
Accurate:		
Facebook Inferences	4.64	1.88
Google Inferences	3.84	2.05

The table below shows Pearson correlations between the inference ratings. The sensible, relevant, and accurate ratings of the inferences were highly correlated with each other. All correlations are statistically significant ($p < .001$).

	Facebook	*Google*
sensible, relevant	0.84	0.93
relevant, accurate	0.85	0.89
sensible, accurate	0.84	0.88

C Interview Protocol and Example Reports

Participants were interviewed about their reactions to the inferences they had been assigned by Facebook or Google using the semi-structured interview protocol on the next page. Each participant was given a report that presented the inferences they had answered questions about in the survey. Responses to the three questions about each inference were averaged, rounded to the nearest whole number, and then color-coded according to the 7-point Likert agreement scale used for the responses to the questions. Inferences in the Facebook reports were displayed in one grouping. Inferences in the Google reports were grouped into higher-level categories that were present in the labels of the inferences parsed from respondents' Ad Settings pages. Within each group, inferences were displayed to participants in order from Strongly Agree-Strongly Disagree. The reports included in this appendix were generated from the inferences and survey data of one of the authors.

Interview Guide and Protocol

General Background and Instructions

The goal of the interviews:

- The goal of the interviews is to collect data about people's beliefs about how the categories about them were assigned, and about their reactions to the categories, IN THEIR OWN WORDS.
- We want to know what they think and believe about the following kinds of things:
    - What the information is on the ad preferences page
    - How the information was chosen
    - How it feels to have this information associated with them
    - What the information IS being used for
    - What the information COULD be used for

About priming or leading the participant:

- It is important that the interviewer does not convey anything to the participant about what they know or believe about how the categories may be determined, or even what kinds of words to use to refer to what the platforms refer to as "categories".
- In other words, don't even call them "categories"; this may prime or bias the participants.
- This means that the interviewer must pay careful attention to the language each participant uses during the interview, and refer to the same concepts the participant talks about using the same kinds of words as the participant.
- If we indicate to them the kinds of vocabulary they should use, by using it ourselves, or if we ask leading questions, the data we collect will be about participants reactions to OUR beliefs, not about their own beliefs.
- The more we guide their responses, the more we will be collecting data about something they're only thinking about because we asked them to think about it. We want to know what they think about this, not what they think about what WE think about this.

Good ways to probe/follow up for more information:

- "Tell me more about [X]"
- "What do you mean by [X]"
- "How do you think [X] happens"
- "Where do you think [X] came from"
- "Can you give me a specific example of [X]"
- "What would [X] look like"

Things to listen for and follow up on in their answers:

The following are all common things that come up in interviews about privacy-related topics. These things are not useful data for this study if people just mention them and then quickly move on. But, if the interviewer can ask follow up questions and probe for their beliefs and thinking related to these things, the data will be more interesting and useful for identifying patterns and differences in people's beliefs.

So, for example, it isn't very useful to learn that people think having inferences assigned to them is creepy, or that they're concerned about it, without knowing more about why. Likewise, it isn't very useful to us to know that people think a certain kind of information is more sensitive than some other kind of information. We already know things like this from previous studies other people have done, and it doesn't help us figure out how people think systems are able to infer things about them or why certain inferences have been assigned to them.

But, what IS useful is learning about what makes it feel creepy, or what makes the information sensitive for them, and connecting that to their beliefs about where the inference categories come from and how they are assigned. So it is especially important to ask follow-up questions about topics like these:

- assumptions about why Facebook or Google assign categories, or how they will be used
- saying something is creepy or invasive

Interview Guide and Protocol

- commenting on what/who Facebook or Google thinks they are, as a person
- emotional reactions, like anger, irritation, amusement, pride, etc.
- talking about accuracy or mistakes related to the categories
- feeling surprised about any of the categories
- talking about what's missing from the list of categories

Tips for effective interviewing:

- *It is OK to wait for people to answer when you've asked them a question.*
  People may need to think for a minute before answering some of these questions, and if there is a pause in conversation while they do that it may feel a bit awkward. This is OK. The best way to give a participant space to answer is to remind yourself to PAUSE and let them think, even if the silence makes you uncomfortable. If you move on, and ask another question, they won't answer the first question! Count to 10 in your head if you have to.
- *Never interrupt the person you are interviewing!*
  This may mean that it feels like the person may be rambling into something that is off topic for the interview protocol. But, that doesn't mean the data won't be useful, and if you cut them off you will never know what they were going to say. People think out loud sometimes, and the process of talking about something is important for the process of thinking about it. Also, interrupting someone conveys to them that you weren't actually that interested in what they were saying, and which is absolutely the LAST thing we want interview participants to feel. We are VERY interested in what they have to say!
- *If you finish the interview feeling like " that was a nice conversation" you're not listening carefully enough.*
  Interviews should be tiring, because you should be trying to pay attention to **everything** the person is saying and thinking about how to follow up. This takes a lot of energy and focus. You shouldn't be thinking about other stuff going on in your life during the interview -- focus all of your attention on the participant, and asking good follow-up questions.
- *We really want the participants to actually think about the inferences during the interview.*
  Sometimes people's first response might be "I don't know" or "I have no idea" or "I've never thought about that before". This is because we're showing them information they really may not have thought about much before! It is important to follow up when they say that, don't just let it go! Some ways to follow up and get them talking about what they're thinking are: "Tell me more about that." or "Why do you think you haven't thought about it?" or "What is it that makes it hard to answer this question?" or "We're really interested in anything you can tell us about your thoughts about this."

Interview Guide and Protocol

Interview Protocol

At the beginning of the interview:

1. Introduce yourself and thank the participant for coming.

2. Tell them a brief overview of the study and why they're there. Something like:
    - We're going to be asking you some questions today about yourself and about your perceptions of the information available about you in popular web platforms like Facebook and Google.
    - Before we get started, I have a consent form here that I need you to look over and sign. It describes the study at a high level, and lets you know about what you will do, what your rights are as a participant, and about the Amazon gift card you'll receive at the end of the study.
    - I want to emphasize that what you say to us during the interview will be kept confidential, and you can stop participating at any time, just let us know.

3. Give them the consent form and a pen, and wait for them to sign it.

4. After they've signed the consent form, check it to make sure they have given consent to the audio recording before turning on the recorders. Ask them if they have any questions before starting.

5. Turn on the audio recorders once they've consented to both the interview AND the recording. DON'T FORGET TO TURN ON THE RECORDERS! Yes, both recorders.

6. Let them know when you've turned on the audio recorders.

1. Warm-Up Questions:

1. Tell me a little bit about yourself…
    - Want to probe for things like
        - What they do for a living
        - What their educational background is
        - How much they use computers and/or mobile devices and the Internet
    - Goal of this question is to put the participant at ease (everybody always starts off a little nervous and unsure about what to expect) and learn background information that will help us to describe the participants in general.

2. You recently filled out a survey for us – that's why you're here! We're going to be talking about some information from that survey in a minute, but I wanted to ask you first about:
    - What do you remember off the top of your head about that [Facebook | Google] web page that the survey asked you to save?
    - Had you ever looked at that web page before you filled out the survey? When was the last time you went to that page? Tell me more about that time… Do you remember why you looked at it? What your impression at that time was? Was that the only time? What other times did you look at it? Etc.

2. Questions about the categories:

1. Give the report to the participant—but don't give them the page comparing their categories to other people's categories yet! Tell them:
    - Here's a report that we generated from the answers you filled out to the questions in the survey you did for us about the information about you on Facebook or Google. Take a few minutes to look this over, and choose two of the categories that we'll start off talking about. Pick whatever ones you like, we're interested in what stands out to you about this information. Let me know when you're done. (If they start talking during this, just go with the flow!)

2. Then ask them to talk about each of the things they picked—first one, then the other one – we want to make sure we end up with answers to each of these questions about each of the categories they choose to talk about:
    - **What** stood out to you about that? Why did you pick it to talk about?
    - **Why** do you think that information was in your list?
    - **How** do you think that information was chosen?

3. Then, ask them to mark or show you any other categories that stand out to them for any reason in the report. Ask them the **what**, **why** and **how** questions about each of these categories too.

4. Next, we want to know about whether any of the information violated their expectations in any way. They may have talked about some of these things (surprising, inaccurate, uncomfortable, missing, confusing) already, if so, you don't need to ask about them again. For each bullet point below that you end up talking with them about, ask the **why** and **how** questions.
    - Is there any information on here that is **surprising** to you? Why?
    - Is there anything you think might be on here by **mistake**? Why?
    - Is there any information here that makes you **uncomfortable**? Why?
    - Is there any information **missing**, that you would have expected to see but isn't on here? Why?
    - Is there any information on here that you find **confusing**? Why?

5. Finally, show them the page comparing their categories with other people's categories. Be really, really careful not to ask leading questions about these categories!
    - Ask about the first set – the categories that are unique to them. *Note that if this list is long and the participant seems like they are getting tired of talking about categories, just ask them which ones stand out and talk about only those.*
        - Ask the **why** and **how** questions about these categories, for the ones you haven't already talked about with them.
    - Compare the first set (unique to you, Section 1) to the second set (common for other people in our survey, but not associated with you, Section 2):
        - What's the first thing that stands out to you? Why?
    - Next, ask them to take a look at Section 3 under "How your information compares with others":
        - This section compares the information that was assigned to you, with the information Google | Facebook assigned to other people who completed the survey. It shows information that everyone had, versus information that you and only a few others had.
        - I'd like you to look at this information and think for a minute about how accurate it might be. How would you say these lists compare, from that perspective?

3. Big-picture questions:

1. What do you think about what the purpose of the information is? Why do you think so? Tell me more about that…
    - *If they have trouble answering the above question, try:* How do you think it is being used by Facebook or Google right now? What do you think it could be used for?
    - If the participant says something about third parties using the information, here are some follow-up questions: Can you give me a specific example of that? Describe for me how that works. What would [third party] be doing with that information?

2. How does the information we've been talking about today from [Facebook | Google] compare with the information you typically put in when you fill out an online profile for a website account? Can you give me an example of the kind of information in the profile of one of your online accounts? How is that similar/different?

3. If there were someone who works for [Facebook | Google] sitting here with us, what would you like to ask them about this?

Interview Guide and Protocol

4. Questions about Facebook or Google usage:

1. Facebook
    - When was the last time you used Facebook? Tell me about that.
    - Is that typical of what you use Facebook for? Can you tell me about another example of when you used Facebook like that?
    - What other ways do you use Facebook? Can you give me a specific example? Tell me about that.
    - How do you use the "Like" button on Facebook? Can you give me an example? Tell me about that.
    - What are your impressions of the advertising that you see on Facebook? Can you give me an example? Tell me about that.
    - What would you do if you wanted to change or remove some of this information about you that Facebook has? *(If they don't know, they may ask you how to do it.)*

2. Google
    - Google has a lot of different "products"; in your survey you mentioned that you used products X, Y, and Z. Which one do you use the most?
    - When was the last time you used X? Tell me about that.
    - Is that typical of what you use X for?
    - What's another Google product that you use often? When was the last time you used Y? Tell me about that.
    - What are your impressions of the advertising that you see in product X or product Y? Can you give me an example? Tell me about that.
    - What would you do if you wanted to change or remove some of this information about you that Google has? *(If they don't know, they may ask you how to do it.)*

5. Wrap Up

1. Ask if the participant has any questions:
    - If the participant has asked any questions, now is the time to answer them. If you don't know the answer, make a note about it and tell them [redacted] will get back to them. You should be able to answer simple questions like "what is this study about", "when do you expect to be done", "can you email me a copy of the report" (yes!), etc. You should also be prepared to assist if the participant asks you how to delete/remove/edit categories.

2. Ask the participant what email address they want to use for receiving the gift card:
    - Tell the participant that you've reached the end of the interview. Then tell them that they'll be sent a $25 Amazon.com gift card for participating, by [redacted]. Ask them what email address they want us to use to send them the gift card, and write it down!

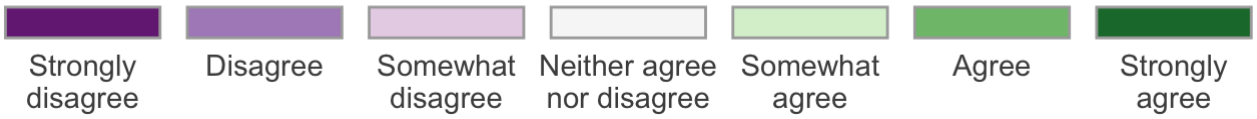
(Participant Name): Facebook Information

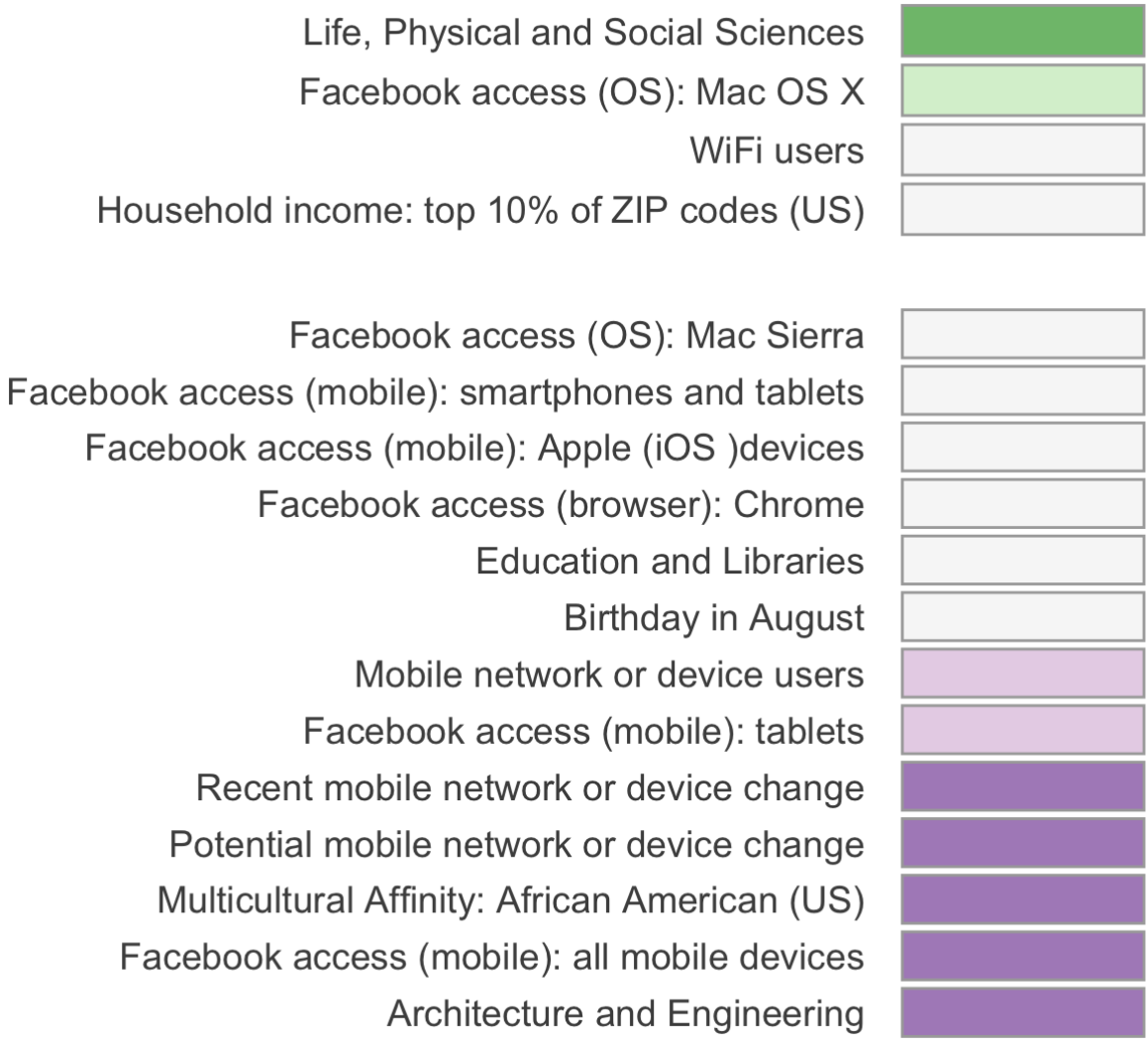
Below is the information you answered questions about when you filled out the survey. You answered three questions for each piece of information:

- [X] is relevant to who I am as a person
- [X] is an accurate description of my everyday activities
- It makes sense that [X] is associated with me

The colors in each table below represent the average of your responses to the three questions for each piece of information. The legend illustrates what the colors mean:

Legend





How your information compares with others

This page shows you how your Facebook information compares with the Facebook information about the other people who filled out our survey.

In total, 44 people from around the East Lansing area completed the survey and submitted their Facebook information.

Section 1

Here is the information that appeared in your file, **but nobody else's**:

Architecture and Engineering

Section 2

Here's the information that **was NOT in your file**, but was in in the Facebook files of **at least 13%** of the people who completed our survey:

Android: 360 degree media supported
Commuters
Engaged Shoppers
Facebook access (mobile): Android devices
Facebook access (network type): 4G
Facebook access (network type): WiFi
Facebook access: older devices and OS
Frequent Travelers
Gmail users
Household income: top 10%-25% of ZIP codes (US)
Likely engagement with US political content (liberal)
Parents (All)
US politics (moderate)
Uses a mobile device (25 months+)

Section 3

And finally, here's how similar your information was to the information in other people's Facebook files.

Everyone *had this information:*
Mobile network or device users, Potential mobile network or device change, Recent mobile network or device change, WiFi users

You and ***a few other people*** *had this information:*
Education and Libraries, Facebook access (OS): Mac OS X, Facebook access (OS): Mac Sierra, Life, Physical and Social Sciences

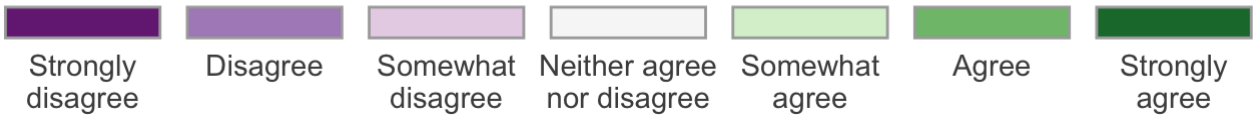
(Participant Name): Google Information

Below is the information you answered questions about when you filled out the survey. You answered three questions for each piece of information:

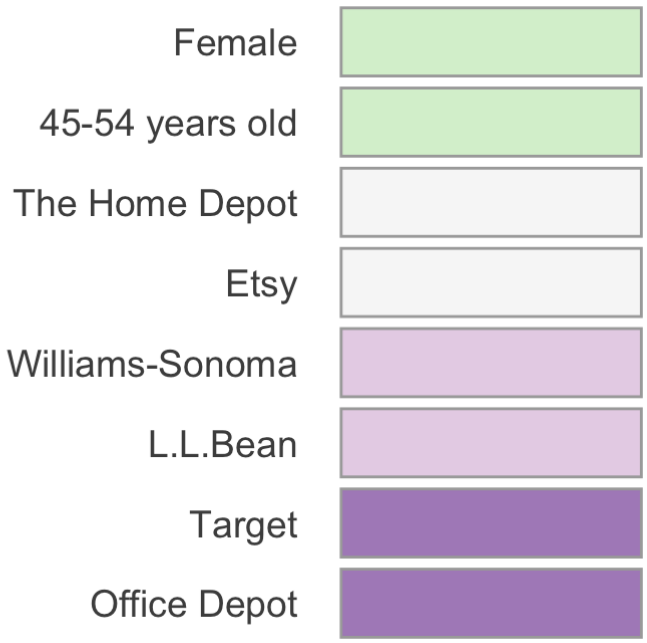
- [X] is relevant to who I am as a person
- [X] is an accurate description of my everyday activities
- It makes sense that [X] is associated with me

The colors in each table below represent the average of your responses to the three questions for each piece of information. The legend illustrates what the colors mean:

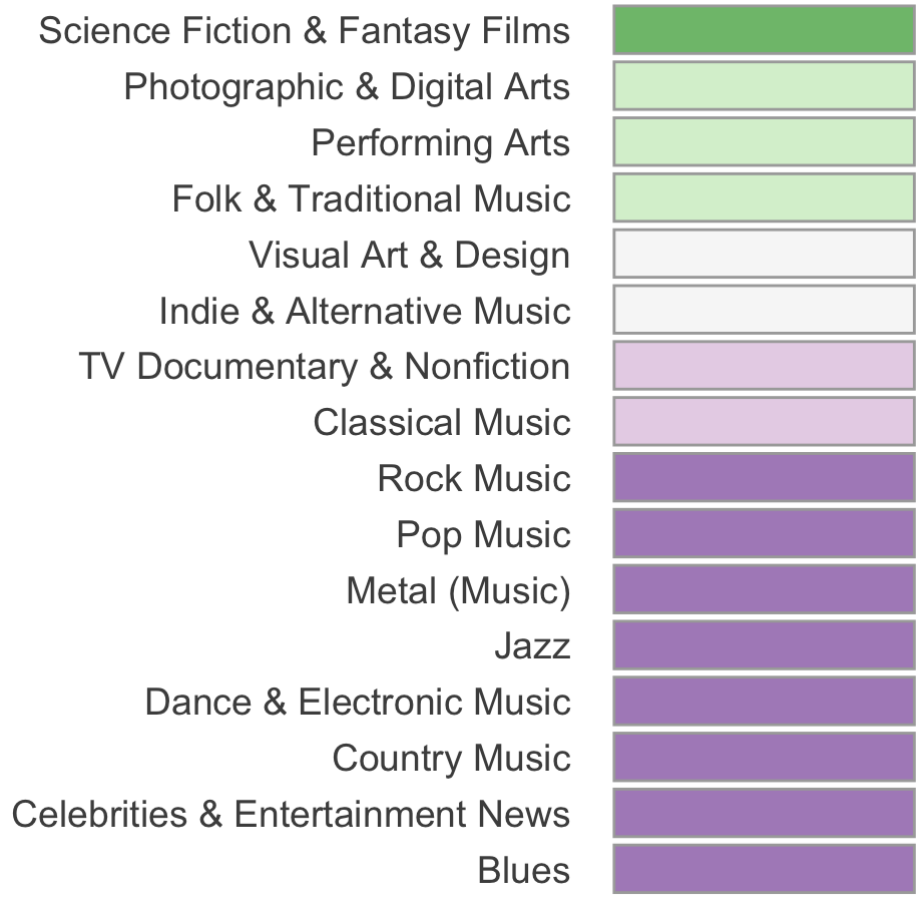
Legend



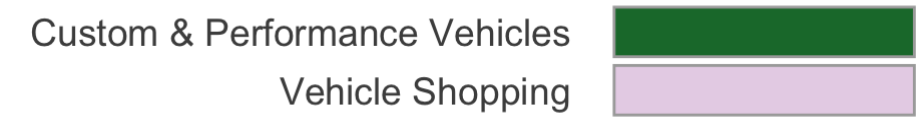
Google has organized some of the information about you into a hierarchy. Where possible, this report presents those pieces of information grouped under the heading that Google has assigned. The first group, below, has no heading:



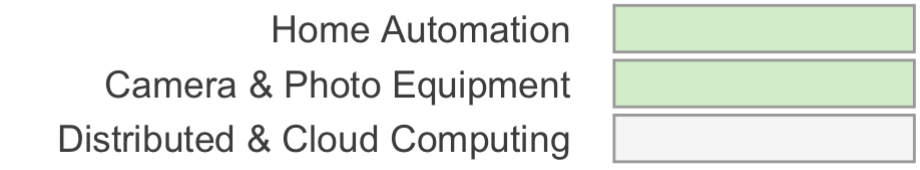
Arts & Entertainment



Autos & Vehicles



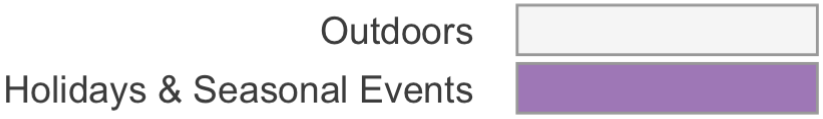
Computers & Electronics



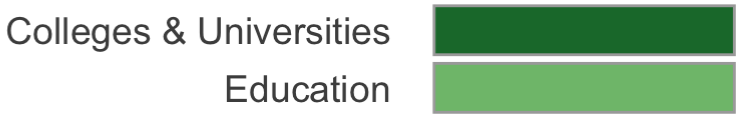
Food & Drink



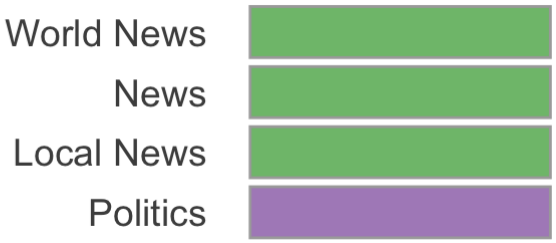
Hobbies & Leisure



Jobs & Education



News



Online Communities



People & Society

Family & Relationships 

Pets & Animals

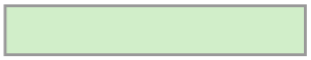
Dogs 

Shopping

Shopping 

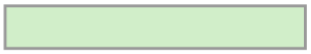
Custom & Personalized Items 

Sports

Baseball 

Australian Football 

Travel

Travel 

How your information compares with others

This page shows you how your Google information compares with the Google information about the other people who filled out our survey.

In total, 54 people from around the East Lansing area completed the survey and submitted their Google information.

Section 1

Here is the information that appeared in your file, **but nobody else's:**

Custom & Personalized Items
Williams-Sonoma

Section 2

Here's the information that **was NOT in your file**, but was in in the Google files of **at least 50%** of the people who completed our survey:

American Football
Autos & Vehicles
Books & Literature
Business Services
Comics & Animation
Computers & Electronics
Coupons & Discount Offers
Credit Cards
Gifts & Special Event Items
Home & Garden
Home Improvement
Parenting
Price Comparisons
TV Game Shows
TV Reality Shows
TV Talk Shows

Section 3

And finally, here's how similar your information was to the information in other people's Google files.

Everyone *had this information:*
Shopping, Holidays & Seasonal Events, Performing Arts, Family & Relationships, Visual Art & Design, Cooking & Recipes, Blues, Gourmet & Specialty Foods, Politics, Travel

You and ***a few other people*** *had this information:*
Photographic & Digital Arts, Camera & Photo Equipment, TV Documentary & Nonfiction, World News, Social Networks, News, Baseball, Colleges & Universities, Custom & Performance Vehicles, Australian Football

Replication: Why We Still Can't Browse in Peace: On the Uniqueness and Reidentifiability of Web Browsing Histories

Sarah Bird
Mozilla

Ilana Segall
Mozilla

Martin Lopatka
Mozilla

Abstract

We examine the threat to individuals' privacy based on the feasibility of reidentifying users through distinctive profiles of their browsing history visible to websites and third parties. This work replicates and extends the 2012 paper *Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns* [48]. The original work demonstrated that browsing profiles are highly distinctive and stable. We reproduce those results and extend the original work to detail the privacy risk posed by the aggregation of browsing histories. Our dataset consists of two weeks of browsing data from ~52,000 Firefox users. Our work replicates the original paper's core findings by identifying 48,919 distinct browsing profiles, of which 99% are unique. High uniqueness holds even when histories are truncated to just 100 top sites. We then find that for users who visited 50 or more distinct domains in the two-week data collection period, ~50% can be reidentified using the top 10k sites. Reidentifiability rose to over 80% for users that browsed 150 or more distinct domains. Finally, we observe numerous third parties pervasive enough to gather web histories sufficient to leverage browsing history as an identifier.

1 Introduction

Web tracking is the process by which parties with visibility into web traffic identify distinctive patterns of navigation in order to attribute browsing history to specific individuals. Third-party trackers remain a major concern; their prevalence and mass tracking activity is well documented [27, 53, 64].

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

This work seeks to reproduce the findings of Olejnik, Castelluccia, and Janc [48] regarding the leakage of private information when users browse the web. The reach of large-scale providers of analytics and advertisement services into the overall set of web properties shows a continued increase in visibility [64] by such parties across a plurality of web properties. This makes the threat of history-based profiling even more tangible and urgent now than when originally proposed.

2 Background and related work

As a replication of prior work, this manuscript presumes some familiarity with the source manuscript [48]. The following sections provide a summary of the original findings (Section 2.1), changes in the overall context, and relevant background for a comparison between the original research and our research (Section 2.2).

2.1 Original paper

Olejnik, Castelluccia, and Janc [48] gathered data in a project aimed at educating users about privacy practices. For the analysis presented in [48] they used the CSS :visited browser vulnerability [8] to determine whether various home pages were in a user's browsing history. That is, they probed users' browsers for 6,000 predefined "primary links" such as `www.google.com` and got a yes/no for whether that home page was in the user's browsing history. A user may have visited that home page and then cleared their browsing history, in which case they would not register a hit. Additionally a user may have visited a subpage e.g. `www.google.com/maps` but not `www.google.com` in which case the probe for `www.google.com` would also not register a hit. The project website was open for an extended period of time and recorded profiles between January 2009 and May 2011 for 441,627 unique users, some of whom returned for multiple history tests, allowing the researchers to study the evolution of browser profiles as well.

With this data, they examined the uniqueness of browsing histories. Each history profile was assembled as a vector, where each index was a boolean indicating the presence of a historical visit to each of the 6,000 domains they probed for. This vector was sorted by the popularity they observed in their dataset. Research questions addressed in their work included (a) How many profiles both unique and non-unique were observed for profiles of different history sizes? (b) Did profile distinctiveness vary with profile history size? and (c) How did the size of the unique domain vector impact the uniqueness metrics?

Notably, they found that with just the 50 most popular sites they were able to get a very similar distribution of distinctive histories compared to complete knowledge of the full 6,000 site list. Their analysis began with 382,269 users who completed their popular site test; 94% of these had unique browsing histories. There were 223,197 profiles with histories of length 4 or more, of which 98% were unique. The results were generalized to an information theoretical representation such that distinctiveness of profiles could be quantified. This analysis allowed the results to be recomputed for a compressed version of history profiles where only a set of 72 interest categories were considered. This resulted in 164,043 distinct profiles of which 88% were attributed to a unique user.

The stability of history profiles was also measured in order examine the possibility of history profiles being a tracking vector. A subset of users revisited their site and they analyzed the change in their history. They found profiles were stable over time although there were limitations to the analysis due to their history detection mechanism.

Finally, they explored the threat of large parties with extensive third-party reach (Google and Facebook) and built unique domain vectors based exclusively on sites present in users' histories that also contained content from Google or Facebook. They found 50% of Google-visible profiles, and 25% of Facebook-visible profiles, were unique.

2.2 Modern context

In early 2020, we are in the midst of an upheaval of the tracking ecosystem, as regulatory oversight and public discourse appear to have reached a tipping point. Work published around the time of the original manuscript provides insight into the state of the tracking ecosystem of that era [1,7,29,55]. More recent work depicts increasingly sophisticated tracking technologies fueling the targeted behavioural advertisement industry [31, 50, 51]. We also see continued increases in scale [56], a profound lack of transparency in disclosure of personal information flows [37], and consolidation of the internet economy to fewer, larger, dominant parties [6]. The concept of a singular web is rendered increasingly obsolete as more and more content is dynamically generated, personalized per visitor, and generated by web visitors themselves.

Meanwhile, concerns from the time of the original manuscript persist. Third-party trackers remain a major presence, with the prevalence of mass tracking activity now better understood [17, 27, 30, 33, 34, 53, 64]. And, while the specific technical exploit used to gather browser histories in the original manuscript no longer exists, in 2018 Smith et al. [58] documented four new history sniffing techniques.

In the modern context, increasing the usability of fingerprinting and transient identifiers is at the forefront of the technical web tracking discussion. Mishra et al. demonstrated that IP addresses can be static for a month at a time [42] which, as we will show, is more than enough time to build reidentifiable browsing profiles. The effectiveness of fingerprinting to *uniquely* identify people has been debated since Gómez-Boix et al. in 2018 [27] estimated that only 36% of desktop users were *uniquely* identifiable compared to Eckersley's 2010 estimate of 88% [22]. However, Gómez-Boix et al.'s paper [27] also showed that 95% of desktop users were in a fingerprinting pool of just 100 users or less. We will show this also has a significant impact on reidentifiability.

Another fundamental change in the web ecosystem has to do with the drive towards cross-device identification [10, 13]. In this context, we see evidence of the original paper's concerns that browser histories may be used as an identifier coming to light. While specifics are generally proprietary, marketing and advertising platforms advertise their ability to build consumer profiles [2–4]. In 2015, Drawbridge, noted for their use of probabilistic tracking [10], launched a public competition for cross-device identification [32]. The dataset included user's website data [20] and the winner of the competition was DataLab, a marketing analytics company [54].

3 Methodology

We designed a methodology that would allow us to not only replicate the original findings, but also extend the analysis towards specific privacy threats raised by the original authors' work. Data was collected from ~52,000 Firefox browser users who elected to share data for research and product development purposes beyond what is outlined in Mozilla's default data collection policies [44]. An opt-in prompt was shown to candidate participants in accordance with Mozilla's policies governing data stewardship and consent management for user research [47]. This prompt provided users with a clear, comprehensible, English-language explanation of the extended data collection [45].

Measurement of browsing data was carried out using a custom browser extension [46] derived from the OpenWPM instrumentation [23]. Data was encrypted on the client and collected via secure infrastructure separated from Mozilla's normal telemetry pipeline enabling highly restricted data access. Data was collected in "pings" transmitted regularly from the browser as data amassed. Each browser was given a unique identifier to enable pings to be joined together to

assemble the dataset. This unique identifier was specific to the opt-in data collection program used by this study and not connected or join-able to other identifiers used by Mozilla.

For each user, we wished to amass two distinct browsing periods of data. Practical considerations pertaining to the total volume of data collected, operational costs, and a desire to measure profile stability motivated our implementation which collected data for 7 calendar days, paused for 7 days, and subsequently resumed for an additional 7 days. We added an additional day to each week-long observation period to ensure collection of delayed pings.

Figure 1 shows the number of unique users who were active each day of the experiment. The final dataset ultimately

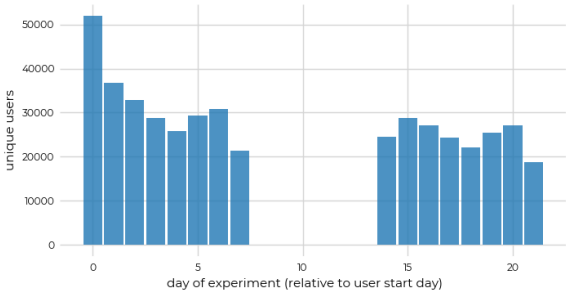


Figure 1: Number of unique active users per day over experiment period (including intentional one-week gap.)

contained browsing data from ~35 million site visits and ~660,000 distinct domains, gathered between July 16 and August 13, 2019. Figure 2 shows the distribution of the number of different domains per user aggregated per collection day. We typically see a median of 8 different domains per user per day; aggregated over the entire collection period results in a median of 34 different domains per user.

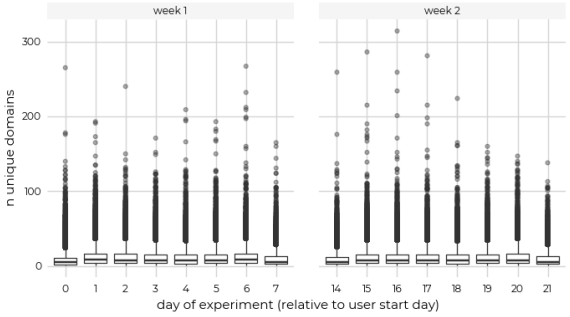


Figure 2: Number of unique domains visited per user per day.

Restricting our domain counts to only a top-site list (details in section 3.1), the median number of domains per client was 18. This is comparable to a median of 10 seen in the original paper, whose data collection methodology required use of a top-site list. We also note that the maximum number of domains per user is 1,116. Although this is a lot, it is not an unrealistic amount for 14 days of browsing and is one indicator that our users were real users rather than bots or other automated tools using Firefox.

Min size	Max size	N users
1	25	21,519
26	50	11,195
51	75	6,750
76	100	4,499
101	125	2,791
126	150	1,766
151	-	3,457
	Total	51,977

Table 1: Number of users by number of unique domain visits

We computed the distribution of the size of profiles across users. Table 1 gives a top-level view of the relative sizes of profiles when considering all unique domains visited. We see a similar exponential trend as in the original work: the largest number of users have a small history size.

We do not present the data for the number of site visits per day. We found some users were strong outliers with hundreds of thousands of page load events. We examined these manually and found that these users had one or two sites that were being hit a very large numbers of times with otherwise normal browser histories. We speculate that these users had browser addon(s) installed for automation of specific workflows, but were otherwise representative of normal, human browsing. As we represent browsing history as a boolean vector of distinct domains, thereby removing all frequency measures, we did not study this further and did not manually remove any users from the dataset.

Our data collection mechanism allowed us to capture all network requests and responses associated with site navigation events. Starting with complete information about the user's browsing history allows us to examine the uniqueness of browsing profiles in full depth, and make strong claims about the stability of profiles and model specific reidentification scenarios. Capturing all the network requests allows us to assess the potential tracking capabilities for wide range of third parties.

Aside from the methodological differences outlined above, we note several sources of potential variation between the specific study populations. Both studies use an opt-in population, though with different methodology and flows. Our cohort provided a priori consent to longitudinal data collection, meaning a low attrition rate compared to the original work which relied on return visitors to the study website. The original study was able to study traffic from different web browsers' users. They did not report on differences between browsers. However, they do mention data from 1,256 users reporting a mobile user agent. Our WebExtension shipped to Firefox versions 67

or 68, which were release-channel versions during the study period, and ensured that the participants' browser was configured to an en-US localization, as the opt-in consent text was only available in English.

3.1 Selection of a history vector

The original study by Olejnik et al. required a pre-selected list of sites used to probe users' histories to see if a user had visited them. They sorted the list of sites by observed popularity in their dataset. The globally ordered vector, with boolean entries for each user was then the main input to their analyses.

As we have complete information of browsing history during the study period, we can create a history vector of the observed domains, ranked by popularity. In our results, we label this as *all observed domains*. However, we also created a pre-defined site list for analysis. We do this for the following reasons: (a) using a pre-defined list allows us to more closely replicate the original paper's methodology; (b) to build a category vector, we use a third-party service (see section 3.2) and we do not want to leak user data to that service by basing our queries on user data; and (c) since history-sniffing attacks, which require a pre-defined vector, are still possible [58], it is relevant to perform analyses with one.

Olejnik et al's list of 6,000 popular sites was "created out of 500 most popular links from Alexa, 4,000 from the Quantcast popular websites list, lists of common government and military websites, and several custom-chosen URLs selected for their demonstration and education potential."

Pre-experiment analysis performed by Zeber [63] examined different top site lists, and led to our site list which is a hybrid of the Alexa [5] and Tranco [36] top 10K site lists. We call it the *Trexa* list and it is made by interleaving the sites from each and dropping duplicate entries [59]. The top 100 Trexa sites overlap with the 100 observed top sites in our user data by 40%. If we expand to look at the top 10,000 sites from each list, the overlap rises to 52%.

3.2 Generation of a category vector

To generate our category vector we used the WebShrinker API [60] to obtain a categorization for domains. The entire Trexa list was run through the WebShrinker API with the Interactive Advertising Bureau categorization returned. The IAB taxonomy has a series of top-level and sub-level categories, with corresponding scores and confidence levels for each domain-category pair. We mapped each domain to its highest-rated specific subcategory, unless no such confident category existed. In that case, we substituted the most relevant higher-level category. If the domain was listed as "Uncategorized," we removed it from the dataset. Finally, we sorted from most to least observed category as we did with the all observed domains vector. Ultimately, our categorical dataset contained 281 categories out of a total of a possible 404 offered in the IAB category standard. The original paper used a similar categorization service called Trend Micro, which yielded 72 interest categories. Although Trend Micro is still around today, we chose WebShrinker because the IAB categorization is in line with the threat of adtech we are interested in exploring.

3.3 Terminology

Throughout the paper, we use the word *user* for convenience and flow, but what we are actually examining is sets of browser histories. The data we collected only guarantees a unique identifier for the browser. There may be multiple browsers per user and there may be multiple users per browser.

We refer to collections of unique user domain (or category) visits in a given time period as a *profile*, as in Olejnik et al. The profiles' underlying *dataset* may be the list of all observed domain visits, Trexa domain visits, or categories. The *size* of a profile is thus the number of unique domains or categories the profile contains.

As in the original work, and described in Section 2.1, we also use the concept of a boolean *profile vector*. We refer to an index of this vector as *rank*; thus, a *subvector-to-rank-k* describes the vector of length k with boolean values representing whether a user has visited each of the top k items in the dataset. Recall that the items in the vector are sorted from the most popular item in the underlying dataset to the least.

4 Replication Results

4.1 Web history profile uniqueness

In Figure 3, we compare the size of profiles when we observe all domains (as presented in Table 1), when we use the Trexa list to specify the set of domains considered, and when we use the category representation of profiles. As expected, the

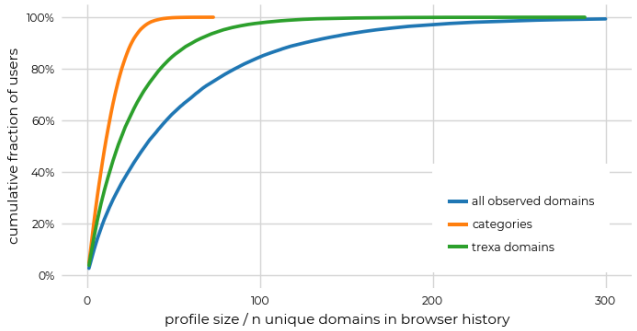


Figure 3: Cumulative proportion of users per profile size.

distribution shifts left as profiles shrink in size, relatively uniformly across the entire population. Shifting from all domains to the Trexa list, the median profile size is reduced to

18 domains. Restricting to just categories, the median profile size shifts further still, to 11. We note that while the maximum profile size for all observed domains (1116) is much higher than its counterpart for Trexa (288) and categories (73), we observe over 99% of all users in this plot.

In Figure 4, we examine the distribution of prevalence of history profiles constructed from the set of all observed domains, Trexa domains, and page categories. The leftmost point on the x-axis represents the most common profile for each underlying dataset. The y-axis represents the number of users with that profile. In the Trexa dataset, for example, the most popular profile consists of only visits to the most popular Trexa domain, `www.google.com`, a profile which is shared by 559 users, when considering all domains in the Trexa list (brown line) and nearly 10,000 users when the domain set to build profiles from is restricted to just the top 10 sites (blue line). When again considering all domains in the Trexa list, the second-most popular profile consists of only visits to `www.youtube.com`, with 150 users. Generally, the smaller the pool of users with the same profile, the more easily a specific user can be identified. When the number of users with a particular profile reaches 1, we call that profile *unique*, as it is only observed for a single user within this data collection. We

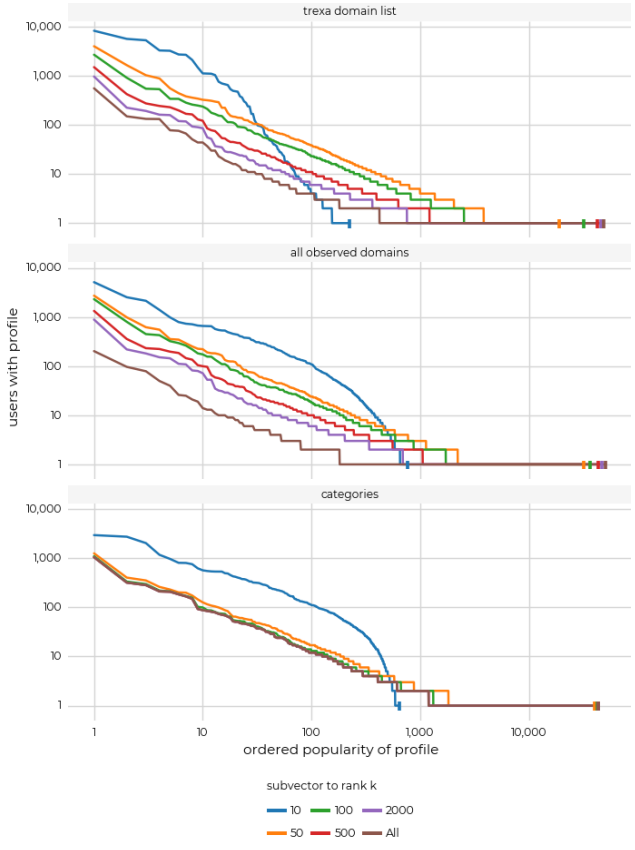


Figure 4: Frequency distributions of distinct profiles ranked by popularity, across multiple subvectors of top domains.

repeated the analysis with several subvectors of the domain history, sampling up to the rank k entries for various values of k. In all three datasets, $k = 50$ is enough to align the slope of the line with the one generated from the entire population. Additionally, the length of the segment at $y = 1$ for each line on the chart indicates the number of profiles deemed unique. This finding is consistent with results presented in Olejnik et al. despite the differences in the underlying datasets.

We now estimate how much identifiability is lost by only looking at a portion of a predefined domain list. When all domains were observed, we saw 51,035 different profiles, corresponding to a 99.65% rate of uniqueness among our users. Restricting to visibility only the top 100 most frequently observed domains allowed us to compute 36,652 profiles (based on available histories) of which 95.31% were unique. Substitution of the observed domain popularity with the Trexa list led to the aggregation of 48,919 profiles of which 99.14% were unique. Further constraining visibility to only the top 100 Trexa domains led to 31,810 profiles of which 92.05% were unique. When using the compressed data representation of just 281 categories, we still observed 43,348 profiles, of which 97.24% were unique.

In Figure 5 we examine the change in proportion of unique profiles with respect to the size of the subvector more closely, looking at all values of k spanning the range 1 to 250. We

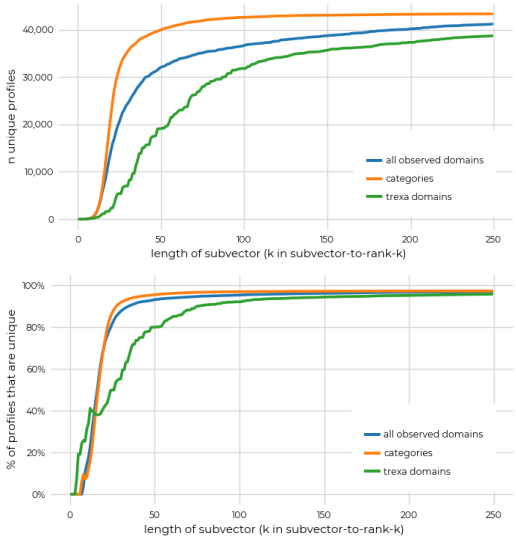


Figure 5: Variation in number and proportion of unique profiles as history vector grows in size.

found, unsurprisingly, that across datasets there are no unique profiles described by a profile subvector of length 1 (in the Trexa dataset, for example, there are 2 such possible profiles– users that did or did not go to `www.google.com`). *All observed domains* and *categories* were ordered by the user data, describing the behavior of our users most precisely, and a large amount of these profiles quickly exhibit uniqueness.

Conversely, the ordering of the predefined Trexa domain vector does not perfectly match the population browsing data, accounting for the lack of smoothness as k increases. Although the category vector has less precision than a history vector, a category covers numerous domains; thus the category vector allows for more browsing behavior to be represented for a given k. We observe that the number of distinct category profiles at the chart's elbow (k~30) is larger than the amount of distinct Trexa profiles by roughly a factor of 3.

Though it's straightforward to classify two profiles as distinct or not, we would like to be able to conceptualize the extent of the distance between them. Olejnik et al. used the Jaccard index to measure similarity. In Figure 6, we use the related Jaccard distance, measured as (1 - Jaccard index), to examine the range of this distance. To generate the distribu-

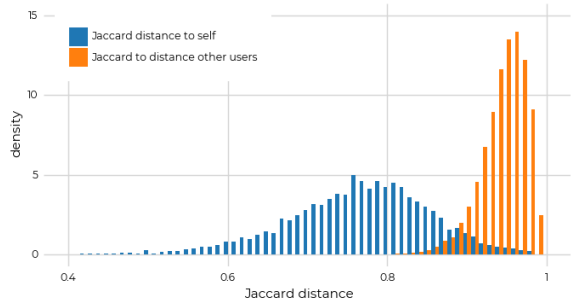


Figure 6: Distribution of Jaccard distances between profiles of a single user and distinct users

tions in Figure 6, we selected users total profile sizes of at least 50, and then split the profile into two observed activity periods. We then compute the Jaccard distance for the user to theirself between period 1 and period 2, and the user in period 1 to all other users in period 2. We plot a sample of this distribution for tractability. Though there is overlap in the distributions, we can reasonably expect that, for a given profile, a profile from the same user has a smaller distance than one from a different user a good portion of the time. Repeated measures from the distribution (if we had several more time periods, for example) would only increase confidence in which distribution the measurements belongs to.

We believe that our results strengthen the original paper's claim that subsets of user data pose a potential for exploitation in user identifiability, despite differences in the underlying data. The original manuscript observed a (likely) longer duration of history for around 7 times as many users, while we were able to collect more detailed information. In the context of this Section, it's important to note that our smaller population increases the chances of observing a unique profile due to fewer overall profiles.

The analysis performed above replicates the distribution of profile sizes from Olejnik et al. Section 4.1.2, the frequency of distributions of different subvectors from Sections 4.1.3 and 4.2.3, and the distribution of Jaccard similarity from 4.1.3. In particular, we see similar trends and measurements of uniqueness, as Olejnik et al. note in Section 4.1.2, 4.1.3, and 4.2.3. Our similar findings regarding the uniqueness of category profiles is particularly relevant in light of the proposal for category-based targeted advertisement as a privacy safeguard put forth by some advertisement companies [18, 26]. In this section we did not directly address the original paper's findings related to surprisal, as in section 5.4 we use a different approach to quantifying information gain from a more detailed profile.

4.2 Stability of profiles

The original paper examined stability of history profiles to understand the potential for browser histories to be used for tracking. The combination of uniqueness and stability being preconditions for reidentifiability is also discussed by Gómez-Boix et al [27]. However, the data collection method employed by the original work hindered a detailed examination of profile stability as it relied on organic return visits to the study page. Although over 368,000 browsing histories were collected, only a small fraction of users could be included in the stability analysis. They report data for ~1,800 returning users on day 1, dropping to ~400 returning users by day 7, and ~150 by day 14. Aside from sample size considerations, two additional challenges impact the interpretation of those findings. Firstly, the history detection technique employed could not detect whether sites collected on first visit were revisited or first-time visits. Secondly, ground truth was established based on reidentifying visitors with a combination of IP Address and UserAgent, perhaps biasing the baseline data to under-represent users accessing the web from multiple locations. In particular, accurate estimation of site revisitation rates is vital to estimating the possibility of reidentifiability.

Our methodology gathered two weeks of browsing data from all our users. Although we do see a drop in the number of users over the course of the study as visible in Figure 1, we have two weeks of browsing data from tens of thousands of users allowing us to model the reidentification of users based on browsing history. Due to the fundamental differences between the datasets, we have not attempted to replicate the original paper's stability analysis. Instead, we extend the original work and model reidentifiability as outlined in Section 5. We note here that our work supports the original finding that browser history profiles are stable.

4.3 Third parties

As our data collection included all requests and responses, we are able to see all third parties that users were exposed to. We find the results from the original paper are not only reproducible, but are stronger today with Google (Alphabet) and Facebook observing large portions of the web. Presenta-

tion of our third-party analyses is available in the extension section 5.6 where we examine the theoretical reidentifiability of the top third parties our participants were exposed to.

5 Reidentification rates extension

As previewed in Sections 4.2 and 4.3, we wish to expand on the work illuminating the privacy risk that Olejnik et al. [48] describe. We do so by directly modeling the reidentifiability of users based on their browsing history. As we move into this section, recall that while we continue to talk about profiles, as described in Section 3.3, we now consider each user separately. If two different users have the same history profile, they appear as two identical rows in our dataset and the matching process, as explained in 5.1, will randomly select between those two users. For all the analyses presented, period 1 and period 2 are the two weeks of browsing data separated by a week.

5.1 Reidentifiability metric

As motivated by the distributions in Figure 6, we define a model for attempting to reidentify a user between observation sessions. For each user, we compute the Jaccard distance between the user's period 1 profile vector and all profile vectors in week 2. The period 2 browsing history vector with the lowest Jaccard distance to the profile of interest is considered the most likely match. (Note: We did not build special handling for ties. We used the pandas `idxmin` function which returns the "index of first occurrence of minimum over requested axis". As the identifier is a randomly generated alpha numeric code, we do not believe this introduced bias into our results.) We then evaluate the match and a user is considered *reidentified* if the correct identification was made. The reidentifiability rate is thus the percentage of all users in a specified pool who are reidentified in this manner.

Though the minimum Jaccard distance is a very simplistic metric to use for this analysis, we chose it for a number of reasons. Firstly, it directly follows from the examination of Jaccard distances between profiles in the original paper. Secondly, it is easy to interpret: this metric picks the period 2 vector with the most overlap between both sets. Lastly, and perhaps most importantly, building a reidentifiability engine is a complex task with several domain-specific algorithmic choices to consider. Subtler distance metrics, multinomial data (modeling the frequency of domain visits), and more detailed browsing metadata (time of browsing, for example, or more privacy-invasive features that could be collected in the wild) could all improve our matching algorithm significantly. However, our goal is to take the next logical step beyond the analysis of profile uniqueness to understand whether that uniqueness implies a level of reidentifiability worth exploring. Because of possible model improvements previously mentioned, we can safely consider the results we obtain to be an underestimate of analagous methods potentially used in industry.

We note that throughout the following analyses we split users into groups by their overall profile size (the number of distinct domains in their total browsing history). It is done once to sub-divide the population. It is not recomputed for each type of analysis and serves solely as a way to stratify the population so that inter-group comparisons are valid.

5.2 Baseline reidentifiability

We start our analysis by looking at the reidentifiability of all users with more than 50 distinct domains in their complete browsing history. We start with this number as a trade-off between two factors: (a) as we will show later in Section 5.4, reidentifiability increases as the number of unique domains in a user's profile increases, and (b) as shown in Figure 3, the smaller the minimum profile size we consider, the larger the pool of users we have. Restricting to a minimum profile size of 50 domains results in 19,263 users, 37% of the total in our dataset. This number is tractable for computation and yields significant reidentifiability rates.

With the data split into two profile vectors for the distinct time periods, we compute the reidentifiability metric at various subvectors of rank k. Once we had the vector of evaluated reidentifications (True/False for each user), we resampled it 10,000 times to find the bootstrapped confidence interval for the rate of reidentifiability. Figure 7 shows these results for both the set of all observed domains and the Trexa list domains. The first thing we can observe from Figure 7 is

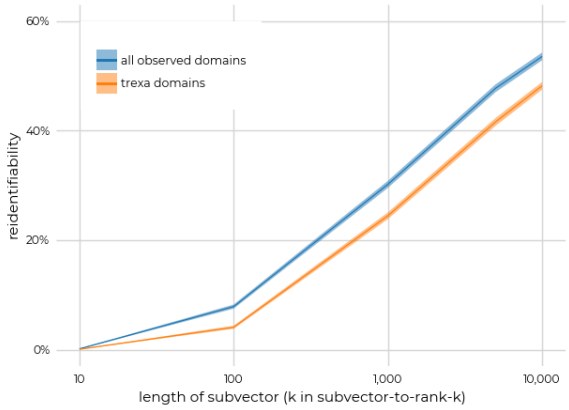


Figure 7: Reidentifiability over two domain lists for 19,263 users with a profile size over 50

the tightness of the 95% confidence intervals due to the population size. The relative width of these intervals can be contrasted with those in Figure 9. Second, we observe that, unsurprisingly, the more domains included in the computation, the higher the reidentifiability rates. This makes intuitive sense: as we increase the rank of the subvector, we include

both less-common sites and combinations of sites that are more likely to be specific to a particular user profile. Interestingly, we note that although a subvector truncated at rank 100 leads to a high proportion of unique profiles (92-95%), the reidentifiability rates are below 10% for both datasets. However, when we include the subvector to rank 10,000, reidentifiability rates grow to ~50%. Additionally, the difference in reidentifiability across the two datasets is smaller than we may have expected given the relatively large difference between the two lists.

The upward trend in reidentifiability as the length of the subvector increases makes sense, but we must also consider the effect of stability of browsing activity. The more consistent a history is, the easier it is to reidentify with smaller amounts of data. For example, consider a light internet user that regularly goes to the same 5 websites, compared with a heavy internet user that spends many hours a day browsing but rarely the same domains. Understanding types of users, and patterns of browsing behavior, is a relevant problem but outside the scope of this paper.

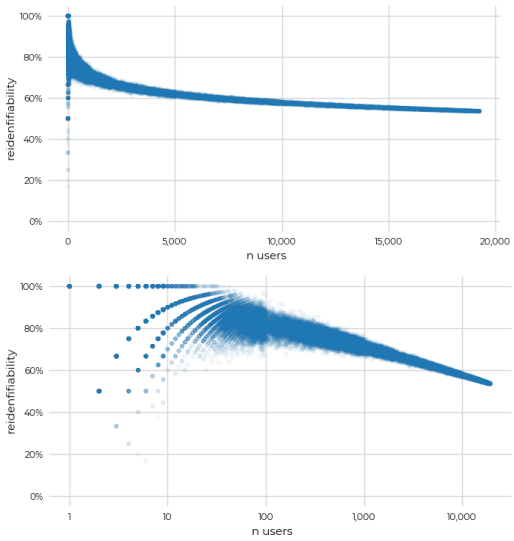


Figure 8: How reidentifiability scales with the number of users in a pool. Simulation for all users with profile size over 50. Linear axis (above) and log axis (below).

5.3 Modeling scaling effects

Another concern for a real-world implementation is the scale of the user pool. In a pool of a single user, reidentifiability is necessarily 100%. As the pool of potential candidates grows, the signal must be increasingly specific in order to correctly match two user profiles.

Measuring scalability is ideally done by collecting the browsing data of millions of people. Unfortunately, this method is infeasible within the limits of our research. Additionally, this type of collection involves privacy risk, although we note that this practice is commonly employed by companies engaged in cross-site tracking activity.

To explore scalability with the data we have available to us, we perform a Monte Carlo simulation with the subset of users with a profile size over 50. We sample between 1 and the max number of users 55,000 times without replacement, and calculate the subvector of observed domains at rank 10,000 for each user's profile. The sampling volumes were designed to give good coverage over the log space of n users. The results are shown in Figure 8. The log scale in the lower plot causes the striations in reidentifiability on the left of the lower plot, which are an artifact of the limited value space for reidentifiability with small n ($n = 1$ has a reidentifiability rate set of $\{1\}$, $n = 2$ has $\{0, 0.5, 1\}$, etc.).

The top chart suggests that the reidentifiability rate reaches an asymptote, bolstered by the linearity on the log scale chart. However, it does not feel prudent to say that the trend is certain for millions of users. We leave this analysis to future research. We can, however, make claims about the effects of reducing the pool of users. Roughly speaking, a 10-fold reduction in the number of users increases reidentifiability by 10%. This is relevant as we now turn to looking at how reidentifiability risk increases as the number of domains in a profile increases.

5.4 Effect of profile size

Understanding the effect of profile size on reidentifiability can illuminate certain threat models. For example, if only a modest threshold for profile size is needed, can a short private browsing mode session provide enough activity to allow reindentifiability, despite the precautions taken?

To understand the effect of profile size alone it is important to constrain our data in new ways. We divide the data into 7 groups with profile sizes as shown in Table 1. We selected buckets that added 25 domains at a time so that the smallest bucket, by number of users, still had a meaningful 1,766 users. As shown in Section 5.3, the number of users in a reidentifiability pool affects the reidentifiability rate, so we constructed our buckets to all be the same size in order to isolate the effect of profile size. Specifically, we sampled without replacement 1,766 users from each bucket. We then computed the reidentifiability rate and the bootstrapped confidence interval (n=10,000) at various subvector lengths as done in Section 5.2. We repeated this process with a second downsampled set to validate that the downsampling did not overly bias the outcome. The results are presented for the first downsampled set in Figure 9. As Figure 9 shows, reidentifiability increases as profile size increases up to ~80% for users with over 150 distinct domains in their profile. We note the wider 95% confidence intervals compared to those in Figure 7, which is expected with the smaller sample in each group. If we look at the reidentifiability rate for the group

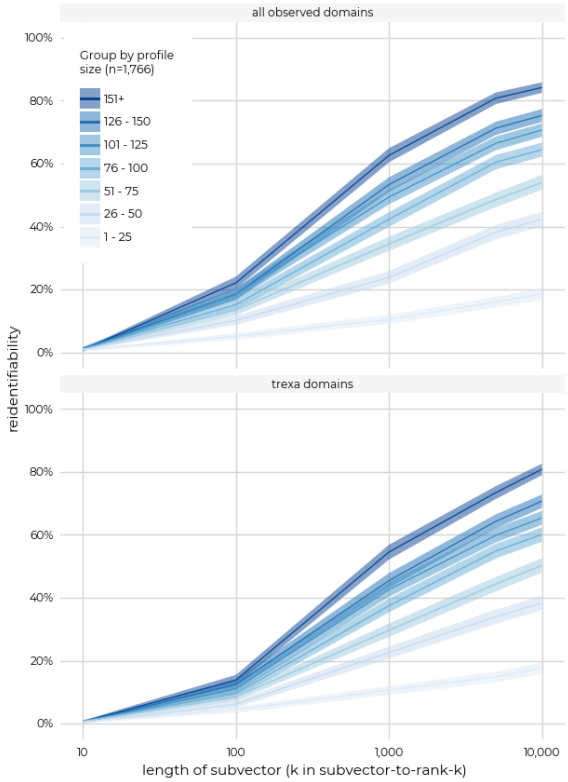


Figure 9: Reidentifiability across varying profile sizes

with profile size 51-75, we see it is a little higher than that shown for the 50 plus group shown earlier in Figure 7. There are two competing forces here: (1) the smaller pool of users in in this analysis drives higher reidentifiability, but (2) the cap on the size of the profile at 75, as opposed to no cap in Figure 7, limits reidentifiability.

We observe that as the size of bucket increases, the reidentifiability rate increases. However, the increase diminishes with each additional increase in profile size. At rank 10,000 the reidentifiability rate jumps ~20% when the profile size increases from 1-25 to 26-50. However, the increase is less than 5% when going from 101-125 to 126-150. The additional separation seen by the largest group (151+) is an artifact of that bucket incorporating some much larger profiles.

As seen in section 5.1, Trexa reidentifiability rates are not dramatically different despite the difference in these domain lists. This similarity highlights the viability of history sniffing attacks as tracking vectors, event though we found Trexa to only have 52% overlap with observed user top sites. In Figure 9, we note the "contraction" in the reidentifiability rates at k=100, which is consistent with our observation that the top 100 Trexa list overlaps with the observed browsing history less than it does for the top 10,000.

5.5 Reidentifiability with category profiles

Using the category profiles from Section 4.1 we computed reidentifiability rates with the full-length 281 category vector for the same buckets of users as Section 5.4. The rates are shown below in Table 2. The reidentifiability for categories

Profile size	Category-based reidentifiability rate (95% confidence interval)
1 - 25	5.8% (4.8% - 6.9%)
26 - 50	10.0% (8.6% - 11.4%)
51 - 75	14.3% (12.7% - 16.0%)
76 - 100	15.7% (14.0% - 17.4%)
101 - 125	16.5% (14.8% - 18.3%)
126 - 150	19.1% (17.3% - 21.0%)
151+	23.3% (21.3% - 25.3%)
51+	8.1% (7.7% - 8.5%)

Table 2: Reidentifiability rate over category profiles

is limited but non-zero. The rates are consistent with domain vectors of a similar length. We note, as in Section 5.2, that although this small number of categories is sufficient to yield a large number of distinct profiles, it does not yield high reidentification using our metric. As noted by Olejnik et al., the category analysis is an obvious candidate for a multinomial approach, using the frequency of category visits instead of the binary vector. However, as outlined in Section 5.1, we chose to stick to a single reidentifiability metric. We look forward to further work examining the impact of a more sophisticated model.

5.6 Third-party reidentifiability

The original paper [48] constrained an examination of third parties to the subset of domains on which third-party scripts from Google or Facebook were observed. Our data collection included all request and response data during the data collection period, allowing a direct identification of all third parties present on a site during a user's visit. However, the very definition of a third-party relationship at the domain level has become increasingly complex. On the web today, individual actors, such as Oracle or Wikipedia, operate their services on multiple domains. This presents three problems: (1) The reach of one third-party domain does not accurately characterize an entity's reach; (2) an entity using a separate domain to serve their own content on their main site, for example media on wikipedia.org is hosted by wikimedia.org will be identified as a third party; and (3) an entity may have a corporate or operational structure such that it has insight into data collection performed by other companies, through common ownership or data exchange agreements. To overcome these challenges we use the webXray domain list [37, 38] which connects domains with corporate entities and codifies corporate structures. For consistency we associate all domains with

the highest level parent entity in the webXray dataset. So, for example, we capture the third-party reach of Alphabet, Google's parent company.

We mark all requests with the parent entity they are associated with. If the parent entities of the request and the navigation domain differ, the request marked as third party. We measured the most prevalent third parties with two metrics: the number of first parties a third-party entity is present on, and the number of users who were exposed to that third party. We took the union of the top 50 by each metric to create a list of 61 prevalent third parties. To preserve privacy, we filtered this list to only contain third-party entities to which more than 5,000 users had been exposed. This step resulted in the removal of 1 third party, leaving a list of 60 third-party entities to analyze.

We started with the top 10,000 observed domains and reduced it to the set visible by the third-party entity. This limits the set of domains we are considering compared to the full reach of the entities. However, it also makes the result directly comparable with Figure 9, along with being more computationally tractable. The results are presented in Figure 10 for three groups of users: the complete set of 19,263 users with over 50 domains in their complete profile shown in Figure 7, the set of 1,766 users with a profile size over 150 domains shown in Figure 9, and a randomly selected set of 1,766 users from those with a profile size over 50. This last set provides a link between the other two, enabling direct comparisons for both number of users and profile size. Figure 10 provides a number of important insights. Firstly, we observe that the steps between the three user groups are consistent across third parties, and are consistent with our results in earlier sections. We note that the results are presented as theoretical reidentifiability rate in order to highlight that we are not claiming the entities presented are all performing this kind of an attack. That said, TapAd and Drawbridge, two leading probabilistic trackers [10, 13], are present in our top third-party list with a theoretical reidentifiability rate of approximately half the rate we obtained with complete information. Alphabet and Facebook, the two entities studied in the original paper, have close to our computed maximum, which is significantly higher than what the original authors found relative to their core metric of uniqueness. We summarize third parties as being highly prevalent and with the potential for deep visibility into browsing activity; that potential for surveillance is not in and of itself proof thereof.

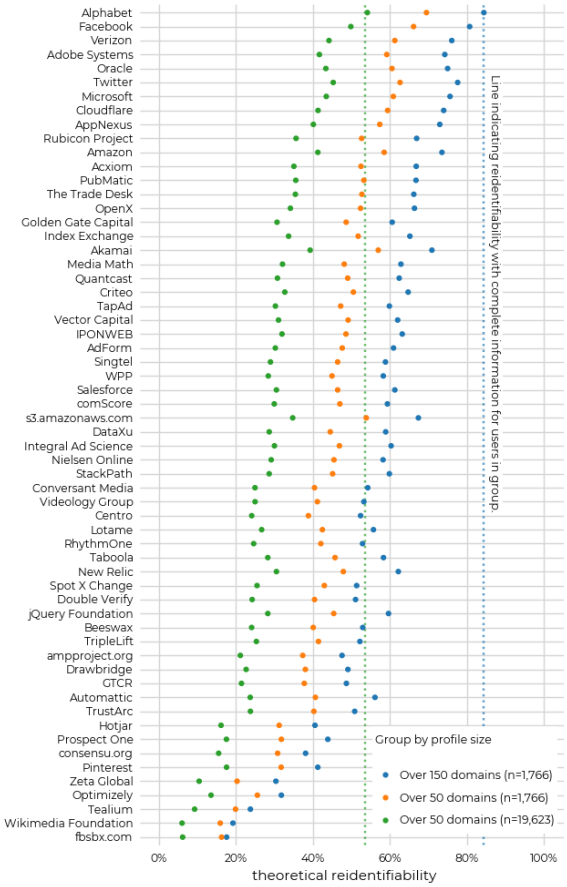


Figure 10: Theoretical third-party reidentifiability rates

6 Discussion

Our findings on profile uniqueness replicate those of Olejnik et al. [48] and our work on reidentifiability provides robust evidence for the viability of browser profiles as a tracking identifier. Interpretation of these extended findings as it pertains to third parties requires a nuanced examination of the underlying data. Notably, Wikimedia Foundation was observed on 395 of the top 10,000 first-party domains. This is low in relation to Alphabet (parent company of Google) at 9,823, and Facebook at 7,348, and the numerous companies from Verizon to TrustArc with a presence on 2,000 - 5,500 of the top 10,000 first-party domains. Nonetheless, 395 first-party domains could still be considered surprising for a non-profit organization likely not engaging in data surveillance activities. The surprisingly large presence of Wikimedia is at least partially due to user-generated content, where individuals make use of web platforms to share content via third-party links. This observation raises important questions about methodological due diligence when interpreting third-party relationships for the purpose of privacy and security research.

To use a history profile as a tracking vector, one must first be created. This means an entity requires some visibility into browsing behaviour via another tracking identifier to enable data collection over a time period, or they must perform a history sniffing attack. In some jurisdictions, history sniffing is likely illegal [9] and in the US led to the FTC bringing charges against a company [16]. This type of attack is very hard for users to prevent, and responsibility rests with researchers and

browser vendors to remain vigilant in monitoring vulnerabilities such as those disclosed by Smith et al. [58] to protect global web users.

The most prevalent tracking identifier is a browser cookie; however, as awareness of cookies has increased so have user protections. Currently, tracking protections built into Firefox, Edge, Safari, and Brave attempt to limit or block tracking content. Protections are not complete for a variety of reasons; often compromises are made to avoid unexpected errors in the way web pages are loaded as web browsers must correctly display web content while selectively blocking content designed by web developers. In January 2020, Chrome made headlines announcing that they will phase out third-party cookies [57]. The W3C Privacy Community Group, with members from all major browsers and a wide-range of stakeholders, proposes going much further with Client-Side Storage Partitioning (CSSP) [28]. CSSP will link access to a wide-range of browser features to a first-party domain. In a simple case, if `tracker.example` sets a cookie on `a.example`, on `b.example`, `tracker.example` will not see that cookie. In principle, CSSP will make the web profoundly more private than it is today, and its scope is far beyond cross-site cookies. Versions of this type of protection can be seen in Safari [61, 62], Firefox Nightly [12], Tor Browser [52], and Brave [11] as of April 2020.

Even if traditional stateful tracking is addressed, IP address tracking and fingerprinting are a real concern as ongoing privacy threats that can work in concert with browser history tracking. We point readers to Mishra et al.'s [42] discussion on IP address tracking and possible mitigations. They observed IP addresses to be static for as long as a *month* at a time, and while not a perfect tracker, IP addresses are trivial to collect. Our reidentifiability metric tracked users from one week to the next using only one week of web browsing history to build a profile (see Section 5). Technical solutions such as VPNs and Tor can offer imperfect protection [42, 49], as can efforts on behalf of Internet service providers to refresh users' IP addresses more frequently.

There is a growing technical emphasis on efforts to mitigate browser fingerprinting. Gómez-Boix et al. [27] found that browser fingerprints are not perfect identifiers, but did observe that 95% of desktop users had fingerprints that matched 100 or fewer other users. Presuming that device-specific properties (screen size, graphics stack) are not correlated with browsing history, then fingerprinting offers a mechanism to greatly narrow a pool of users in which to associate browsing histories. As we have shown, the smaller the pool of users under consideration, the higher the reidentifiability rates achievable with browsing history alone (see Section 5.3).

We are encouraged that fingerprintability is a consideration in web specifications [19], but a significant amount of work remains if we are to prevent fingerprinting and we hope researchers and browser vendors will push aggressively on anti-fingerprinting measures.

6.1 User-facing recommendations

Until the state of the web has improved, the onus of ensuring privacy often falls on the user. However, evidence suggests that general audiences are not aware that privacy protection tools exist, let alone understand their function or the threats they protect against [14]. Additionally, Fagan et al. [24] describe the trade-off between convenience and security that often results in non-compliance with security measures unless the risk can be understood and evaluated.

Currently, users can limit their exposure to third-party tracking by opting into tools such as privacy addons [25, 39, 41], containers [43], and by modifying their browser settings to emphasize existing privacy features that may result in a diminished browsing experience. Blocking fingerprinting via the Disconnect list in Edge and Firefox offers users protection, but relies on a blocklist of known trackers. Privacy addons offer fingerprinting protections through lists such as EasyList and EasyPrivacy [21], or a fine-grained control over locally curated exclusion lists, which require substantial effort by users to maintain. Tor Browser includes changes to reduce the entropy of the browser by normalizing the return values of various APIs [52]. This functionality is also available in Firefox behind the `privacy.resistFingerprinting` configuration flag, but introduces compatibility issues with some websites. Chrome's Privacy Sandbox project [15] proposes the concept of a privacy budget [35] that only allows a certain amount of entropy to accumulate, along with reducing entropy and removing fingerprinting surfaces.

As the existence of these tools alone is not enough to prevent a tracking threat, well-researched user experience practices should be considered in order to increase their adoption and make their usage more seamless while mitigating drawbacks. Mathur et al. [40] propose that browsers should automatically provide comprehensive protection in particularly sensitive contexts such as private browsing mode. We did not collect data when a private browsing mode session was active, and so cannot validate our threat model in this context. However, the potentially sensitive nature of private browsing indicates that the need to avoid tracking may overpower the need to avoid breakage or inconvenience the user. Developers of products with private browsing modes should not only encourage their use when appropriate, but enforce short, task-focused sessions to reduce the number and diversity of domains visited, potentially by purging a session when a threshold has been reached or by encouraging good user habits via messaging, suggestions, or other interaction design.

Our research leaves a number of open questions that must be considered in the face of designing user interventions for browser history tracking. In particular, there are questions about how to educate users and offer meaningful privacy protections against the kind of complex and abstract threat model we outline. The many tools that currently exist to preserve privacy should be encouraged within the context of a well-

explained, specific threat to empower users to opt in to these tools on an ongoing basis.

6.2 Limitations and future work

It is our hope that these findings will be appropriately interpreted despite, and in context of, limitations to the data collection methodology, reidentifiability metric, and the analysis of third-party relationships. As with any opt-in data collection, our data represents a biased sample. We cannot know the exact effect of this bias; however, we suggest that our sample may exhibit greater homogeneity than we expect from the overall internet-connected population. This may have resulted in an underestimation of the ease with which participants can be reidentified based on browsing histories.

We believe our replication provides an important and complementary extension to the original, as we were able to directly study reidentification threats. Although we have similar concerns regarding the feasibility with which such tracking threats can be scaled to billions of web users, our discussion in Section 6 provides anecdotal evidence that it is feasible to combine independent tracking vectors, which would can dramatically reduce the required scale for viability.

Our model leverages a very basic metric to demonstrate reidentifiability. It was our intention to focus on the question of moving from uniqueness of profiles to a baseline reidentifiability rate. As described in section 5.1, there is a multitude of more advanced techniques available. For example, Banse et al. [7] achieved an 88.2% reidentification rate among 2,100 users based on daily browsing data using a Naive Bayes classifier and cosine distances. Perhaps more importantly, companies engaged in this kind of activity have access to the full range of tracking data: IP addresses, cookie-based identifiers, identifiers from cookie-syncing, click attribution identifiers, browser fingerprints, and user-generated identifiers such as social media handles or email addresses used for logins. We do not want to over-speculate on the relative impact of these factors, but we do believe that the approach presented in this work, in spite of the above caveats, is conservative when it comes to the potential for using browser history as a tracking profile.

We would like to highlight a core limitation of our approach to measuring third parties at the entity level and why we believe it is still preferable to other strategies for discerning relationships between web resources. This limitation is exemplified by the presence of `fbsbx.com`, a Facebook-owned domain, on our top third-party list. This domain was observed on only 351 of the top 10,000 first-party domains, but its inclusion in our list was based on its visibility to users as a third party. This was driven by the fact that the webXray dataset does not currently associate `fbsbx.com` as being part of the Facebook entity and so requests to `fbsbx.com` were marked as third-party in nature when they originated from visits to `facebook.com` and other popular Facebook owned first-party sites. It is difficult to know how many times this has happened during our data analysis; we consider the implications of this finding and caveat our results accordingly. If a domain is not associated with an entity in the webXray list it has the following possible effects: (1) The presence / traffic of that domain in a third-party context is not correctly associated with the entity, thereby lowering the power of the third-party in our reidentifiability analysis, and (2) Missing the third-party relationship with a domain only affects the reidentifiability analysis by the number of first-party domains the third party has been misidentified on. In the case of Facebook not being linked with `fbsbx.com`, we may have incorrectly labeled it as having a third-party relationship with Facebook domains such as `facebook.com`, `instagram.com`, `whatsapp.com`. The fact that `fbsbx.com` is not correctly associated with Facebook has lowered the Facebook reidentifiability rate by up to 350 domains, but it has raised the `fbsbx.com` domain rate by only a handful of domains. We are comfortable with this trade-off despite the fact that using an entity list introduces a dependency on its accuracy and completeness. Correctly characterising a third-party relationship in the context of web research is, itself an immensely complex undertaking and there is an abundance of future work to be done in this space.

Finally, we did not find all profiles to be unique or reidentifiable. Future work should leverage the common traits in these non-unique profiles in order to inform strategies for privacy tools and education development.

7 Conclusion

In summary, we set out to replicate and expand upon the ideas put forth in Olejnik et al.'s 2012 paper *Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns*. The original paper observed a set of ~400,000 web history profiles, of which 94% were unique. Our set of 48,103 distinct browsing profiles, of which 99% are unique, followed similar distributions as the original. Likewise, these patterns held when we used a public top-site list and category mappings to restrict visibility into the number of domains considered, mimicking the methodology of the original work.

Olejnik et al. found evidence for profile stability among a small pool of returning users. We extend this work and modeled reidentifiability directly for nearly 20,000 users. We reidentify users from two separate weeks of browsing history, and examine the effect of profile size, and how reidentifiability scales with the number of users under consideration. Our reidentifiability rates in a pool of 1,766 were below 10% for 100 sites despite a >90% profile uniqueness across datasets, but increased to ~80% when we consider 10,000 sites. Finally, while Olejnik et al. show somewhat lower uniqueness levels for profiles of pages tracked by Google and Facebook, we show theoretical reidentifiability rates for some third-party entities nearly as high as those we achieve with complete knowledge of all visited domains.

Acknowledgments

We extend a sincere thanks to Olejnik, Castelluccia, and Janc, the authors of the original manuscript [48], for providing the inspiration for this work and their contributions to the advancement of privacy technology. In particular, we thank Lukasz Olejnik for his communication and collegiality throughout the duration of this research.

The authors also thank the numerous individuals who participated in reviewing both the codebase enabling this work, and earlier versions of the manuscript. Thanks to David Zeber, Fred Wollsén, and Victor Ng on the Firefox Machine Learning team for supporting this research line as well as Mozilla's data pipeline engineering, data operations, and quality assurance teams.

Finally, the authors would like to thank all of the Firefox users who continue to trust us with their private data through their voluntary enrollment into experimental initiatives.

References

[1] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, page 674–689, New York, NY, USA, 2014. Association for Computing Machinery.

[2] Acxiom. Omnichannel identity graphs. Technical report, Acxiom, 2017.

[3] Acxiom. New face of data driven marketing - buyers-guide. Technical report, Acxiom, 2019.

[4] Acxiom. What is a customer data platform? is a cdp the best solution? Technical report, Acxiom, 11 2019.

[5] Amazon. Alexa - Top sites. https://www.alexa.com/topsites, 2019.

[6] Kalvin Bahia and Stefano Suardi. Internet Society Global Internet Report: Consolidation in the Internet Economy. Technical report, The Internet Society, January 2019. Date accessed: 08-02-2020.

[7] Christian Banse, Dominik Herrmann, and Hannes Federrath. Tracking users on the internet with behavioral patterns: Evaluation of its practical feasibility. In *IFIP Advances in Information and Communication Technology*, pages 235–248. Springer Berlin Heidelberg, 2012.

[8] David Baron. 147777 - visited support allows queries into global history. https://bugzilla.mozilla.org/show_bug.cgi?id=147777, May 2005.

[9] Steve Bellovin. Federal trade commission - history sniffing. https://www.ftc.gov/news-events/blogs/techftc/2012/12/history-sniffing, 2012.

[10] Ricardo Bilton. Cross-device tracking, explained, August 2015.

[11] Brave. OK Google, don't delay real browser privacy until 2022. https://brave.com/ok-google/.

[12] Martin Brinkman. Mozilla adds Dynamic First Party Isolation option to Firefox 77 - gHacks Tech News. https://www.ghacks.net/2020/04/17/mozilla-adds-dynamic-first-party-isolation-option-to-firefox-77/, April 2020.

[13] Justin Brookman, Phoebe Rouge, Aaron Alva, and Christina Yeung. Cross-Device Tracking: Measurement and Disclosures. *Proceedings on Privacy Enhancing Technologies*, 2017(2):133–148, 2017.

[14] Farah Chanchary and Sonia Chiasson. User perceptions of sharing, advertising, and tracking. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 53–67, Ottawa, July 2015. USENIX Association.

[15] Chromium. The Privacy Sandbox - The Chromium Projects. https://www.chromium.org/Home/chromium-privacy/privacy-sandbox.

[16] Federal Trade Commission. FTC Approves Final Order Settling Charges Against Epic Marketplace, Inc. https://www.ftc.gov/news-events/press-releases/2013/03/ftc-approves-final-order-settling-charges-against-epic, March 2013.

[17] John Cook, Rishab Nithyanand, and Zubair Shafiq. Inferring tracker-advertiser relationships in the online advertising ecosystem using header bidding. *Proceedings on Privacy Enhancing Technologies*, 2020(1):65 – 82, 2020.

[18] Elizabeth Denham. Update report into adtech and real time bidding. Technical report, UK Information Commissioner's Office, 2019.

[19] Nick Doty. Mitigating Browser Fingerprinting in Web Specifications. https://w3c.github.io/fingerprinting-guidance/.

[20] Roberto Díaz-Morales. Cross-Device Tracking: Matching Devices and Cookies. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1699–1704, 2015.

[21] EasyList. EasyList - Overview. https://easylist.to/.

[22] Peter Eckersley. How unique is your web browser. Technical report, Electronic Frontier Foundation, San Francisco, CA, 2010.

[23] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 1388–1401, New York, NY, USA, 2016. ACM.

[24] Michael Fagan and Mohammad Maifi Hasan Khan. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 59–75, Denver, CO, June 2016. USENIX Association.

[25] Gertjan Franken, Tom Van Goethem, and Wouter Joosen. Who Left Open the Cookie Jar? A Comprehensive Evaluation of Third-Party Cookie Policies. In *Proceedings of the 27th USENIX Security Symposium*, page 19, 08 2018.

[26] Melissa Gallo. Taxonomy: The most important industry initiative you've probably never heard of. Technical report, Interactive Advertising Bureau (IAB), 2016.

[27] Alejandro Gómez-Boix, Pierre Laperdrix, and Benoit Baudry. Hiding in the crowd: An analysis of the effectiveness of browser fingerprinting at large scale. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 309–318, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[28] W3C Privacy Community Group. Client-Side Storage Partitioning. https://privacycg.github.io/storage-partitioning/, 2020.

[29] Dominik Herrmann, Christoph Gerber, Christian Banse, and Hannes Federrath. Analyzing characteristic host access patterns for re-identification of web user sessions. In *Information Security Technology for Applications, NordSec 2010*, volume 7127, pages 136–154. Springer, Berlin, 01 2010.

[30] Muhammad Ikram, Hassan Jameel Asghar, Mohamed Ali Kâafar, Balachander Krishnamurthy, and Anirban Mahanti. Towards seamless tracking-free web: Improved detection of trackers via one-class learning. *CoRR*, abs/1603.06289, 2016.

[31] Umar Iqbal, Zubair Shafiq, and Zhiyun Qian. The ad wars: Retrospective measurement and analysis of anti-adblock filter lists. In *Proceedings of the 2017 Internet Measurement Conference*, IMC '17, page 171–183, New York, NY, USA, 2017. Association for Computing Machinery.

[32] Kaggle. ICDM 2015: Drawbridge Cross-Device Connections. https://kaggle.com/c/icdm-2015-drawbridge-cross-device-connections.

[33] Andrew J. Kaizer and Minaxi Gupta. Towards automatic identification of javascript-oriented machine-based tracking. In *Proceedings of the 2016 ACM on International Workshop on Security And Privacy Analytics*, IWSPA '16, page 33–40, New York, NY, USA, 2016. Association for Computing Machinery.

[34] Arjaldo Karaj, Sam Macbeth, Rémi Berson, and Josep M. Pujol. Whotracks.me: Monitoring the online tracking landscape at scale. *CoRR*, abs/1804.08959, 2018.

[35] Brad Lassey. Privacy budget. https://github.com/bslassey/privacy-budget, 2019.

[36] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. *arXiv e-prints*, 1806:arXiv:1806.01156, June 2018.

[37] Timothy Libert. An automated approach to auditing disclosure of third-party data collection in website privacy policies. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 207–216, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[38] Timothy Libert. webxray domain ownerlist. https://github.com/timlib/webXray_Domain_Owner_List, 2020.

[39] Arunesh Mathur, Jessica Vitak, Arvind Narayanan, and Marshini Chetty. Characterizing the Use of Browser-Based Blocking Extensions To Prevent Online Tracking. In *Fourteenth Symposium on Usable Privacy and Security ({SOUPS} 2018)*, pages 103–116, 2018.

[40] Arunesh Mathur, Jessica Vitak, Arvind Narayanan, and Marshini Chetty. Characterizing the use of browser-based blocking extensions to prevent online tracking. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 103–116, Baltimore, MD, August 2018. USENIX Association.

[41] G. Merzdovnik, M. Huber, D. Buhov, N. Nikiforakis, S. Neuner, M. Schmiedecker, and E. Weippl. Block Me If You Can: A Large-Scale Study of Tracker-Blocking Tools. In *2017 IEEE European Symposium on Security and Privacy (EuroS P)*, pages 319–333, 04 2017.

[42] Vikas Mishra, Pierre Laperdrix, Antoine Vastel, Walter Rudametkin, Romain Rouvoy, and Martin Lopatka. Don't count me out: On the relevance of IP addresses in

the tracking ecosystem. In *Proceedings of the 29th International Conference on World Wide Web*, WWW '20. International World Wide Web Conferences Steering Committee, April 2020.

[43] Mozilla. Multi-Account Containers. https://support.mozilla.org/en-US/kb/containers.

[44] Mozilla. Mozilla Privacy Policy. https://www.mozilla.org/en-US/privacy/, May 2018.

[45] Mozilla. Introducing firefox pioneer. https://addons.mozilla.org/en-US/firefox/pioneer, 2019.

[46] Mozilla. Jestr pioneer shield study. https://github.com/mozilla/jestr-pioneer-shield-study, 2019.

[47] Mozilla. Firefox data collection - mozillawiki. https://wiki.mozilla.org/Firefox/Data_Collection, 2020.

[48] Lukasz Olejnik, Claude Castelluccia, and Artur Janc. Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns. In *5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2012)*, July 2012.

[49] Ramakrishna Padmanabhan, Amogh Dhamdhere, Emile Aben, kc claffy, and Neil Spring. Reasons Dynamic Addresses Change. In *Proceedings of the 2016 Internet Measurement Conference*, pages 183–198. Association for Computing Machinery, 11 2016.

[50] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *The World Wide Web Conference*, WWW '19, page 1432–1442, New York, NY, USA, 2019. Association for Computing Machinery.

[51] Panagiotis Papadopoulos, Nicolas Kourtellis, Pablo Rodriguez Rodriguez, and Nikolaos Laoutaris. If you are not paying for it, you are the product: How much do advertisers pay to reach you? In *Proceedings of the 2017 Internet Measurement Conference*, IMC '17, page 142–156. Association for Computing Machinery, 2017.

[52] Mike Perry, Erinn Clark, Steven Murdoch, and Georg Koppen. The Design and Implementation of the Tor Browser [DRAFT]. https://2019.www.torproject.org/projects/torbrowser/design/.

[53] S. Pretorius, A. R. Ikuesan, and H. S. Venter. Attributing users based on web browser history. In *2017 IEEE Conference on Application, Information and Network Security (AINS)*, pages 69–74, Nov 2017.

[54] PRWeb. 1st place finish for datalab usa at icdm 2015: Drawbridge cross-device connections. https://www.prweb.com/releases/2015/09/prweb12938138.htm.

[55] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. Detecting and defending against third-party tracking on the web. In *Presented as part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, pages 155–168, San Jose, CA, 2012. USENIX.

[56] Sebastian Schelter and Jérôme Kunegis. On the ubiquity of web tracking: Insights from a billion-page web crawl. *The Journal of Web Science*, 4(4):53–66, 2018.

[57] Justin Schuh. Building a more private web: A path towards making third party cookies obsolete. https://blog.chromium.org/2020/01/building-more-private-web-path-towards.html.

[58] Michael Smith, Craig Disselkoen, Shravan Narayan, Fraser Brown, and Deian Stefan. Browser history re:visited. In *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18)*, 2018.

[59] Firefox Machine Learning Team. Trexa. https://github.com/mozilla/trexa.

[60] Webshrinker. Webshrinker apis. https://www.webshrinker.com/apis/, 2020.

[61] John Wilander. Full third-party cookie blocking and more. https://webkit.org/blog/10218/full-third-party-cookie-blocking-and-more/.

[62] John Wilander. Introducing storage access api. https://webkit.org/blog/8124/introducing-storage-access-api/.

[63] David Zeber. Top site list comparison analysis. https://github.com/mozilla/research-repo-webconf-crawl-representativeness.

[64] David Zeber, Sarah Bird, Camila Oliveira, Walter Rudametkin, Ilana Segall, Fredrik Wollsén, and Martin Lopatka. The representativeness of automated web crawls as a surrogate for human browsing. In *Proceedings of the 29th International Conference on World Wide Web*, WWW '20. International World Wide Web Conferences Steering Committee, 2020.

How Risky Are Real Users' IFTTT Applets?

Camille Cobb
Carnegie Mellon University

Milijana Surbatovich
Carnegie Mellon University

Anna Kawakami
Wellesley College

Mahmood Sharif
NortonLifeLock

Lujo Bauer
Carnegie Mellon University

Anupam Das
North Carolina State University

Limin Jia
Carnegie Mellon University

Abstract

Smart-home devices are becoming increasingly ubiquitous and interconnected with other devices and services, such as phones, fitness trackers, cars, and social media accounts. Built-in connections between these services are still emerging, but end-user-programming tools such as *If-This-Then-That* (IFTTT) have existed for almost a decade, allowing users to create rules (called *applets* in IFTTT) that dictate interactions between devices and services. Previous work found potential secrecy or integrity violations in many applets, but did so without examining how individual users interact with the service. In this work, we study the risks of real-world use of IFTTT by collecting and analyzing 732 applets installed by 28 participants and participants' responses to several survey questions. We found that significantly fewer applets than previously thought pose realistic secrecy or integrity risks *to the users who install them*. Consistent with this finding, participants were generally not concerned about potential harms, even when these were explained to them. However, examining participants' applets led us to identify several new types of privacy risks, which challenge some assumptions inherent in previous analyses that focus on secrecy and integrity risks. For example, we found that many applets involve monitoring *incidental users*: family, friends, and neighbors who may interact with someone else's smart-home devices, possibly without realizing it. We discuss what our findings imply for automatically identifying potentially harmful applets.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

1 Introduction

Smart home technology has made its way into public consciousness and widespread use [3]. On their own, smart-home devices typically allow users to control them via dedicated apps, possibly creating schedules, routines, or triggering notifications from the apps on users' phones. Additionally, many smart-home devices enhance their capacity for home automation by interfacing with end-user programming tools such as If-This-Then-That (IFTTT), Stringify, and WebHooks. Such tools allow users to create trigger-action "rules" that react to and/or control their IoT devices and services like social media, cloud storage, or news. This enables users to accomplish home automation tasks that would not be possible otherwise. For example, a user could create a rule to automatically turn on all their smart lights when they arrive home, even if those lights were made by a variety of manufacturers. While these tools can enable creative, beneficial uses of smart-home technologies, they may also introduce security and privacy risks.

Prior work found that as many as 50% of applets shared on the IFTTT webpage could lead to secrecy or integrity violations (i.e., leak private information or allow unauthorized access to a user's devices and services) [35]. That study, and others (e.g., [8, 10, 11, 28, 38]), sought to understand and measure the prevalence and magnitude of security and privacy risks of end-user programming with trigger-action rules, and they have proposed automated ways of identifying risky rules—rules that have the potential to cause harm—with an end-goal of mitigating risks. However, these studies have relied on publicly available data (e.g., applets shared on the IFTTT webpage) and have not evaluated risks in the context of individual users' sets of rules, the contexts in which those rules are applied, or the individuals' privacy preferences.

In this paper, we seek to better contextualize our understanding of the ways that users employ end-user programming in order to answer open questions about the secrecy, integrity, and other security and privacy risks their rules may create. To do so, we focus specifically on IFTTT, which is the most popular end-user-programming tool [25]. We recruited 28 IFTTT

users via popular home-automation message boards. Participants allowed us to collect data about their IFTTT applets and responded to a short survey. Survey questions addressed the context in which the applets are used (e.g., who cloud storage documents are shared with), participants' understanding and perception of secrecy and integrity risks (e.g., if they had considered certain risks when setting up rules, if they had experienced any harms, and if they believed certain risks were possible for a particular rule), and how they would react to specific violations identified in prior work.

Using automated information-flow-based analysis, we found that about 57% of participants' IFTTT rules had potential secrecy or integrity violations (see Section 4.3), which is consistent with the findings of prior work analyzing applets shared on the IFTTT website. In Section 4.4, we examine participants' rules in more detail, considering context such as their titles. This more detailed analysis revealed that although many applets might technically have secrecy or integrity violations, they are rarely harmful because of these violations. Only about 10% of the secrecy-violating rules (just over 3% of all rules) could lead to secrecy harms, and just under 20% of integrity-violating rules (8.6% of all rules) present serious integrity-related risks. Consistent with our manual evaluation, participants did not believe that their rules were likely to lead to secrecy- or integrity-related harms, though they did care about the security and privacy of their rules.

Our contextualized analysis of trigger-action rules and their security and privacy risks is a key contribution of this work and also led to unexpected findings. Although secrecy and integrity violations rarely pose risks to IFTTT users, IFTTT rules pose other types of security and privacy risks that have not been identified through automated analysis. For example, IFTTT rules can create surveillance risks to *incidental users*—people besides the IFTTT user who created the rule. In Section 5, we discuss these other types of risks, as well as other limitations of the information-flow analysis. From our findings we draw guidelines for how automated analysis tools could better distinguish between practically *risky* and merely theoretically *violating* trigger-action rules. We also propose future research to better understand incidental users' preferences regarding their interactions with smart-home devices. Identifying contextual factors needed for more accurate automated analyses and previously unexplored categories of risks are also key contributions of this study.

2 Background

2.1 Security of Smart-Home Technology

In recent years, researchers have investigated the security and privacy risks imposed by home IoT ecosystems. Most of these efforts consider the IoT ecosystem either at the application level or at the network level. At the application level, researchers have found that many applications built on emerging programming platforms such as Samsung's SmartThings [4] are over-privileged due to design flaws in their permission models [15, 17]. User-centric and context-aware permission systems have been developed for appified IoT platforms to address their coarse-grained permission flaws [16,23,37]. Systems utilizing static analysis [10, 28], model-checking [11], and data provenance graphs [38] have been proposed to help identify incorrect or inconsistent application behavior. Many research groups have proposed network-traffic-analysis-based security mechanisms [9, 12, 13, 29, 33, 34, 40]; many of these were introduced in light of the infamous Mirai attack, which took advantage of insecure IoT devices to launch a distributed denial of service (DDoS) attack [20, 30].

Differently from these studies, our work focuses on risks introduced by end-user programming. That is, we find that potential harms persist even under the assumption that technical vulnerabilities do not exist or are sufficiently unlikely.

2.2 Privacy Concerns in Smart Homes

In spite of their widespread adoption, users continue to surface privacy concerns about smart-home devices. To understand what concerns users have about smart-home technology, several interview- and survey-based studies investigated users' experiences and preferences [6, 7, 14, 36]. When IoT devices are installed in multi-person households, new security, privacy, and usability challenges emerge. Recent research has sought to identify user requirements in these multi-user settings and proposed potential solutions [19, 39, 41] such as making it easier for everyone in a household to control the devices and how they are configured [41]. Others have studied desirable access controls for smart-home devices [21, 32]. Our study also attempts to understand privacy concerns in a smart-home setting (including multi-user setting), but more so in the context of using automation services like IFTTT which can inadvertently cause harms.

2.3 End-User Programming for IoT Devices

Several end-user programming tools—including IFTTT ("If This, Then That") [1], Microsoft Flow [2] and Zapier [5]—enable users to connect multiple services by constructing simple trigger-action programs [24]; IFTTT is by far the most popular of these [25].

2.3.1 IFTTT

An IFTTT rule or "applet" (previously called "recipe") consists of a "trigger" and an "action." The trigger is the "this" and the action is the "that" in "if this then that." Shortly before our study's data collection, IFTTT added a feature to allows a single applet to have more than one action. Each trigger and action belongs to a "channel," which specifies the service provider who created the trigger or action (e.g., IoT device

manufacturer, social media company). As of November 2019, IFTTT offered 1,228 channels [1]. Some actions and triggers have additional fields that must be specified by a user. For example, in the trigger *"Amazon Alexa :: say a specific phrase,"* a user-configured field specifies the specific phrase. When users set up rules, they can edit a plaintext description of the applet, which we refer to as the applet's "title."

2.3.2 Information-Flow Analysis of IFTTT Applets

Although our main focus is to understand the potential harms of real users' IFTTT rules, we sought to ground our assessment in terms of previous estimates. Doing so enables us to assess the efficacy of previous methods of assessing IFTTT rules and to contribute insights that may improve automated analysis methods. In particular, we build on a prior study by Surbatovich et al. that applied information-flow analysis to IFTTT applets to automatically determine which rules contain potential secrecy or integrity violations [35].

Secrecy violations occur when a rule allows information to flow from a more private source (the trigger) to a less private sink (the action), possibly leaking private information to a wider audience than intended. For example, a rule that posts to Facebook each time motion is detected at the user's front door could unintentionally broadcast when the user arrives home at a suspicious time (e.g., late at night or when they should be at work).

Integrity violations occur when a rule allows a more trusted action to be *controlled* by a less trusted trigger, thus possibly allowing unintended people to perform actions they would not otherwise be able to (e.g., allowing an adversary to control a user's smart-home devices or post to their social media pages). For example, a rule that unlocks the user's home when an email is received that contains a pre-specified keyword could allow an adversary who guesses the keyword to compromise the user's home security.

Information-flow analysis as used in previous work [35] consists of three steps: (1) creating a set of secrecy and a set of integrity labels and arranging each set into a lattice that describes whether information flows are safe or constitute secrecy or integrity violations; (2) manually assigning *secrecy* and *integrity* labels to each trigger and action, conveying who can observe (secrecy) or control (integrity) the trigger or action; and (3) given the labels of trigger-action pairs and the secrecy and integrity lattices, determining whether the trigger-action pair constitutes a security or integrity violation.

Surbatovich et al. used four secrecy labels to describe who may have observed that a trigger or action event has occurred: *private* (e.g., only the IFTTT user), *restricted physical* (e.g., people in a user's home), *restricted online* (e.g., the user's Facebook friends), and *public* (e.g., anyone in the world). Similarly, they used six integrity labels to describe who can cause a trigger or action to occur: *trusted* (e.g., only the IFTTT user), *trusted other* (e.g., trusted news sources or weather reports), *untrusted* (e.g., anyone in the world), *untrusted group* (e.g., the unrestricted group people could collectively cause a trigger event such as *"reddit :: New hot post in subreddit")*, and *restricted physical* and *restricted online* (same as for secrecy labels). Detailed explanations of the labels are included in Appendix B. We replicate this analysis with minor adjustments as described in Section 3.1.

When a precise secrecy or integrity label cannot be determined (e.g., because context, such with whom a Google Sheet is shared, would affect the label), multiple labels may be assigned to denote that the label could be *any* of these. If some combination of labels for an applet constitutes a secrecy or integrity violation, we refer to it as "potentially violating."

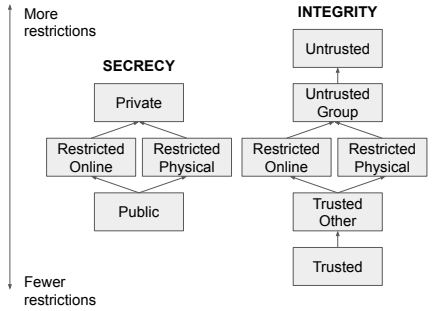


Figure 1: Secrecy and integrity lattices from prior work [35].

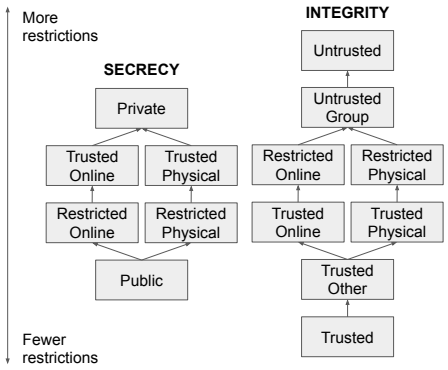


Figure 2: Modified secrecy and integrity lattices used in our analysis.

2.3.3 Other Automated Analyses of Applets

Several other research efforts have sought to identify and characterize security and privacy risks for IFTTT users. Bastys et al. found that applets shared on IFTTT for others to use are a potential attack vector [8]. In user-shared rules, filters and parameters set by the rule creator are not visible to users who adopt the rule, but can be manipulated to execute URL attacks. Wang et al. and Hsu et al. considered how "chains" of trigger-action rules that cause other rules to execute can lead to significant complexity in evaluating whether a system of connected

smart-home devices is working as intended [22,38]. They proposed solutions that provide data provenance [38] or that use dynamic model-checking techniques to prevent hidden attack chains from executing [22]. IFTTT's use of long-term OAuth tokens to connect various services creates another potential source of risk. To prevent the misuse of these tokens, Fernandes et al. proposed a Decentralized Trigger-Action Platform (DTAP) with fine-grained rule-specific tokens [18]. Our study differs from these in that we focus on analysis at the rule level; that is, we consider only risks that could arise while the applets themselves execute correctly as defined.

3 Method

Our study consisted of an online survey and automated collection of IFTTT data for 28 participants. Both were conducted with participant consent and approved by our institution's IRB. We recruited participants via posts on Home Automation Reddit and SmartThings and Stringify forums, which advertised it as a study about "the habits and behavior of IFTTT users." Participants received a $5 Amazon gift card, and one out of ten participants chosen at random received an additional $50 Amazon gift card. Participants were required to have an IFTTT account with at least five applets, in order to ensure that participants were active IFTTT users and because many survey questions referenced participants' specific applets.

Of the 67 people who started our study, 41 completed it (i.e., 61% completion rate). The main contributor to dropout was people opting out of downloading a Chrome extension to collect their IFTTT data. We also excluded 13 responses from people who did not have the requisite five applets to qualify. Thus, our final data set included IFTTT data and survey responses from 28 participants. The survey was open for eight weeks in April and May of 2018. The median time to complete the study was 18 minutes (IQR 16-24 minutes). At the end of the survey, we revealed its purpose as a security-based survey to "see users' perceptions and awareness of applets that are possibly insecure or safe," and offered participants to opt out. No participants opted out at this stage.

3.1 IFTTT Data Collection and Analysis

We created a Chrome extension to collect the applets participants had installed or created, specifically, the trigger and trigger channel, the action(s) and action channel(s), and the applet title. Participants downloaded our extension and signed into their IFTTT account so that we could collect their applets; the extension was subsequently automatically removed.

Post-hoc data analysis had two phases: (1) automated information-flow analysis to identify applets with potential secrecy and privacy violations, and (2) a qualitative analysis of each trigger and action, to group similar triggers and actions.

3.1.1 Information-Flow Analysis

Surbatovich et al. shared with us their data and codebase [35], which allowed us to reproduce the assessment of secrecy or integrity violations as they would have been evaluated in that study, as described in Section 2.3.2. This allowed us to directly compare the secrecy and integrity violations in the recipes we collected with what was previously measured. Because many new triggers and actions were added to IFTTT after the earlier study, only about 20% of participants' rules could be assessed with the pre-existing secrecy and integrity labels.

To compensate for this and for changes to the functionality of devices and services may have altered the appropriate label for existing triggers and actions (e.g., smart-home devices that added new online access capabilities for multiple household members), we relabelled *all* of the triggers and actions, rather than only adding labels for new triggers and actions. To relabel, four researchers met in three labeling sessions, totaling over six hours of meeting time. We looked up and discussed device functionality, which always allowed us to reach consensus. For example, the secrecy and integrity labels for a device can depend on whether the device's default app supports just one or multiple users.

During relabelling, we found that the original information-flow analysis was limited in its ability to differentiate between those *restricted online* groups that could be used to grant family access to smart-home devices and the much larger *restricted online* groups such as all Facebook friends. The previous analysis also did not consider situations when two *different* restricted groups had simultaneous access (e.g., smart lights that can be controlled by people in a *restricted physical* space *and* by household members via an app).

To address these limitations, we made two adjustments to the security lattices used in prior work. First, we added two labels, both of which apply to both secrecy and integrity—*trusted online* and *trusted physical*—to represent very small, trusted groups of individuals. Second, we adapted the security lattice to include *unions* of two labels (as is standard in information-flow analyses [26,27]), so that we could indicate when, for example, a *restricted physical* and *trusted online* group could *both* control a device, as is the case in many smart-home devices with apps. The modified secrecy and integrity lattices are shown in Figure 2. The information-flow labels for all triggers and actions are included in Appendix C.

3.1.2 Semantic Labelling of Triggers and Actions

To evaluate whether applets that were potentially violating are likely to lead to harm, we identified *semantic labels* to group together similar triggers and actions. For example, 20 different actions control smart lights, and the distinction between a voice command given to Alexa versus Google Assistant is likely unimportant for evaluating risks. In terms of evaluating the riskiness of rules, these semantic labels allow us to examine what an attacker could learn in the event of a secrecy

violation; and what an attacker could control in the event of an integrity violation. A single researcher identified semantic labels via open coding after repeatedly and iteratively discussing trigger and action categories collectively with the other researchers.

Semantic labels for triggers are: *Weather or time*, *News-ish*, *Sensing IoT device state*, *Environment sensing*, *Intentional trigger*, *Voice command*, *Incoming communication*, *Sensing online account state*, *Actions with personal devices*, and *Other automations*. Semantic labels for actions are: *Change IoT device state* (with optional sub-labels of *Home security* and *Lights*), *Log or notify*, *Change personal device state*, *Outgoing communication*, and *Other automations*. We intentionally created distinct categories for particularly popular types of triggers and actions (e.g., giving *Voice commands* their own label despite their similarity to *Intentional triggers*, because they are so prevalent). The labels' titles generally sufficiently describe their meaning, but more detailed explanations are included in Appendix B. Semantic labels for all triggers and actions are included in Appendix C.

Two researchers independently applied these labels to to all 160 triggers and 112 actions, with a high degree of agreement (Cohen's $\kappa = 0.93$, calculated separately for triggers and actions). We subsequently discussed and came to a consensus about disagreements. Details about rules with each semantic label are shown in Tables 3 and Table 4. In Section 4.4, we use these semantic labels as the basis for evaluating the riskiness of participants' rules, frequently referring to representative examples of the triggers and actions included in each set.

3.2 Participant Survey

Participants answered survey questions addressing three broad topics: (1) how they choose applets, (2) their beliefs and preferences about security and privacy properties of their applets, (3) harms they had experienced from using IFTTT. The full survey instrument is available in Appendix D.

After they downloaded the browser extension, we asked participants general questions about their use of IFTTT, such as how often and whether they prefer to create their own applets and if they ever turn on applets based on a friend or colleague's recommendation. We also asked participants, on a 5-point Likert scale, whether they agreed or disagreed that they would be comfortable with friends, colleagues, or "anyone" knowing what applets they use (see Table 6).

Next, we asked each participant a series of questions about up to five randomly chosen applets that were violating according to the analysis from prior work [35]. For each applet, participants were prompted to consider four situations (or five situations for applets that involve a physical device) in which the applet might contribute to harm and rate whether this would make them very upset, slightly upset, or not upset. The situations were chosen to reflect potential harms identified in previous work [35] and are listed in Table 5. Near the end of the survey, we explained the concept of secrecy and integrity violations (using lay language) and asked if considering this changed participants' desire to keep using any of their applets.

Age	25-34 (17), 35-44 (6), 45-54 (5)
Gender	Male (27), Female (1)
Education	High school graduate (2), bachelor's degree (16), mboxassociate's degree (2), professional degree (Master's/PhD) (8)
Household size	1 (2), 2 (7), 3 (10), 4 (6), 5 (2), 7 (1)
Other household members	Family (24), housemates (1), other (1), live alone (2)
IUIPC Score (Avg.)	Overall 6.01±0.66 Control 6.25±0.66 Awareness 6.58±0.41 Collection 5.23±1.44

Table 1: Summary of participant demographics.

Several additional questions, spread throughout the survey, asked participants whether they had experienced concerns or incidents related to their applets' security and privacy. For example, we asked participants whether they ever experienced an incident in which an applet made them feel unsafe or that their privacy was violated and whether they had ever manually deleted anything that was posted automatically by an applet. Participants who answered affirmatively were asked to elaborate. We concluded the survey by asking participants to provide demographic information. The survey included additional questions that we do not discuss because they were not directly applicable to the specific focus of this work.

4 Results

4.1 Participant Characteristics

Participants were predominantly men and highly educated. Table 1 shows participant demographics. Most participants (24) lived with at least one other person, typically a family member. Details about participants' living situations can help inform our risk analysis of their applets. For example, since only two participants live alone, we can assume that physical devices connected with their rules are likely to be seen and interacted with by other members of the household (e.g., ambiguity between the labels of *trusted physical* or *trusted* in the information-flow analysis could be resolved).

4.2 Characteristics of Participants' Applets

Participants had a total of 732 applets. Each participant had between 5 and 66 applets (average=26, sd=20). Most applets had a single trigger with a single action; however, seven applets had a single trigger with multiple actions. Multi-action applets were a newly-added IFTTT feature at the time of the survey. For example, P24's applet titled *"Google Home Find My Phone"* used a voice command to (1) set the Android device's ringtone volume (presumably increasing phone volume so it was easier to find) and (2) receive a phone call. For

most analyses, we use a demultiplexed version of the data, where multi-action applets are treated as multiple rules with the same trigger and different actions. To differentiate, we refer to the demultiplexed version of data as "rules" rather than "applets." There were 743 rules in our data set.

Frequency of Triggers and Actions. Participants' rules included 68 unique trigger channels with 160 unique triggers and 64 unique action channels with 112 unique actions—only a small fraction of the total triggers and actions available from IFTTT. 396 unique combinations of triggers and actions were represented in participants' rules. For reference, we include a list of the top ten channels, triggers, actions, and trigger-action pairs (i.e., those used by the most participants) in Appendix A.

Frequency of Trigger-Action Pairs. Participants used diverse trigger-action pairs, and every participant had at least one trigger-action pair that was unique to them. 63% of unique trigger-action pairs occurred only once (251 unique pairs, 34% of rules). When trigger-action combinations appeared multiple times, it was often because the same person used the same trigger-action pair repeatedly. For example, P13 had three rules with the same trigger (*Alexa :: Say a specific phrase*) and action (*Philips Hue :: Set a scene in a room*). User-specified titles differentiated these rules, in this case conveying what the specific phrase was: "trigger bright mode," "trigger sleep mode," or "trigger read mode."

89 unique trigger-action pairs were used multiple times by only a single participant, making up 251 rules (22% of unique trigger-action pairs, 34% of rules). Only 56 trigger-action combinations (14% of unique trigger-action pairs, 32% of rules) were used by more than one person, and most of these were used more than once by some participants.

Cloud Storage Sharing Settings. Participants frequently had applets with triggers or actions that access or modify cloud storage files. For example, 51 rules updated Google Sheets or Google Drive and 17 updated Dropbox. Other cloud storage tools included Evernote and Day One. In our survey, we automatically identified 41 rules involving Google Sheets, Google Drive, or Dropbox, and asked the participants about the document's or cloud storage space's sharing settings. Participants unanimously stated that the cloud storage space or document was not shared with *anyone* else. In Section 4.4.1, we discuss the implication of these cloud storage files being private on our assessment of the potential risks of these rules.

4.3 Information-Flow-Based Analysis

Table 2 shows the breakdown of rules and unique trigger-action pairs that were found to be violating using the secrecy and integrity labels from prior work [35], and the new labels and lattice we generated for this study. Exactly replicating the analysis from previous work to evaluate the 158 fully-labelled rules, we found that about 35% of rules were potentially violating, compared to about 50% in previous work [35]. Using the updated lattice and labels (see Section 3.1), applied to all 743 rules, we found that 57% of rules were potentially violating. In total, 319 rules have potential integrity violations and 269 rules have potential secrecy violations.

	Total rules	Potentially violating	Secrecy-violating	Integrity-violating
2017 lattice	158	55 (34.8%)	41 (25.9%)	48 (30.4%)
Updated lattice	743	423 (56.9%)	269 (36.2%)	319 (42.9%)

Table 2: Secrecy and integrity violations reported by the information-flow-based analysis.

However, a key research question remains: are the rules labelled as violating actually something to worry about and are the remaining rules actually innocuous?

4.4 Secrecy and Integrity Risks in Context

In this section, we assess the riskiness in practice of the rules deemed potentially violating by information-flow-based analysis. We use semantic labels (described in Section 3.1.2) to structure this analysis and leverage contextual information such as applets' titles to better understand specific rules.

4.4.1 Evaluating Secrecy-Violating Rules

Out of the 269 rules with potential secrecy violations, almost one third are unlikely to actually carry sensitive information, and 42% (some of which overlap with that third) have actions that are probably not observable by unintended people. We judge these rules to be unlikely to be harmful based on the semantic labels assigned to either their triggers or their actions. We next describe why rules with specific semantic labels are unlikely to be harmful. In total, only around 10% of secrecy violating rules (a subset of those with actions that have semantic labels of *Outgoing communication* or *Other automations*) are likely to lead to significant secrecy risks.

No Secret Information. 47 rules of the 164 rules with *Voice command* triggers and 24 of the rules with *Intentional triggers* (totaling 27% of rules with potential secrecy violations) have potential secrecy violations according to information-flow-based analysis. This is because the *Voice command* or *Intentional trigger* could be done privately, and they result in an action with wider observability. For example, 10 of these rules control *Lights* (as their action), which might be observed by neighbors outside of the user's home, potentially revealing when a person is home or which rooms they are using. These rules act as an alternative light switch and do not introduce additional secrecy risks beyond those of a normal light switch. Also, because the user actively decides to cause this action each time the rule executes, they can evaluate at time-of-use whether the potential secrecy leak is a problem.

Twenty one *News-ish* rules were found to have potential secrecy violations (8% of secrecy-violating rules). Although many *News-ish* triggers have an secrecy label of *public* and

can, therefore, not lead to secrecy violations, these 21 rules have triggers that require additional context to determine whether they utilize public or restricted information. For example, the trigger *"Twitter :: New tweet by a specific user"* should have a secrecy label of *restricted online* if the specific user's Twitter account is set to private and *public* if not. P20's rule *"Save every tweet from the US President"* triggers based on public tweets and is therefore not secrecy-violating. We manually evaluated the titles for these 21 secrecy-violating *News-ish* rules, which were not considered in the analysis from previous work or the automated information-flow-based analysis; based on their titles, we determined that 13 are not actually violating (4.9% of all secrecy-violating rules).

Action Remains Private. 113 rules (42% of secrecy-violating rules) have an action with the semantic label *Log or notify*. This includes rules that update Google spreadsheets (45 rules) or Dropbox files or folders (17 rules), add calendar events (15 rules), or send the user a notification (24 rules). It is possible that these actions could leak information to untrusted parties who have access to the documents, folders, or calendars, or can observe when the user receives a notification. Based on participants' survey responses, we know that most of participants' applet-connected Google Drive and Dropbox files were not shared with other people. Hence, rules connected to private cloud storage spaces are not actually secrecy-violating (based on how cloud storage content is shared) although they could be (if the cloud storage content was widely shared), and hence were detected as potentially violating secrecy by the information-flow-based analysis.

Notifications have a *restricted physical* secrecy label, because there are many situations in which someone else might see a user's phone screen (e.g., if their phone is being used for navigation in a car or to play music at a party). Rules that have actions that send a notification could potentially leak information if their action is private. Although this is a real risk, most users are routinely exposed to this risk, even without using IFTTT—many smart phone apps such as email and SMS have (by default) the side-effect of showing a notification, often with a message preview. Thus, rules that have an action that sends a notification directly are no more risky than those that send the user an email or an SMS.

Secrecy Risk Is Limited by the Expressivity of the Action. For 76 rules (28% of secrecy-violating rules), if they leaked sensitive information, they would do so by *Changing the State of a Personal Device* or *Changing the state of an IoT device* (e.g., 21 secrecy-violating rules control *Lights* and 4 change the user's phone volume). The extent to which these rules could leak private information is limited by the expressivity of these devices—many have only an "on" or "off" option. Additionally, there are a plethora of *other* ways the action could occur (e.g., triggered by another rule, controlled directly from the device's dedicated app, or through physical interaction with the device). Private information *could* be leaked via these rules, but the risk is typically low.

Sending "Outgoing" Communication. 37 potentially secrecy-violating rules (14%) share information via *Outgoing communication* (i.e., social media, SMS or email), which could leak sensitive or private information to other people. As we discuss again in terms of its implications for rules with potential integrity violations, participants regularly use *Outgoing communication* actions to send information to themselves. For example, P20 has a rule called *"Receive an email diagnosis from Dash if your car experiences an issue"* that uses the action *"Gmail :: Send an email"* (rather than the action *"Email :: Send me an email"*). In several cases, even when the rule's title does not specifically state that the outgoing content is sent to the user themself, we can infer that this is the case. For example, 14 of these 37 rules belong to P28 and post to Slack based on *Sensing IoT device states* (e.g.,*"If basement Sliding Door closed then post a message to a Slack service"*); these rules probably post to a private Slack channel.

There are only 14 secrecy-violating rules (5%) which are likely to *actually* send *Outgoing communication*, which could be risky. For example, because of the rule *"Tumblr Likes to Pinterest,"* P9 could accidentally "like" a post on Tumblr that would be embarrassing if it was sent to his Pinterest followers. These secrecy risks exist even for rules with triggers and actions that *both* use the *same* channel, or service. For example, P18's rule *"Save Facebook photos you're tagged in to your own album"* could result in unflattering photos of him being added to his album, possibly with a broader audience on Facebook than the original post.

Other Secrecy-Violating Rules. Of all 269 rules found to potentially violate secrecy, the previous discussion has addressed all but 11 rules (4%). These 11 rules all have actions that are *Other [non-IFTTT] automations*. In particular, one rule has both a trigger and an action that are *Other automations* (*"If maker Event 'mancave sleep' then run a Stringify Flow"* by P13). Without additional details about the automations, we cannot evaluate the potential harms of these rules. It is probably pertinent to warn users who install such applets that no automated analysis could evaluate secrecy properties, which could make these rules especially risky.

4.4.2 Evaluating Integrity-Violating Rules

Although 319 rules have potential integrity violations, according to information-flow analysis, very few of them are actually likely to lead to integrity-related harms. 64% of the integrity-violating rules do nothing more than update a digital log or notify the user, which is unlikely to be harmful even if it is caused by an adversary. The rules that are most likely to be potentially risky include 40 rules with *Other automation* actions and 21 rules that would potentially allow an adversary to control smart home devices other than lights—totaling just under 20% of the integrity-violating rules.

Trigger Is Sufficiently Trusted. Many rules have triggers with *trusted other* integrity labels and actions labeled *trusted*.

While such rules are flagged as integrity violations, we found that rules with *trusted other* triggers would not typically create integrity risks. For example, the trigger *"Best Buy :: product price changes"* is controlled by a company (Best Buy) that is unlikely to change the price of a product with the goal of adversarially triggering someone's applet. Out of 319 rules that have potentially integrity violations, 71 rules (23%) have similar triggers to the one discussed above, with *trusted other* integrity. These predominantly have the semantic labels *Weather or time* (30 rules) or *News-ish* (29 rules). The triggers *"Nest Protect :: Battery is low"* and *"Fitbit :: Daily activity summary,"* which is sent at the same time each day regardless of the activity summary's contents, account for the remaining 12 rules with other semantic labels.

An additional 31 integrity-violating rules (10%) have an action with the semantic label *News-ish*, but not *trusted other* integrity. Based on manual examination of the applet titles, we determined that for 26 of these 31 rules the trigger has *trusted other* integrity in practice (8% of integrity-violating rules). For example, a new item in an RSS feed *could* come from an *untrusted* source; however, in P20's rule *"Text me if the CDC reports a zombie outbreak,"* if the update comes from the United States Centers for Disease Control (CDC), as suggested by the title, then *trusted other* would be a more appropriate integrity label for this trigger.

Creating a Log or Notifying the User. The majority of integrity-violating rules (64%; 204 rules) have an action with the semantic label *Log or notify*. Some of these violations could sometimes lead to practical harm. For example, as noted in prior work, an adversary could potentially fill up cloud storage space [35], or an ill-timed notification could disrupt an important meeting. However, their titles reveal that many of these applets *intentionally* trigger based on other people's actions. For example, in P3's rule *"If office Nest Protect battery is low, then send a notification,"* other people with physical access to the home *could* cause this rule to execute (e.g., by repeatedly touching the device to keep the screen on and drain the battery more quickly). The user would likely still want this warning so that they know to replace the battery.

Sending "Outgoing" Communication. 27 rules (8% of integrity-violating rules) have an action that sends *Outgoing communication*. Someone who controls the trigger could, for example, spam the user's friends with emails or create unwanted social media posts, if that is what the applets were set up to do. In practice, however, as previously discussed regarding secrecy-violating rules, many of these do not actually send *outgoing* messages. The titles of these 27 rules reveal that almost all of them likely send information only to the IFTTT user. For example, P10's rule *"Have Alexa email you your shopping list"* probably emails P10, despite using the action *"Gmail :: send an email"* (*Outgoing communication*) rather than *"Email :: send me an email"* (*Log or notify*).

Controlling Smart-Home Devices. 37 rules with potential integrity violations (12%) have an action that *Changes an IoT device's state*, including 16 that control *Lights* and 9 that control devices related to *Home security*. If these rules were triggered maliciously, the extent and type of harm, is predominantly determined by the capabilities of the device they control and other contextual factors. An additional 41 integrity-violating rules (13%) have an action that causes a non-IFTTT automation to execute. For 21 of these 41 rules, the titles suggest that they change the state of home IoT devices (e.g., P27's rule *"If You say 'Set Sonos to 10 percent' then run a Stringify Flow"*).

Rules that control lights, could be harmful if they are used at inopportune times—e.g., lights coming on at 2am causing someone to lose sleep, lights turning off while someone is walking down stairs, potentially causing an injury. A more likely risk is that lights are left on more than they normally would be, consuming electricity or causing the device to wear out more quickly. This creates a financial risk bounded by the cost of the device plus the cost of the light being on constantly. Other types of smart-home devices might have a greater potential to cause expensive and/or dangerous damage. For example, P7 and P12 have rules that mention turning on a waffle iron and "cat heaters," respectively, which could potentially start a fire. P9's rules that turn on an irrigation systems cause costly damage if used during freezing weather.

Users may be especially protective of rules that affect their home security (40 total rules, 9 of which have potential integrity violations). For example, P27's rule *"If You say 'Disarm Blink' [to Google Assistant] then disarm Outside Blink system"* is a rule that might warrant a warning to the user; an unintended person could potentially speak loudly enough from outside the home to disarm the system. However, many users might still decide that this risk is acceptable or sufficiently unlikely given the placement of their smart assistant.

Controlling the User's Personal Device. 10 integrity-violating rules (3%) *Change [the user's] personal device state* (e.g., changing the volume level or launching an app like music or navigation). If properly timed, an adversary could cause this rule to execute during an important meeting, causing embarrassment or punishment. However, similar risks exist any time a user's phone is on, unrelated to IFTTT (e.g., repeatedly calling during an important meeting). Alternately, an adversary could cause these rules to execute with the goal of draining the user's phone battery more quickly than usual, which could be dangerous in some contexts (e.g., if the user is in an unfamiliar place and will need directions or a ride home). Launching certain apps could utilize cell data, which might be limited or expensive for the user.

Other Integrity-Violating Rules. 40 integrity-violating rules (13%) have not yet been addressed. All of these rules have actions that are *Other [non-IFTTT] automations*. As with secrecy-violating rules that have *Other automation* actions, we cannot evaluate the possible integrity harms that could be associated with these rules. Therefore, these rules should be treated with caution.

Semantic trigger label	# (%) of rules	# potentially violating rules (secrecy / integrity)
Weather or time	101 (13.6%)	30 (0 / 30)
News-ish	121 (16.3%)	60 (21 / 60)
Sensing IoT device state	111 (14.9%)	68 (45 / 57)
Environment sensing	46 (6.2%)	40 (7 / 37)
Intentional trigger	38 (5.1%)	24 (24 / 0)
Voice command	164 (22.1%)	72 (47 / 72)
Incoming communication	19 (2.6%)	15 (17 / 14)
Sensing online account state	57 (7.7%)	49 (46 / 20)
Actions with personal devices	58 (7.5%)	33 (33 / 0)
Other automations	30 (4.0%)	29 (29 / 29)

Table 3: The distribution of rules with each of five semantic trigger labels, and the breakdown of rules with each label with secrecy or integrity violations.

Semantic action label	# (%) of rules	# potentially violating rules (secrecy / integrity)
Change IoT device state	254 (34.2%)	79 (72 / 37)
*Home security**	40 (5.4%)	17 (14 / 9)
*Lights**	66 (8.9%)	25 (21 / 16)
Log or notify	368 (49.5%)	246 (113 / 204)
Change personal device state	28 (3.8%)	11 (4 / 10)
Outgoing communication	41 (5.5%)	39 (37 / 27)
Other automations	52 (7.0%)	49 (43 / 41)

Table 4: The distribution of rules with each of five semantic action labels, and the breakdown of rules with each label with secrecy or integrity violations. Labels denoted with an asterisk (*) are secondary labels for *Change IoT device state*.

4.5 Survey Responses

We now consider participants' responses to survey questions, to help further contextualize our findings about their applets. Are participants' assessments of and experiences with their own applets consistent with our finding that most applets are unlikely to lead to harm due to secrecy or integrity violations? What harms have participants encountered, including but not limited to those that arise because of secrecy or integrity violations?

4.5.1 Choosing Applets

Prior work hypothesized that applets shared publicly on IFTTT are a potential attack vector [8]. Most participants (16, 57%) reported a preference to create their own applets. 25 participants (89%) reported creating *some* applets themselves, and seven participants had created 20 or more applets. Thus, although preventing malicious applets from being available on the IFTTT webpage can mitigate some security and privacy risks, it is also important to be able to identify potential risks in applets created by users themselves.

In addition to using existing applets shared on IFTTT, 9 participants (32%) said that they sometimes or often turn on applets based on friends' or colleagues' recommendations. When taking suggestions from trusted sources, users might be less likely to consider the potential harms.

Would you be upset if this applet contributed to the following situation occurring:	Very upset	Slightly upset	Not upset
TOTAL	105 (41.7%)	**114 (45.2%)**	33 (13.1%)
You no longer directly control what files are downloaded from email or social media, possibly spreading malware on your computer?	**32 (69.6%)**	9 (19.6%)	5 (10.9%)
Your electronic device is used in a way it wasn't designed for (such as being toggled on/off very rapidly), possibly reducing its longevity or damaging it? *	**21 (55.3%)**	16 (42.1%)	1 (2.6%)
Private information gets posted online unintentionally, possibly embarrassing you?	**28 (52.8%)**	17 (32.1%)	8 (15.1%)
Data gets uploaded to your cloud storage more often than you thought, possibly causing you to run out of space?	14 (26.9%)	**26 (50%)**	12 (23.1%)
You consume more resources (e.g., electricity, phone data, cloud storage space), possibly increasing your bills or otherwise causing you to spend more money?	10 (15.9%)	**46 (73%)**	7 (11.1%)

Table 5: Participants typically stated that they would be very or slightly upset if an applet contributed to a harmful situation. The asterisk (*) denotes that this question applied only to rules that utilize a physical device in their trigger or action.

4.5.2 Participants Believe Their Applets Are Safe

In general, participants did not express concerns about security and privacy risks arising from *their own* use of IFTTT, though they seemed to be aware of the possibility that applets could lead to security and privacy risks. This is consistent with our assessment of their applets. Almost all participants (96%) believed that their applets work as expected and are safe to use. Only six participants (21%) changed their views on the riskiness of applets, even after we explained the definition of secrecy and integrity violations and how this might manifest in applets (at the end of the survey). All six reported increased caution about their applets' secrecy; P16 also reported increased concern about integrity violations.

4.5.3 Harms Experienced from IFTTT Rules

Despite an overall sense that their applets are safe, four participants noted that they *had* experienced harms that they attributed to their applets or expressed that their applets sometimes did not work as expected. The harms they described were not the result of secrecy or integrity violations. Rather, they described instances in which the app or service malfunctioned or in which they had misconfigured a rule. P10's door was unlocked when location updates functioned in an unexpected way: *"... I was in Disneyland and had just turned on my phone's location settings, so when the IFTTT app received the first location broadcast, it was before GPS and cell location could be locked in and so it assumed I was still at home."* P26 explained that she *"used to use an applet that posted*

my door lock to a google calendar. That got annoying." P23 recalled an applet that accidentally flooded his Twitter with posts: *"[it] was only supposed to trigger in some situations, I set it up wrong, realized it was posting too often."* Illustrating the complexity that can exist in a set of rules, P20 reported that one of his applets created an undesired Facebook post, though he was *"unsure of which applet did it."* In Section 5, we discuss other types of harm, including but not limited to the ones these participants experienced.

Selection bias may have contributed to our finding that so few participants experienced harms due to their applets. Users who fear or experience harms due to their applets may be less likely to participate in online message boards about home automation, where we reached potential participants.

4.5.4 Participants Value Applets' Security and Privacy

Although they mostly believed their applets were safe and did not change their level of caution based on explanations of potential secrecy and integrity violations, participants conveyed that the security and privacy of their applets *is* important to them. 21 participants (75%) said they would be upset if an applet triggered when they did not intend it to, which could happen through an integrity violation (or misconfiguration or incorrect behavior of the rule or connected services).

When asked about whether they would be upset if a specified, potentially-violating applet contributed to one of five undesirable outcomes, they only reported that they would not be upset for a total of 13% of applets (see Table 5). Comparing across the different types of harmful outcomes, they were less concerned about the possibility of using up cloud storage space and consuming extra resources than they were about an applet possibly posting private information that would embarrass them, spreading malware on their computer, or damaging their physical smart-home devices.

Many participants were uncomfortable with certain other people knowing which applets they have, especially strangers (i.e., 46% of participants disagreed that "[they are] comfortable with anyone knowing what applets [they] use," as shown in Table 6). strangers having this information (see Table 6). In Section 5, we re-examine the implications of this finding.

Survey Prompt	Strongly or somewhat agree	Neither agree nor disagree	Strongly or somewhat disagree
I am comfortable telling my **friends** what applets I use.	**24 (85.7%)**	3 (10.7%)	1 (3.6%)
I am comfortable telling my **colleagues** what applets I use.	**23 (82.1%)**	4 (14.3%)	1 (3.6%)
I am comfortable with **anyone** knowing what applets I use.	11 (39.3%)	4 (14.3%	**13 (46.5%)**

Table 6: Participants were more likely to report that they would be comfortable with friends or colleagues knowing about their IFTTT rules than strangers. The answer choice with the most responses is shown in bold.

5 Discussion

In contrast with the results of the automated analysis (Section 4.3) and the findings of previous work [35], our analysis of real users' applets suggests that the majority of applets are unlikely to lead to significant risks due to secrecy or integrity violations. Nevertheless, our finding that participants *are* concerned about the security and privacy of their applets emphasizes the importance of identifying potentially harmful applets more accurately.

With this in mind, we next consider what *additional* harms applets could introduce that may not be encompassed by the previous focus on secrecy or integrity risks, as they were defined in the automated analysis that we leveraged. We additionally re-surface key take-aways from Section 4, addressing limitations of the existing automated analysis that lead to non-risky applets being labeled as potentially violating. Finally, we conclude with guidelines for future tools that seeks to more accurately identify risky (or non-risky) applets.

5.1 Risks Beyond Secrecy and Integrity Violations Against the IFTTT User

Mismatch Between Reality and Expectations When Setting up an Applet. Several participants described scenarios in which their experiences using IFTTT applets did not match their expectations. This could occur for many reasons, which do not necessarily involve secrecy or integrity violations or adversarial actions: IFTTT services might not work as expected (e.g., not sensing location appropriately when devices connect and disconnect from the Internet), users might unintentionally misconfigure their applets (e.g., mis-spelling an applet parameter such as a search term), or users might not anticipate the impact of an applet even when it is configured as intended (e.g., P26 in retrospect decided that adding a calendar entry every time her door locked was "annoying").

Surveillance Risks to Incidental Users. Many rules cause data to be collected about people *other* than the IFTTT user, possibly without their awareness. We refer to these other people as *incidental users*. For example, P11's rule *"If [name] presence detected, then create Journal entry"* in effect monitors [name]'s location and schedule. Furthermore, all IFTTT rules create a record, accessible after the fact, whenever they execute. Hence, P9's rule *"If - Front Door locked then switch off Entryway Light"* creates a record, through IFTTT, of each time the front door is locked. P9 could use this log to determine exactly when his family members arrive or leave.

Rules especially likely to trigger based on the actions of incidental users include those that *Sense IoT device states* (e.g., when a door is locked) or *Sense changes to the environment* (e.g., if the temperature increases when someone is home), as

well as those that trigger based on *Incoming communication*. Collectively, these make up 162 rules, or 22% of participants' rules, and include 53 rules that are non-violating according to information-flow analysis. Additionally, some rules with other semantic trigger labels might also present risks to incidental users. For example, several of P2's rules toggle a SmartThings Switch when three distinct people's phones connect to or disconnect from his WiFi (semantic trigger label *Actions with personal devices*). One of the phone owners is referred to as a "[baby/pet]sitter" in the applet title.

Incidental users can be family members, household visitors, employees, or neighbors. For example, P28's "Motion alert!" rule (which sends a notification when motion is detected) could inadvertently capture information about neighbors' daily schedules if it reacts to motion on a communal sidewalk. People other than household members might be especially unlikely to realize that they are being monitored via these rules. We have no reason to believe that any participants in our study used IFTTT in abusive ways to intentionally collect data about incidental users, but risks and harms, such as chilling effects [31], can exist even when surveillance is not intended adversarially. When such data *is* collected with malicious intent, as it might be in an abusive relationship, this type of surveillance could be especially harmful.

Our study was not designed to explore research questions related to incidental users and, thus, our conclusions on this topic are limited. Instead, our findings advocate for research that more thoroughly explores the privacy preferences and experiences of incidental users. Future work in this area would benefit from data collection of a larger scale and broader scope, including input from a variety of stakeholders other than people who use IFTTT or own smart home devices.

5.2 Limitations of Current Information-Flow-Based Automated Analysis

Trustworthiness of Information. Rules that convey trusted information could be used to trick users into trusting attacker-supplied information, particularly when this information is conveyed to the user over a less trusted channel. Several users, for example, used rules triggered by official CDC or weather information updates whose actions propagated this information to the user via email or another similar channel. In these cases, the source of the information is trusted, but the delivery channel used by the rule is not. An attacker could easily create an email message indistinguishable from one created by the rule, thereby tricking a user who set up the rule into believing that a trusted source supplied the information (and hence caused the email to be sent).

Such risks are not discovered by previous information-flow analyses: these would flag as a violation a flow from a lower integrity trigger to a higher integrity action; here, in contrast, the potential danger comes from the *user* treating an action as if it was as trustworthy as the trigger. Although this risk could apply to any rule with a somewhat trusted trigger, *Log or notify* (193 rules) are explicitly designed to convey information and may be particularly susceptible.

Reconsidering What Is "Secret." The automated information-flow-based analysis we use assumes that IFTTT rules adopted by users are public (i.e., known to the attacker). Similarly, it assumes that secret information would flow from the trigger to the action. However, viewing real users' rules has led us to challenge these assumptions. What risks are substantially mitigated, or introduced, if we assume that adversaries do not know a rule and its configuration? Could adversaries infer sufficient details about rules to take advantage of otherwise unlikely risks?

Many potential secrecy and integrity violations in participants' rules could only be exploited if the adversary knew the rule. For example, P27 had the rule *"If You say 'Disarm Blink' then disarm Outside Blink [alarm] system."* Without knowing the phrase to use to disarm the alarm system, an adversary may technically be able to disarm the alarm system but would have difficulty doing so. Similarly, without knowing about P11's rule *"If daily Steps goal not achieved by 10:15 pm, then send me an SMS,"* an adversary who sees (or hears) that P11 receives an SMS at 10:15pm would be unlikely to guess the *meaning* of the SMS (i.e., that P11 had been less active that day) without knowing that he was using this rule.

There are many potential ways that an adversary could infer details about the configuration of a rule (e.g., overhearing P27 use the smart assistant phrase to disarm their alarm system), and there are a variety of aspects of the rule that a user might care about keeping secret. In some cases, the secret or sensitive information may be entirely contained in the rule and only *implied* by the triggers or actions. For example, in the case of P11's rule *"If new [Craigslist] post from search Search URL then send me an SMS at [number],"* the user may wish to protect the secrecy of the specific search term they are following. The trigger is not secret, since Craigslist posts are publicly available; observations of the action could reveal this potentially sensitive detail about the rule.

By assuming that the IFTTT rule was public, the information-flow-based analysis on the one hand failed to identify rules that potentially leak sensitive information regarding the rule itself and on the other hand overestimated the probability of secrecy or integrity violations of other rules. Modeling the components of IFTTT rules on a more fine-grained level (e.g., specifying a secrecy and integrity label for the rule's parameters or the rule itself) could potentially address these limitations.

Rethinking the Granularity of Labels. In participants' applets, we found triggers and actions whose labels were both too fine grained and too coarse grained. For example, triggers based on official weather reports or the time of day were labelled as *trusted other*. This led to innocuous applets such as P17's *"Get the weather forecast every day at 7:00 AM"* being marked as potentially violating. In practice, distinguishing

between *trusted* and *trusted other* was unnecessary.

On the other hand, for some rules that have the same secrecy or integrity label for both their trigger and action—which causes information-flow analysis to judge them as safe—contextual details are needed to determine whether the people who can observe or control the trigger are in fact the same as those who can observe or control the action. For example, in P2's rule *"If kitchen Lights switched off then turn off lights in Kitchen"* both the trigger and the action have *restricted physical* secrecy, so this rule is considered non-secrecy violating. Based on the title, we can infer that this interpretation is correct—the trigger and action occur in *same* physical space (the kitchen). A similar rule that triggers when bedroom lights are switched off would also be marked as non-secrecy violating; however, that rule might reveal to people in the kitchen when someone enters the bedroom, and hence does constitute a secrecy violation (albeit perhaps not a particularly harmful one). 224 (125) such non-violating rules have triggers and actions with the same secrecy (integrity) label (*restricted online*, *restricted physical*, *trusted online*, or *trusted physical*).

In order to more accurately identify when rules are actually violating, choices of secrecy and integrity labels should be better-informed by a deeper understanding of contextual factors (e.g., devices' relative locations).

Challenges in Labelling External Services. Many services that access or update online content include attribution: they capture *who* added or edited content. The information-flow analysis specifies only that integrity labels describe "who could cause the event" [35]. We found that determining an integrity label for this type of action requires both a nuanced definition of "the event" and a deep understanding of (e.g., Google Calendar's) functionality. For example, the integrity label for the action *"Google Calendar :: Quick add event"* should denote "who could add an event to this calendar" (*definition 1*). However, Google Calendar events include a "created by" field, specifying the user (account) that created the event. Since the IFTTT rule creates a calendar event that shows up as "created by [username]" (where [username] is the IFTTT user's Google account name), an alternative integrity label would denote "who could add a calendar event to this calendar that appears to be created by [username]" (*definition 2*).

Although definition 1 is consistent with the action's description in IFTTT, using it for information-flow analysis will not capture erroneous attribution of the calendar entry. If someone who can add events to a shared calendar (from their own user account) instead uses, e.g., Avery's rule to create a calendar event, they could add events with embarrassing or offensive titles that other people would attribute to Avery.

5.3 Guidelines for Automatically Identifying Risky Applets

Based on our findings, we compiled guidelines to help more accurately identify potentially risky trigger-action rules:

- Be aware of gaps between users' intent and installed rules (Section 5.1).
- Consider risks to incidental users, and consider a variety of potential adversaries such as abusive partners, which might include the IFTTT user themselves (Section 5.1).
- Analysis should not assume that other people (e.g., potential attackers) know the rule or its configuration (Section 5.2).
- Expect that appropriate secrecy and integrity labels are sensitive to contextual details that may be difficult to determine automatically, such as the settings of users' external services or the location of their physical devices (Section 5.2).
- For giving intuitive warnings to users, semantic labels may be more useful than fine-grained analyses, because they correlate with risky rules (e.g., about *Home security*) and can explain how a rule could lead to harm. Simple heuristics could effectively complement more theoretically grounded analyses (Section 4.4).

Future work should seek to incorporate these guidelines into automated analyses and to provide a deeper understanding of users' experiences. Our survey was designed to inquire about the types of harms posited in prior work; future work should seek to explore the other types of potential harm that we identified. For example, while we identified the possibility of IFTTT rules with surveillance risks to incidental users, we do not yet know if they recognize that these risks exist or what their security and privacy preferences regarding other peoples' rules are. Our participant sample was demographically skewed; future studies could assess whether different groups of IFTTT users might have rules with different potential risks. Since incorporating contextual details may be necessary to determine appropriate secrecy and integrity labels, future work might consider how this information could be obtained, either by asking users directly or through automated (technical) means.

6 Conclusion

We evaluated the possible risks and harms associated with real users' IFTTT applets. Applets were less risky than was previously shown through automated analysis that sought to identify secrecy and integrity violations; however, we discovered new types of potential harm not previously considered in that automated analysis. Additionally, we outline some of the ways that the automated analysis falls short even in its ability to accurately identify secrecy and integrity risks. Finally, we discuss guidelines for creating a better tool (future work) that would identify risky applets—both from the standpoint of more accurately identifying secrecy and integrity violations and in terms of identifying other types of risk.

Acknowledgments

This work was supported in part by gifts from Google and the CyLab Security and Privacy Institute at Carnegie Mellon University; by a CyLab Presidential Fellowship and a Symantec Research Lab fellowship; and by DARPA and the Air Force Research Laboratory under agreement number FA8750-15-2-0277.

References

[1] If This Then That (IFTTT). https://ifttt.com.

[2] Microsoft Flow. https://flow.microsoft.com/en-us/.

[3] Research: 90 https://www.iottechnews.com/news/2018/may/15/research-us-consumers-smart-home-device/.

[4] SmartThings. https://www.smartthings.com/.

[5] Zapier. https://zapier.com/.

[6] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than smart speakers: Security and privacy perceptions of smart home personal assistants. In *Proceedings of the 15th Symposium on Usable Privacy and Security (SOUPS)*, 2019.

[7] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. Discovering smart home internet of things privacy norms using contextual integrity. *Proceedings of ACM Interaction Mobile Wearable Ubiquitous Technology*, 2(2):59:1–59:23, 2018.

[8] Iulia Bastys, Musard Balliu, and Andrei Sabelfeld. If this then what? Controlling flows in IoT apps. In *Proceedings of the 25th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1102–1119, 2018.

[9] Suman S. Bhunia and Mohan Gurusamy. Dynamic attack detection and mitigation in IoT using sdn. In *Proceedings of the 27th IEEE International Telecommunication Networks and Applications Conference (ITNAC)*, pages 1–6, 2017.

[10] Z. Berkay Celik, Patrick McDaniel, and Gang Tan. Soteria: Automated IoT safety and security analysis. In *Proceedings of the 2018 USENIX Annual Technical Conference (USENIX ATC)*, pages 147–158, 2018.

[11] Z. Berkay Celik, Gang Tan, and Patrick McDaniel. IoTGuard: Dynamic enforcement of security and safety policy in commodity IoT. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS)*, 2019.

[12] Nicholas DeMarinis and Rodrigo Fonseca. Toward usable network traffic policies for IoT devices in consumer networks. In *Proceedings of the 2017 Workshop on Internet of Things Security and Privacy (IoTS&P)*, pages 43–48, 2017.

[13] Soteris Demetriou, Nan Zhang, Yeonjoon Lee, XiaoFeng Wang, Carl Gunter, Xiaoyong Zhou, and Michael Grace. Hanguard: Sdn-driven protection of smart home WiFi devices from malicious mobile apps. In *Proceedings of the 10th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec)*, pages 122–133, 2017.

[14] Pardis Emami-Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Faith Cranor. Exploring how privacy and security factor into IoT device purchase behavior. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)*, 2019.

[15] Earlence Fernandes, Jaeyeon Jung, and Atul Prakash. Security analysis of emerging smart home applications. In *Proceedings of the 37th IEEE Symposium on Security and Privacy (SP)*, pages 636–654, 2016.

[16] Earlence Fernandes, Justin Paupore, Amir Rahmati, Daniel Simionato, Mauro Conti, and Atul Prakash. Flowfence: Practical data protection for emerging IoT application frameworks. In *Proceedings of the 25th USENIX Security Symposium (USENIX Security)*, pages 531–548, 2016.

[17] Earlence Fernandes, A. Rahmati, Jaeyeon Jung, and Atul Prakash. Security implications of permission models in smart-home application frameworks. *IEEE Security Privacy Magazine*, 15(2):24–30, 2017.

[18] Earlence Fernandes, Amir Rahmati, Jaeyeon Jung, and Atul Prakash. Decentralized action integrity for trigger-action IoT platforms. In *Proceedings of the 25th Annual Network and Distributed System Security Symposium (NDSS)*, 2018.

[19] Christine Geeng and Franziska Roesner. Who's in control?: Interactions in multi-user smart homes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)*, pages 268:1–268:13, 2019.

[20] Garrett M. Graff. Now a dorm room minecraft scam brought down the Internet. https://www.wired.com/story/mirai-botnet-minecraft-scam-brought-down-the-internet/, December 2017.

[21] Weijia He, Maximilian Golla, Roshni Padhi, Jordan Ofek, Markus Dürmuth, Earlence Fernandes, and Blase Ur. Rethinking access control and authentication for

the home internet of things (IoT). In *Proceedings of the 27th USENIX Security Symposium (USENIX Security)*, pages 255–272, 2018.

[22] Kai-Hsiang Hsu, Yu-Hsi Chiang, and Hsu-Chun Hsiao. Safechain: Securing trigger-action programming from attack chains. *IEEE Transactions on Information Forensics and Security*, 14(10):2607–2622, 2019.

[23] Yunhan Jack Jia, Qi Alfred Chen, Shiqi Wang, Amir Rahmati, Earlence Fernandes, Z. Morley Mao, and Atul Prakash. ContexIoT: Towards providing contextual integrity to appified IoT platforms. In *Proceedings of the 24th Annual Network and Distributed System Security Symposium (NDSS)*, 2017.

[24] Thorin Klosowski. Automation showdown: IFTTT vs Zapier vs Microsoft Flow. https://lifehacker.com/automation-showdown-ifttt-vs-zapier-vs-microsoft-flow-1782584748, 2016.

[25] Xianghang Mi, Feng Qian, Ying Zhang, and XiaoFeng Wang. An empirical characterization of IFTTT: Ecosystem, usage, and performance. In *Proceedings of the 2017 Internet Measurement Conference (IMC)*, pages 398–404, 2017.

[26] B. Montagu, B. C. Pierce, and R. Pollack. A theory of information-flow labels. In *IEEE 26th Computer Security Foundations Symposium (CSF)*, 2013.

[27] Andrew C. Myers and Barbara Liskov. Complete, safe information flow with decentralized labels. In *IEEE Symposium on Security and Privacy (Cat. No.98CB36186)*, 1998.

[28] Chandrakana Nandi and Michael D. Ernst. Automatic trigger generation for rule-based smart homes. In *Proceedings of the 2016 ACM Workshop on Programming Languages and Analysis for Security (PLAS)*, pages 97–102, 2016.

[29] Sukhvir Notra, Muhammad Siddiqi, Hassan Habibi Gharakheili, Vijay Sivaraman, and Roksana Boreli. An experimental study of security and privacy risks with emerging household appliances. In *Proceedings of the 2014 IEEE Conference on Communications and Network Security*, pages 79–84, 2014.

[30] Danny Palmer. Mirai botnet adds three new attacks to target IoT devices. https://www.zdnet.com/article/mirai-botnet-adds-three-new-attacks-to-target-iot-devices/, May 2018.

[31] Jonathon W. Penney. Whose speech is chilled by surveillance? https://slate.com/technology/2017/07/women-young-people-experience-the-chilling-effects-of-surveillance-at-higher-rates.html, July 2017.

[32] Roei Schuster, Vitaly Shmatikov, and Eran Tromer. Situational access control in the Internet of Things. In *Proceedings of the 25th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1056–1073, 2018.

[33] Anna Kornfeld Simpson, Franziska Roesner, and Tadayoshi Kohno. Securing vulnerable home IoT devices with an in-hub security manager. In *Proceedings of the 15th IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 551–556, 2017.

[34] Vijay Sivaraman, Hassan Habibi Gharakheili, Arun Vishwanath, Roksana Boreli, and Olivier Mehani. Network-level security and privacy control for smart-home IoT devices. In *Proceedings of the 11th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 163–167, 2015.

[35] Milijana Surbatovich, Jassim Aljuraidan, Lujo Bauer, Anupam Das, and Limin Jia. Some recipes can do more than spoil your appetite: Analyzing the security and privacy risks of IFTTT recipes. In *Proceedings of the 26th International World Wide Web Conference (WWW)*, 2017.

[36] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford. "I don't own the data": End user perceptions of smart home device data practices and risks. In *Proceedings of the 15th Symposium on Usable Privacy and Security (SOUPS)*, 2019.

[37] Yuan Tian, Nan Zhang, Yueh-Hsun Lin, XiaoFeng Wang, Blase Ur, Xianzheng Guo, and Patrick Tague. SmartAuth: User-centered authorization for the Internet of Things. In *Proceedings of the 26th USENIX Security Symposium (USENIX Security)*, pages 361–378, 2017.

[38] Qi Wang, Wajih U. Hassan, Adam Bates, and Carl Gunter. Fear and logging in the internet of things. In *Proceedings of the 25th Annual Network and Distributed System Security Symposium (NDSS)*, 2018.

[39] Yaxing Yao, Justin Reed Basdeo, Oriana Rosata Mcdonough, and Yang Wang. Privacy perceptions and designs of bystanders in smart homes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):59:1–59:24, November 2019.

[40] Tianlong Yu, Vyas Sekar, Srinivasan Seshan, Yuvraj Agarwal, and Chenren Xu. Handling a trillion (unfixable) flaws on a billion devices: Rethinking network

Rank	Channel (# of participants)
1	Notifications (21)
2	Email (18)
	Weather Underground (18)
4	Date and Time (16)
5	Google Sheets (15)
6	IFTTT (13)
	SmartThings (13)
8	Amazon Alexa (12)
	Location (12)
10	RSS Feed (11)

Table 7: The top ten channels in rank order by the number of participants using them.

Rank	Channel :: Trigger (# of participants)
1	Amazon Alexa :: Say a specific phrase (12)
	IFTTT :: New trigger or action published by service (12)
3	Date and Time :: Every day at __ (11)
	Weather Underground :: Tomorrow's forecast calls for __ (11)
5	Button Widget :: Button press (10)
	Google Assistant :: Say a simple phrase (10)
	RSS Feed :: New feed item (10)
8	Location :: You exit an area (9)
9	Location :: You enter an area (8)
10	Date and Time :: Every day of the week at __ (7)

Table 8: The top ten triggers in rank order by the number of participants using them.

security for the Internet-of-Things. In *Proceedings of the 14th ACM Workshop on Hot Topics in Networks (HotNets)*, pages 5:1–5:7, 2015.

[41] Eric Zeng and Franziska Roesner. Understanding and improving security and privacy in multi-user smart homes: A design exploration and in-home user study. In *Proceedings of the 28th USENIX Security Symposium (USENIX Security)*, pages 159–176, 2019.

A Appendix A: Frequently-Used Applet Components

Participants had a total of 743 IFTTT rules. Although there was substantial diversity in their rules, certain channels, triggers, actions, and trigger-action pairs were used by several participants. The top ten channels, triggers, actions, and trigger-action pairs (i.e., used by the largest number of participants) are shown in Tables 7 to 10.

B Appendix B: Descriptions of Information-Flow and Semantic Labels

Each trigger and action is labelled in three ways, based on its: integrity properties, secrecy properties, and semantic meaning.

Rank	Channel :: Action (# of participants)
1	Notifications :: Send a notification from the IFTTT app (2)
2	Email :: Send me an email (18)
3	Google Sheets :: Add row to spreadsheet (15)
	SmartThings :: Switch off (10)
5	Phone Call (US Only) :: Call my phone (9)
	SmartThings :: Switch on (9)
7	Google Calendar :: Quick add event (7)
	SMS :: Send me an SMS (7)
9	Android Device :: Set ringtone volume (6)
	Dropbox :: Add file from URL (6)
	Gmail :: Send me an email (6)

Table 9: The top eleven actions in rank order by the number of participants using them.

Rank	Trigger→Action Pair (# of participants)
1	IFTTT :: New trigger or action published by service
	→ Email :: Send me an email (12)
2	Amazon Alexa :: Say a specific phrase
	→ Phone Call (US Only) :: Call my phone (7)
	Weather Underground :: Tomorrow's forecast calls for __
	→ Notifications :: Send a notification from the IFTTT app (7)
4	RSS :: New feed item → Email :: Send me an email (6)
5	Space :: ISS passes over a specific location
	→ Notifications :: Send a notification from the IFTTT app (5)
	Weather Underground :: Today's weather report
	→ Notifications :: Send a notification from the IFTTT app (5)
7	Facebook :: You are tagged in a Facebook photo
	→ Dropbox :: Add file from URL (4)
	Fitbit :: New sleep logged → Google Sheets :: Add row to spreadsheet (4)
	Google Assistant :: Say a simple phrase → SmartThings :: Switch off (4)
	Google Assistant :: Say a simple phrase → SmartThings :: Switch on (4)
	IFTTT :: New applet published by service → Email :: Send me an email (4)

Table 10: The top eleven trigger-action pairs in rank order by the number of participants using them.

Rank	Trigger Label → Action Label (# of participants)
1	News-ish → Log or notify (19)
2	Weather or time → Log or notify (15)
3	Voice command → Log or notify (14)
4	Sensing IoT device state → Log or notify (11)
	Environment sensing → Log or notify (11)
	Sensing online account state → Log or notify (11)
	Voice command → Change IoT device state (11)
8	Intentional trigger → Log or notify (8)
9	Actions with personal devices → Log or notify (7)
	Weather or time → Change IoT device state (7)
	Sensing online account state → Outgoing communication (7)

Table 11: The top eleven most prominent combinations of semantic trigger label – semantic action label pairs in rank order by the number of participants with this type of applet.

Although the labels themselves are descriptive, we include more thorough explanations of each label in Tables 12–14.

Information-Flow Label	Who could cause the trigger or action event to occur (*integrity*) or know that it occurred (*secrecy*)?
Public (pub)	*Secrecy*: Anyone in the world. E.g., an action that posts a public Tweet.
Untrusted (unt)	*Integrity*: Anyone in the world. E.g., the trigger *"Android SMS :: Any new SMS received"* (anyone can *send* the user an SMS).
Untrusted group (unt_g)	*Integrity*: Anyone in the world, but requiring group coordination. E.g., the trigger *"reddit :: New top post in subreddit"* could not be caused by a *single* untrusted user, but the users who could potentially coordinate to cause the event to occur are unrestricted.
Restricted Physical (rp)	*Secrecy* and *Integrity:* People with physical access to a particular space or device. E.g., an action that turns off a smart light has restricted physical secrecy and integrity, because anyone who can see the light could know that the action has occurred, and anyone with physical access to the device can control it.
Restricted Online (ro)	*Secrecy* and *Integrity*: A restricted group of online users. E.g., an action that creates a Facebook post visible to only friends has restricted online secrecy.
Trusted Physical (tp)	*Secrecy* and *Integrity*: Similar to restricted physical, but within a more trusted group. E.g., the trigger *"Automatic Pro :: Entered an area"* can be caused by someone with physical access to the car and the ability to drive it (typically only trusted friends or family).
Trusted Online (to)	*Secrecy* and *Integrity*: Similar to restricted online, but within a more trusted group. E.g., household members explicitly given access to remotely control or monitor a particular smart home device.
Trusted Other (t_oth)	*Integrity*: Trusted sources such as weather or news reports.
Trusted (t)	*Integrity*: The IFTTT user.
Private (priv)	*Secrecy*: The IFTTT user.

Table 12: Descriptions of information-flow labels. Shorthand for each semantic label, used in Appendix C, is shown in parentheses.

Semantic Trigger Label	Description	# Distinct Triggers
Weather or time (WT)	Official weather (e.g., not the weather sensed by a home IoT device) or a pre-specified, likely recurring time or date.	13
News-ish (N)	Actual news sources like the New York Times, updates from official sources about new products or apps, personal news such as shipment status updates or new RSS Feed items, or new posts on websites like Craigslist or Reddit.	35
Sensing IoT device state (DS)	Based on the state of an IoT device that is meant to be controlled by a person, such as whether a door is open, closed, locked, or unlocked.	32
Environment sensing (E)	Based on the state of an IoT device that is meant to reflect to the ground truth state of the environment it is in (including detecting motion, temperature, CO2, etc.).	19
Intentional trigger (I)	Pressing a particular button in a phone app or widget, or sending an SMS or Voicemail to IFTTT (which only has the purpose of acting as a trigger).	4
Voice command (V)	Triggers that can be caused by verbal interaction with a smart assistant. Includes items being added to Alexa lists and alarms going off, as well as arbitrary phrases spoken to the voice assistant.	10
Incoming communication (IC)	The IFTTT user receives communication from others. For example, receiving an SMS or an email.	6
Sensing online account state (OAcc)	Triggers that react to online account updates. Includes, for example, being tagged in a photo on Facebook, having new "liked" videos on YouTube, or a new sleep being logged by FitBit	24
Actions with personal devices (P)	Includes mainly location sensing (e.g., moving through space with a physical device), but also sensing that the user has sent an arbitrary SMS from their device, took a screenshot, etc. Includes location sensing for the IFTTT user's device as well specified family members' devices. Includes location inference via devices coming within range of a specific area (e.g., connecting to home WiFi).	15
Other automations (OAu)	"Stringify flow runs" or "Webhooks, receive a web request."	2

Table 13: Descriptions of semantic labels of triggers. Shorthand for each semantic label, used in Appendix C, is shown in parentheses.

Semantic Action Label	Description	# Distinct Actions
Change IoT device state (DS)	Action alters or sets the state of an IoT device. E.g., turns lights on or off, locks or unlocks a door, turns on the thermostat.	59
Home security (DS:S)	A subset of "Change IoT device state" which affect IoT devices related to home security (e.g., door locks, surveillance cameras).	12
Lights (DS:L)	A subset of "Change IoT device state" which control lights.	20
Log or notify (L)	Creates a record of the trigger or notifies the user when the trigger occurs. Notifications can happen via notifications, emails, or phone calls where the action specifies that it is "to me." Logs can be saved to cloud services such as Google Sheets or in other personal accounts like making a "Journal Entry" in Day One or adding a calendar event to a Google calendar.	19
Change personal device state (P)	Changes the state of a personal device (e.g., phone), for example by launching an app (maps, music), changing the phone volume, or turning the WiFi on or off.	8
Outgoing communication (OC)	Sends information to other people, including by sending email or SMS, or updating social media accounts.	10
Other automations (OAu)	These actions act as triggers for a Stringify flow, Webhooks request, Wink shortcut, or Nexia automation.	4

Table 14: Descriptions of semantic labels of actions. Shorthand for each semantic label, used in Appendix C, is shown in parentheses.

C Appendix C: Trigger and Action Labels

We reached consensus in our labelling of triggers and actions used in participants' applets as described in Section 3. Tables 15 and 16 show all triggers and actions and their respective secrecy, integrity, and semantic labels.

Table 15: Semantic and information-flow labels of triggers used in participants' applets.

Trigger channel :: trigger (# of rules with this trigger)	Semantic label	Secrecy label	Integrity label
Amazon Alexa :: Ask whats on your Shopping List (4)	V	rp∪to	rp
Amazon Alexa :: Item added to your Shopping List (1)	V	rp∪to	rp∪to
Amazon Alexa :: Item added to your To Do List (4)	V	rp∪to	rp∪to
Amazon Alexa :: New song played (2)	V	rp∪to	rp∪to
Amazon Alexa :: Say a specific phrase (69)	V	rp∪to	rp
Amazon Alexa :: Your Alarm goes off (7)	V	rp∪to	rp∪to
Amazon Alexa :: Your Timer goes off (2)	V	rp∪to	rp
Ambient Weather :: Daily Rain rises above (1)	E	rp	rp
Android Device :: Connects or disconnects from any WiFi network (1)	P	rp	t
Android Device :: Connects to a specific WiFi network (1)	P	priv	t
Android Phone Call :: Any phone call answered (1)	IC	rp	t
Android SMS :: Any new SMS received (4)	IC	rp	unt
Android SMS :: Any new SMS sent (1)	P	rp	t
Android SMS :: New SMS received matches search (3)	IC	priv	unt
Apple App Store :: New app featured in a collection (1)	N	pub	t_oth
Apple App Store :: New app from search (1)	N	pub	unt | t_oth
Apple App Store :: Top ten app goes on sale (1)	N	pub	unt_g
AppZapp :: Top App gone free in Google Play (2)	N	pub	unt_g
AppZapp :: Top App gone free in the Apple App Store (3)	N	pub	unt_g
Arlo :: Motion detected (1)	E	to	rp
August :: Lock locked (2)	DS	rp∪to	rp∪to
August :: Lock unlocked (1)	DS	rp∪to	rp∪to
Automatic Classic :: Check engine light turned on (3)	DS	rp∪to	tp
Automatic Classic :: New trip completed (1)	DS	rp∪to	tp
Automatic Pro :: Check engine light turned on (2)	DS	rp∪to	tp
Automatic Pro :: Entered an area (1)	DS	rp∪to	tp
Automatic Pro :: Exited an area (1)	DS	rp∪to	tp
Automatic Pro :: Ignition turned off in area (4)	DS	rp∪to	tp
Automatic Pro :: Ignition turned on (2)	DS	rp∪to	tp
Automatic Pro :: Ignition turned on in area (2)	DS	rp∪to	tp
Automatic Pro :: New trip completed (2)	DS	rp∪to	tp
Best Buy :: New product in category (1)	N	pub	t_oth
Best Buy :: Product price changes (3)	N	pub	t_oth
Boxoh Package Tracking :: Any shipping status change (1)	N	priv	t_oth
Button widget :: Button press (29)	I	rp	t
Camera widget :: Any new photo (2)	P	rp	t
Classifieds :: New post from search (5)	N	pub	unt
Dash :: Check engine light turned on (3)	DS	rp∪to	tp
Dash :: New trip completed (1)	OAcc	rp∪to	tp
Date and Time :: Every day at (20)	WT	pub	t_oth
Date and Time :: Every day of the week at (10)	WT	pub	t_oth
Date and Time :: Every hour at (3)	WT	pub	t_oth
Date and Time :: Every month on the (2)	WT	pub	t_oth
Date and Time :: Every year on (1)	WT	pub	t_oth
Dominos :: Order out for delivery (2)	N	priv	t_oth
Dropbox :: New photo in your folder (2)	OAcc	ro | priv | pub	ro | t | unt
ecobee :: Thermostat enters Smart Home/Away (3)	DS	rp∪to	rp∪to
ecobee :: Thermostat indoor temperature is greater than (1)	E	rp∪to	rp∪to
ecobee :: Thermostat indoor temperature is less than (2)	E	rp∪to	rp∪to
ecobee :: Thermostat outdoor temperature is less than (1)	E	pub	rp
ESPN :: New game start (1)	N	pub	t_oth
ESPN :: New in-game update (2)	N	pub	t_oth
eWeLink Smart Home :: 1 Channel Plug turned on or off (1)	DS	rp∪to	rp∪to
eWeLink Smart Home :: 1 Channel Switch turned on or off (3)	DS	rp∪to	rp∪to
Facebook :: New photo post by you (1)	OAcc	priv | pub | ro	t
Facebook :: You are tagged in a photo (10)	OAcc	priv | pub | ro	unt
Facebook :: Your profile changes (2)	OAcc	priv | pub | ro	t
Fitbit :: Daily activity summary (3)	OAcc	pub	t_oth
Fitbit :: Daily goal not achieved by __:__ (1)	OAcc	priv	t
Fitbit :: New sleep logged (8)	OAcc	priv	t
Fitbit :: New weight logged (1)	OAcc	priv	t
Gmail :: New email in inbox from (2)	IC	priv | ro	unt
Gmail :: New email in inbox from search (5)	IC	priv | ro	unt
Gmail :: New email in inbox labeled (4)	IC	priv | ro	unt
Google Assistant :: Say a phrase with a number (1)	V	rp∪to	rp
Google Assistant :: Say a phrase with a text ingredient (5)	V	rp∪to	rp
Google Assistant :: Say a simple phrase (69)	V	rp∪to	rp
Google Calendar :: Any event starts (2)	OAcc	ro | to | priv | pub	t_oth | t | ro | unt
Google Calendar :: Event from search starts (1)	OAcc	ro | to | priv | pub	t_oth | t | ro | unt

Google Calendar :: New event added (2)	OAcc	ro | to | priv | pub	t_oth | t | ro | unt
Google Wifi :: Device Connects (3)	P	tp	tp
Google Wifi :: Device Disconnects (5)	P	tp	tp
HomeSeer :: A device is turned off (1)	DS	rp∪to	rp∪to
HomeSeer :: A device is turned on (3)	DS	rp∪to	rp∪to
IFTTT :: Daily recommended Applet for you (3)	N	pub	t_oth
IFTTT :: New Applet published by service (9)	N	pub	t_oth
IFTTT :: New IFTTT update (6)	N	pub	t_oth
IFTTT :: New trigger or action published by service (23)	N	pub	t_oth
Instagram :: Any new photo by you (1)	OAcc	ro | pub	t
iOS Calendar :: New event added to specific calendar (1)	OAcc	ro | to | priv | pub	t_oth | t | ro | unt
iOS Contacts :: Any new contact (9)	OAcc	priv	t
iOS Photos :: Any new photo (1)	P	rp | to | priv	t
iOS Photos :: New photo added to album (1)	OAcc	rp | ro | priv	t
iOS Photos :: New screenshot (2)	P	rp | priv	t
Leeo :: Temperature below threshold (1)	E	rp∪to	rp∪to
Life360 :: First family member arrives at a specific place (2)	P	rp∪to	to
Life360 :: Last family member leaves a specific place (3)	P	rp∪to	to
Location :: You enter an area (12)	P	rp	t
Location :: You enter or exit an area (5)	P	rp	t
Location :: You exit an area (12)	P	rp	t
Manything :: Motion detected (2)	E	priv	rp
MyQ :: Door closed (1)	DS	rp	rp
MyQ :: Door opened (1)	DS	rp	rp
Nest Protect :: Battery is low (10)	DS	rp∪to	t_oth
Nest Protect :: Carbon monoxide emergency (3)	E	rp∪to	rp
Nest Protect :: Carbon monoxide warning (8)	E	rp∪to	rp
Nest Protect :: Smoke alarm warning (10)	E	rp∪to	rp
Nest Thermostat :: Nest set to Away (3)	DS	rp∪to	rp∪to
Nest Thermostat :: Nest set to Home (3)	DS	rp∪to	rp∪to
Nest Thermostat :: Temperature drops below (2)	E	rp∪to	rp∪to
Nest Thermostat :: Temperature rises above (2)	E	rp∪to	rp∪to
Netatmo Weather Station :: Temperature rises above (3)	E	pub	rp
NJ Transit :: New bus advisory (1)	N	pub	t_oth
NPR :: New story published (2)	N	pub	t_oth
OhmConnect :: An #OhmHour starts (2)	N	pub	t_oth
Phone Call (US only) :: Leave IFTTT any voicemail (3)	I	priv	t
ProPublica :: Congress is scheduled to vote on a bill (1)	N	pub	t_oth
ProPublica :: The president signs a new bill into law (5)	N	pub	t_oth
RainMachine :: Device is offline (1)	DS	rp∪to	rp∪to
reddit :: Any new post in subreddit (1)	N	pub | ro	unt | ro
reddit :: New hot post in subreddit (1)	N	pub | ro	unt_g | ro
reddit :: New post from search (1)	N	pub | ro	unt | ro
reddit :: New post saved by you (1)	OAcc	priv	t
reddit :: New top post in subreddit (1)	N	pub | ro	unt_g | ro
RSS Feed :: New feed item (14)	N	pub | priv	unt | t
RSS Feed :: New feed item matches (9)	N	pub | priv	unt | t
Slice :: Any new shipment (2)	N	priv	unt
Slice :: Shipment status changes (1)	N	priv | rp	unt
SmartThings :: Any new motion (1)	E	to	rp
SmartThings :: Closed (7)	DS	rp∪to	rp
SmartThings :: Locked (3)	DS	rp∪to	rp
SmartThings :: Moisture detected (1)	E	rp∪to	rp
SmartThings :: Opened (14)	DS	rp∪to	rp
SmartThings :: Presence detected (4)	P	to	tp
SmartThings :: Presence no longer detected (2)	P	to	tp
SmartThings :: Switched off (7)	DS	rp∪to	rp
SmartThings :: Switched on (17)	DS	rp∪to	rp
SmartThings :: Temperature drops below (2)	E	rp∪to	rp
SmartThings :: Temperature rises above (2)	E	rp∪to	rp
SmartThings :: Unlocked (3)	DS	rp∪to	rp
SMS :: Send IFTTT an SMS tagged (3)	I	priv	t
SMS :: Send IFTTT any SMS (3)	I	priv	t
Space :: Astronomy picture of the day by NASA (1)	N	pub	t_oth
Space :: Breaking news by NASA (1)	N	pub	t_oth
Space :: ISS passes over a specific location (6)	N	pub	t_oth
Spotify :: New track added to a playlist (1)	OAcc	pub | priv	t
Square :: Any new payment (1)	OAcc	rp	tp
Stringify :: Stringify Flow runs (7)	OAu	priv | rp | ro | pub	t | t_oth | rp | ro | unt_g | unt
The New York Times :: New article from search (1)	N	pub	t_oth
The New York Times :: New popular article in section (3)	N	pub	unt_g
TiVo :: SKIP segment detected (4)	DS	rp	rp
Todoist :: New completed task (2)	OAcc	priv	t
Tumblr :: Any new post (1)	OAcc	pub | ro	t
Tumblr :: New like (1)	OAcc	priv | ro | pub	t
Twitter :: New liked tweet by you (1)	OAcc	pub | ro	t
Twitter :: New tweet by a specific user (4)	N	pub | ro	t_oth
Weather Underground :: Todays weather report (4)	WT	pub	t_oth
Weather Underground :: Tomorrows forecast calls for (1)	WT	pub	t_oth
Weather Underground :: Tomorrows weather report (1)	WT	pub	t_oth
Weather Underground :: Current condition changes to (4)	WT	pub	t_oth

Weather Underground :: Current pollen count rises above (10)	WT	pub	t_oth
Weather Underground :: Current UV index rises above (30)	WT	pub	t_oth
Weather Underground :: Sunrise (14)	WT	pub	t_oth
Weather Underground :: Sunset (1)	WT	pub	t_oth
Webhooks :: Receive a web request (23)	OAu	priv | rp | ro | pub	t | t_oth | rp | ro | unt_g | unt
WeMo Smart Plug :: Switched off (1)	DS	rp	rp
WeMo Smart Plug :: Switched on (1)	DS	rp	rp
Wireless Tag :: Temperature is too high (1)	E	rp	rp
Wireless Tag :: Temperature is too low (2)	E	rp	rp
YouTube :: New liked video (3)	OAcc	pub | priv	t

Table 16: Semantic and information-flow labels of actions used in participants' applets.

Action channel :: action (# of rules with this action)	Semantic label	Secrecy label	Integrity label
abode :: Change mode (1)	DS:S	rp∪to	rp∪to
Android Device :: Launch Google Maps Navigation (2)	P	rp	t | rp
Android Device :: Mute ringtone (4)	P	rp	rp
Android Device :: Play a specific song (3)	P	rp	t | rp
Android Device :: Play music (1)	P	rp	t | rp
Android Device :: Set ringtone volume (11)	P	rp	rp
Android Device :: Turn off WiFi (2)	P	rp	rp
Android Device :: Turn on WiFi (3)	P	rp	rp
Android Device :: Update device wallpaper (2)	P	rp	t
Android SMS :: Send an SMS (2)	OC	ro	t
Android Wear :: Send a notification (4)	L	rp	t | unt
Arlo :: Start recording (2)	DS:S	to | (to∪rp)	to | rp
Blink :: Arm system (4)	DS:S	rp∪to	rp∪to
Blink :: Disarm system (4)	DS:S	rp∪to	tp∪to
CNCT Life :: Toggle on/off (4)	DS	rp∪to	rp∪to
D-Link Smart Plug :: Turn off (2)	DS	rp∪to	rp∪to
D-Link Smart Plug :: Turn on (2)	DS	rp∪to	rp∪to
Day One :: Create Journal entry (5)	L	priv	t
Dropbox :: Add file from URL (9)	L	ro | pub | priv	ro | t | unt
Dropbox :: Create a text file (8)	L	ro | pub | priv	ro | t | unt
ecobee :: Resume thermostat program (2)	DS	rp∪to	rp∪to
ecobee :: Set thermostat comfort profile until next transition (4)	DS	rp∪to	rp∪to
Email Digest :: Add to daily email digest (4)	L	priv	t
Email Digest :: Add to weekly email digest (9)	L	priv	t
Email :: Send me an email (75)	L	priv	unt
Evernote :: Append to note (1)	L	priv | to | ro	t | to | ro
Evernote :: Create image note from URL (1)	L	priv | to | ro	t | to | ro
eWeLink Smart Home :: Turn 1 Channel Switch on or off (2)	DS	rp∪to	rp∪to
Facebook :: Upload a photo from URL (1)	OC	priv | ro | pub	t | to | ro
Fitbit :: Log your weight (1)	L	priv	t
Garageio :: Close garage door (7)	DS:S	rp∪to	rp∪to
Garageio :: Open garage door (5)	DS:S	rp∪to	rp∪to
Gmail :: Send an email (10)	OC	priv | ro	unt | t
Gmail :: Send yourself an email (1)	L	priv	unt | t
Google Calendar :: Create a detailed event (1)	L	ro | to | priv | pub	t | t_oth | ro | to | unt
Google Calendar :: Quick add event (16)	L	ro | to | priv | pub	t | t_oth | ro | to | unt
Google Contacts :: Create new contact (2)	L	priv	t
Google Drive :: Upload file from URL (1)	L	priv | to | ro | pub	t | to | ro | unt
Google Photos :: Upload photo to album (2)	OC	priv | to | ro | pub	t | to | ro | unt
Google Sheets :: Add row to spreadsheet (50)	L	priv | to | ro | pub	t | to | ro | unt
Google Wifi :: Prioritize Device (3)	DS	to | (toUrp)	to
Harmony :: End activity (6)	DS	rp∪to	rp∪to
Harmony :: Start activity (7)	DS	rp∪to	rp∪to
iOS Calendar :: Create a calendar event (1)	L	ro | priv | pub | to	t | ro | to
iOS Health :: Log weight (1)	L	priv	t
iOS Reading List :: Add item to Reading List (2)	L	priv	t
iOS Reminders :: Add reminder to list (4)	L	priv | ro | to	t | ro | to
Leeo :: Change light color (1)	DS:L	rp∪to	rp∪to
LIFX :: Blink lights (6)	DS:L	rp∪to	rp∪to
LIFX :: Breathe lights (1)	DS:L	rp∪to	rp∪to
LIFX :: Change color of lights (3)	DS:L	rp∪to	rp∪to
LIFX :: Toggle lights on/off (1)	DS:L	rp∪to	rp∪to
LIFX :: Turn lights off (5)	DS:L	rp∪to	rp∪to
LIFX :: Turn lights on (6)	DS:L	rp∪to	rp∪to
Lockitron :: Lock Lockitron (1)	DS:S	rp∪to	rp∪to
Lockitron :: Unlock Lockitron (2)	DS:S	rp∪to	rp∪to
Lutron Caseta and RA2 Select :: Activate scene (2)	DS:L	rp∪to	rp∪to
Lutron Caseta and RA2 Select :: Set light level (4)	DS:L	rp∪to	rp∪to
MagicHue :: Switch to dynamic mode for your Lights (2)	DS:L	rp∪to	rp∪to
Manything :: Start recording (2)	DS:S	rp∪to	rp∪to
Manything :: Stop recording (2)	DS:S	rp∪to	rp∪to
Nest Thermostat :: Set temperature (2)	DS	rp∪to	rp∪to
Nest Thermostat :: Set temperature range (1)	DS	rp∪to	rp∪to

Nest Thermostat :: Turn on fan for 15 minutes (1)	DS	rp∪to	rp∪to
Nexia :: Run a Nexia automation (2)	OAu	priv | rp | ro | pub	t | t_oth | rp | ro | unt_g | unt
Noon Home :: Change scene (2)	DS:L	rp∪to	rp∪to
Noon Home :: Turn off room (2)	DS:L	rp∪to	rp∪to
Notifications :: Send a notification from the IFTTT app (117)	L	rp	t
Notifications :: Send a rich notification from the IFTTT app (1)	L	rp	t
Philips Hue :: Blink lights (3)	DS:L	rp∪to	rp∪to
Philips Hue :: Change color (2)	DS:L	rp∪to	rp∪to
Philips Hue :: Dim lights (8)	DS:L	rp∪to	rp∪to
Philips Hue :: Set a scene in a room (3)	DS:L	rp∪to	to
Philips Hue :: Toggle lights on/off (8)	DS:L	rp∪to	rp∪to
Philips Hue :: Turn off lights (3)	DS:L	rp∪to	rp∪to
Philips Hue :: Turn on color loop (1)	DS:L	rp∪to	to
Philips Hue :: Turn on lights (3)	DS:L	rp∪to	rp∪to
Phone Call (US only) :: Call my phone (19)	L	rp	unt
Pinterest :: Add Pin to board (2)	OC	priv | ro | pub	t | ro | to
Pocket :: Save for later (1)	L	priv	t
Pushbullet :: Push a file (1)	L	priv	t
Pushbullet :: Push a note (10)	L	priv	t
RainMachine :: Start a program (2)	DS	rp∪to	rp∪to
reddit :: Submit a new text post (2)	OC	pub | ro	unt | ro
Slack :: Post to channel (14)	OC	ro | priv	ro | t
Smart Life :: Turn off (2)	DS	rp∪to	rp∪to
Smart Life :: Turn on (3)	DS	rp∪to	rp∪to
SmartThings :: Lock (4)	DS:S	rp∪to	rp∪to
SmartThings :: Switch off (20)	DS	rp∪to	rp∪to
SmartThings :: Switch on (33)	DS	rp∪to	rp∪to
SmartThings :: Unlock (6)	DS:S	rp∪to	rp∪to
SMS :: Send me an SMS (9)	L	priv	t
Spotify :: Add track to a playlist (5)	OC	priv | pub | ro	t | ro | to
Spotify :: Save a track (1)	L	priv	t
Stringify :: Run a Stringify Flow (28)	OAu	priv | rp | ro | pub	t | t_oth | rp | ro | unt_g | unt
TiVo :: Display message (7)	L	rp	to
TiVo :: Send remote control key (7)	DS	rp	rp
Todoist :: Create task (1)	L	priv | ro	t | ro | to
TP-Link Kasa :: Toggle (1)	DS	rp∪to	rp∪to
TP-Link Kasa :: Turn off (1)	DS	rp∪to	rp∪to
TP-Link Kasa :: Turn on (1)	DS	rp∪to	rp∪to
Twitter :: Post a tweet with image (1)	OC	ro | pub	t
Twitter :: Update profile picture (2)	OC	pub	t
VoIP Calls :: Call my device (5)	L	rp	t
Webhooks :: Make a web request (15)	OAu	priv | rp | ro | pub | to | tp	t | t_oth | rp | ro | unt_g | unt | to | tp
WeMo Insight Switch :: Toggle on/off (1)	DS	rp∪to	rp∪to
WeMo Smart Plug :: Turn off (7)	DS	rp∪to	rp∪to
WeMo Smart Plug :: Turn on (6)	DS	rp∪to	rp∪to
Wink: Nimbus :: Set dial label (26)	DS	rp∪to	rp∪to
Wink: Shortcuts :: Activate shortcut (7)	OAu	priv | rp	t | rp
Wyze :: Disable motion detection (1)	DS	to	to
Wyze :: Enable motion detection (1)	DS	to	to

D Appendix D: Full Survey Instrument

D.1 Survey Flow

- The survey flow included:
  - Informed consent procedures and instructions for downloading our browser extension, which collected information about participants' applets (not included in this document)
  - Several questions about participants' general use of IFTTT and preferences about their applets (General Questions)
  - Several sets of looping questions pertaining to specific applets (Looping Questions)
  - Questions pertaining specifically to secrecy and integrity (Explicitly Asking About Secrecy and Integrity)
  - Demographic and IUIPC scale questions (not included in this document)
- Blue text was not shown to participants.
- Answer choices are shown in *[italicized square brackets]* after each question.
- Questions included in the analysis in our paper are shown in **bold**, and the section number where these results are conveyed is included in parenthesis. Questions were omitted from discussion for various reasons, including: (1) looping questions for applet categories that only a few participants used (social media), (2) evidence that participants did not understand our questions in the way we intended them, (3) the questions were not relevant to our specific research questions, which shifted from our original vision based on our analysis of participants' applets.

D.2 General Questions

1. **(Section 4.5.1) How many of your IFTTT applets did you create yourself (as opposed to using ones others have created)? If you're not sure, please make your best guess.**
   [Numeric dropdown]

2. **(Section 4.5.1) Do you prefer to create your own applet or search for one that already exists?**
   [Prefer to create, Prefer to search, No preference]

3. Have you ever turned on (i.e., started using) an applet for one-time use or a specific event (e.g., to easily upload photos during a concert or trip)?
   [Yes, No, Don't remember]

4. **(Section 4.5.1) How often do you turn on applets based on a friend's or colleague's recommendation?**
   [Never, Rarely, Sometimes, Often, Don't remember]

5. Have you ever not turned on an applet (or turned off one that you were already using) because you thought it might be unsafe? (E.g., you were uncomfortable with the permissions it asked for, you thought it might compromise account security)
   [Yes, No, Don't remember, I've never wanted to use an applet I thought was unsafe]

6. **(Section 4.5.3) (If yes to question 5) What were your concerns?**
   [Free response]

7. Have you ever had an incident where an applet made you feel unsafe or you felt it violated your privacy (e.g., unlocked your door when you weren't home, posted a picture to Facebook that you didn't want there)?
   [Yes, No, Not sure]

8. **(Section 4.5.3) (If yes to question 7) Please describe:**
   [Free response]

9. Have you made any of the applets you've created publically available?
   [Yes, No, Don't remember, I haven't created any applets]

10. Do you use any other task automation services besides IFTTT? (For example, services like Tasker, Zapier, Stringify, Microsoft Flow, etc.)
    [Yes, No, Don't remember]

11. (If yes to question 10) Which services?
    [Free response]

12. (If yes to question 10) Can you integrate IFTTT applets within the other service's programs? For example, the service Stringify allows the user to add IFTTT applets as part of a Stringify flow.
    [Yes, No, Don't know]

13. (If yes to question 12) Do you use this feature with any of your applets?
    [Yes, No, Don't remember]

14. How much do you agree with the following statements?
    [Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree]
    (a) **(Section 4.5.2) The applets I turn on behave as I would expect from their description.**
    (b) **(Section 4.5.4 and Table 6) I am comfortable telling my friends what applets I use.**
    (c) **(Section 4.5.4 and Table 6 I am comfortable telling my colleagues what applets I use.**
    (d) **(Section 4.5.4 and Table 6) I am comfortable with anyone knowing what applets I use.**
    (e) I would be upset if an applet triggered when I didn't intend it to.

(f) A stranger could trigger some of my applets.

(g) If an applet I was using didn't do what I thought it would do when I installed it, I would notice right away.

(h) If an applet weren't that useful, I would turn it off or delete it.

(i) (Section 4.5.2) **I think the applets that I have turned on are safe to use.**

(j) I have been concerned about the permissions an applet asked for.

(k) When choosing an applet to turn on, my primary criterion is how useful it will be.

15. (If "strongly agree" for statement 11 in question 14.) Since usefulness is not necessarily your primary criterion, what other considerations do you have?
[Free response]

16. Have you ever tried to link the behavior of multiple applets ("chain" them together)? For example, if you get close to your house, your thermostat is set to home mode, and if your thermostat is set to home mode, your lights will turn on.
[Yes, No, Don't remember]

17. (If yes to question 16) Which applets?
[Free response]

18. (If yes to question 16) Did it work as expected?
[Yes, Sometimes yes sometimes no, No, Don't remember]

19. (If "no" or "sometimes yes, sometimes no" to question 18) Please explain what went wrong:
[Free response]

20. Have you ever **unintentionally** made a "chain" between applets? For example, if the temperature gets above a threshold, set the thermostat to cool, and if the thermostat is set below a specific temperature, open the windows.
[Yes, No, Don't remember]

21. (If yes to question 20) Which applets?
[Free response]

22. Have you ever **manually** deleted anything that was posted **automatically** (e.g., a social media post or cloud storage update) by an applet?
[Yes, No, Don't remember, I don't use this type of applet]

23. **(Section 4.5.3) (If yes to question 22) Which applet created the post, and why did you want to delete it? If you do not remember all the exact details, please explain as much as you can.**
[Free response]

24. Would you consider some applets to be more sensitive than others (i.e., you care more about who knows about them or who can trigger them)?
[Yes, No, Not sure]

25. (If yes to question 24) Which applets?
[Free response]

D.3 Looping Questions

D.3.1 Looping Set 1: Applets Using Physical Devices

Since you had applets that used physical devices, we will now ask a set of questions about physical devices, for up to 5 devices.

(The next set of questions (26-29) loops up to five times, for a randomly-chosen set of the participant's applets that used physical devices, as determined by service categories on the IFTTT website)

Asking about physical device [#] of up to 5.

26. Consider the applet "[applet title]". In which room is the [device] device used in the applet [trigger/action]? If the device has multiple components/sensors/etc. (e.g., a smart hub or lighting system), please list all the rooms that you can remember.
[Free response]

27. Does more than one person in the household commonly access the [device] device used in the applet [trigger/action] (either online or in person)? "Access" could mean performing an activity that the device senses; e.g., opening a door that has a sensor attached or walking into a room with a motion sensor.
[Yes, No, Not sure, I live alone]

28. Does the location of the [device] device make you more protective of who knows about the applet "[applet title]"? (For example, an applet that unlocks a ground floor window might be considered more sensitive than one that unlocks a second floor window.)
[Yes, No, Not sure]

29. (If yes to question 28) Please explain:
[Free response]

D.3.2 Looping Set 2: Cloud Storage Applets

Since you had applets that use cloud storage, we will now ask a set of questions about the details of the cloud storage, for up to 5 applets.

(The next set of questions (30-31) loops up to 5 times, for a randomly-chosen set of the participant's applets that used cloud storage, as determined by service categories on the IFTTT website)

Asking about cloud applet [#] of up to 5.

30. **(Section 4.2) Is the file or folder that the applet "[applet title]" updates accessible only to you, or is it shared with others (e.g., housemates, family)?**
[Only me, A group, Don't remember]

31. How often do you check the file or folder used in this applet to see the updates?
[Never, Rarely, Sometimes, Often, Don't remember]

D.3.3 Looping Set 3: Social Media Applets

You have some applets where both the trigger and action use social media or blogging services. We will ask a small set of more detailed questions for up to 5 applets.

(The next set of questions (32-33) loops up to 5 times, for a randomly-chosen set of the participant's applets that used cloud storage, as determined by service categories on the IFTTT website)

Asking about social media applet [#] of up to 5.

32. In the applet "[applet title]", do the people who follow your [service associated with action] account also follow your [service associated with trigger] account?
[Yes, No, Not sure]

33. (If yes to question 32) What are the main differences between the audiences, in your view?
[Free response]

D.3.4 Looping Set 4: Violating Applets

You will now be asked a set of questions about your thoughts and perceptions of various side-effects applets can have. This set will repeat for up to 5 different applets.

(The next set of questions (34-44) loops up to 5 times for a randomly-chosen set of the participant's applets that violate security principles, determined by information-flow analysis)

Asking detailed questions for applet [#] of [#].

34. Consider the applet "[applet title]." How likely is it that this applet could do the following, in your opinion?
[Definitely impossible, Probably impossible, Probably possible, Definitely possible]

    (a) Be triggered by someone outside of your household?

    (b) Cause monetary loss? (e.g., by increasing your electric bill or using up data) OR (e.g., by increasing your electric bill or causing you to replace devices more frequently) (if applet uses a physical device)

    (c) Cause an undesired event if you forget that you have it turned on?

    (d) Spread sensitive information online?

    (e) Cause you embarrassment?

    (f) (Displayed only if applet uses physical device, determined by service categories on the IFTTT website) Damage the physical device that it uses?

    (g) Be used to undermine your home security?

35. Have you ever experienced any of the above consequences or other harmful side-effects when using this applet?
[Yes, No, Not sure]

36. **(Section 4.5.3) (If yes to question 35) Please describe the incident as best you recall, including which applet(s) were involved and what side-effects occurred.**
[Free response]

37. **(Section 4.5.4 and Table 5) Would you be upset if the applet contributed to the following situations occurring:**
[Very Upset, Slightly Upset, Not Upset, This type of harm is impossible for this applet]

    (a) Private information gets posted online unintentionally, possibly embarrassing you.

    (b) You no longer directly control what files are downloaded from email or social media, possibly spreading malware on your computer.

    (c) (Displayed only if applet uses physical device, determined by service categories on the IFTTT website) Your electronic device is used in way it wasn't designed for (such as being toggled on/off very rapidly), possibly reducing its longevity or damaging it.

    (d) Data gets uploaded to your cloud storage more often than you thought, possibly causing you to run out of space.

    (e) You consume more resources (e.g., electricity, phone data, cloud storage space), possibly increasing your bills or otherwise causing you to spend more money.

38. Did you consider the possibility of some of the preceding consequences when deciding to turn on the applet "[applet title]"?
[Yes, No, Don't remember]

39. (If yes to question 38) Which questions?
[Free response]

40. Would you be upset if a *friend* knew you had the applet "[applet title]"?
[Yes, No, A little upset, Not sure]

41. Would you be upset if a *colleague* knew you had this applet?
[Yes, No, A little upset, Not sure]

42. Would you be upset if a *stranger* knew you had this applet?
[Yes, No, A little upset, Not sure]

43. Who is meant to be able to *purposefully* trigger this applet?
[Myself; Trusted individuals, such as my spouse; A wider circle of known individuals, such as my Facebook friends or house guests, Unknown third parties, such as websites or strangers]

44. **(Section 4.5.4) Would you be upset if someone not in the intended group purposefully triggered this applet?**
[Yes, No, Not sure]

D.4 Explicitly Asking About Secrecy and Integrity

D.4.1 Secrecy

(These questions pertain to up to 9 applets with secrecy violations plus 1 safe one, determined by information-flow analysis, selected randomly from the participant's applets. Titles of applets were displayed to participants.)

45. Some applets indirectly pass information from a smaller, more restricted group to a larger, more open one. For instance, an applet that posts to twitter when the user reaches a Fitbit fitness goal is leaking information that only the user knew to all their twitter followers. Out of your applets below, do any of them allow information to leak from a smaller audience to a larger one (whether inadvertently or on purpose)?
[Yes, No, Not sure]

46. **(Section 4.5.2) Thinking about the possible data leakage, has your desire to keeping using any of these applets changed?**
[Yes, I am more cautious of some applets now; No, my desire to use these applets has not changed; I'm not sure.]

47. **(Section 4.5.3) (If yes to question 46) Which applets?**
[Free response]

D.4.2 Integrity

(These questions pertain to up to 9 applets with integrity violations plus 1 safe one, determined by information-flow analysis, selected randomly from the participant's applets. Titles of applets were displayed to participants.)

48. Some applets allow devices or services usually accessed by members of a smaller, trusted group to be indirectly controlled by members of a larger, less trusted group. For instance, an applet that adds a photo to a Google Drive folder whenever the user is tagged in a Facebook photo is essentially allowing any of the user's Facebook friends to add files to the user's private folder (which by default only the user would have access to). Out of your applets below, do any of them allow less trusted groups to control devices or services usually controlled by more trusted groups (whether inadvertently or on purpose)?
[Yes, No, Not sure]

49. **(Section 4.5.2) Thinking about the possible loss of control, has your desire to keeping using any of these applets changed?**
[Yes, I am more cautious of some applets now; No, my desire to use these applets has not changed; I'm not sure.]

50. **(Section 4.5.3) (If yes to question 49) Which applets?**
[Free response]