USENIX Association

Proceedings of the Fourteenth Symposium on Usable Privacy and Security

August 12–14, 2018
Baltimore, MD, USA

Symposium Organizers

General Chair
Mary Ellen Zurko, *MIT Lincoln Laboratory*

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

Invited Talks Chair
Adam Aviv, *U.S. Naval Academy*

Technical Papers Co-Chairs
Sonia Chiasson, *Carleton University*
Rob Reeder, *Google*

Technical Papers Committee
Yasemin Acar, *Leibniz University Hannover*
Nalin Asanka Gamagedara Arachchilage, *University of New South Wales*
Adam Aviv, *United States Naval Academy*
Rebecca Balebako, *RAND Corporation*
Joseph Bonneau, *NYU*
Pam Briggs, *University of Northumbria*
Joe Calandrino, *Federal Trade Commission*
Marshini Chetti, *Princeton University*
Jeremy Clark, *Concordia University*
Heather Crawford, *Florida Institute of Technology*
Alexander De Luca, *Google*
Serge Egelman, *UC Berkeley/International Computer Science Institute*
Sascha Fahl, *Ruhr-University Bochum*
Alain Forget, *Google*
Marian Harbach, *Audi AG*
Apu Kapadia, *Indiana University Bloomington*
Katharina Krombholz, *SBA Research*
Janne Lindqvist, *Rutgers University*
Michelle Mazurek, *University of Maryland*
Andrew Patrick, *Prisus Research*
Heather Patterson, *Intel*
Michael Reiter, *UNC Chapel Hill*
Manya Sleeper, *Google*
Jessica Staddon, *Google*
Mary Theofanos, *NIST*
Blase Ur, *University of Chicago*
Emanuel von Zezschwitz, *University of Bonn*
Yang Wang, *Syracuse University*
Rick Wash, *Michigan State University*
Heng Xu, *Penn State University*

Lightning Talks and Demos Chair
Heather Crawford, *Florida Institute of Technology*
Scott Ruoti, *MIT Lincoln Laboratory*

Karat Award Chair
Jose Such, *Kings College London*

Posters Co-Chairs
Yasemin Acar, *Leibniz University Hannover*
Kent Seamons, *Brigham Young University*

Tutorials and Workshops Co-Chairs
Elissa Redmiles, *University of Maryland*
Florian Schaub, *University of Michigan*

Publicity Co-Chairs
Joe Calandrino, *Federal Trade Commission*
Patrick Gage Kelley, *University of New Mexico*

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

Email List Chair
Lorrie Cranor, *Carnegie Mellon University*

USENIX Liaison
Casey Henderson, *USENIX Association*

Steering Committee
Lujo Bauer, *Carnegie Mellon University*
Konstantin Beznosov, *University of British Columbia*
Robert Biddle, *Carleton University*
Sonia Chiasson, *Carleton University*
Sunny Consolvo, *Google*
Patrick Gage Kelley, *Google*
Jaeyeon Jung, *Samsung Electronics*
Apu Kapadia, *Indiana University Bloomington*
Rob Reeder, *Google*
Heather Richter Lipford, *University of North Carolina at Charlotte*
Matthew Smith, *University of Bonn, Fraunhofer FKIE*
Rick Wash, *Michigan State University*
Mary Ellen Zurko, *MIT Lincoln Laboratory*

External Reviewers

James Nicholson
Lynne Coventry
Yaxing Yao
Natã Barbosa
Mahmood Sharif
Chris Fennell
Tousif Ahmed
Pamela Wisniewski
Yasmeen Rashidi
Qatrunnada Ismail
Rakibul Hasan
Christian Tiefenau
Maria Muszynska
Arunesh Mathur
Bela Genge
Nathan Malkin
Danny Yuxing Huang
Meghan C McLean
Hua Deng
Can Liu

SOUPS 2018
Fourteenth Symposium on Usable Privacy and Security
Message from the Chairs

Welcome to SOUPS 2018!

We are delighted to bring SOUPS into its 14th year. As long-time SOUPS attendees and participants ourselves, we are happy to see the growth in the range of topics covered at SOUPS and in the number of people joining the community. Technical paper presentations form the core of the SOUPS program, but the conference also includes workshops, posters, lightning talks, and a keynote.

In 2016, SOUPS became an independent conference body. For the last three years, we have partnered with USENIX for hosting and administrative support, a move that has enabled continued growth for the conference. This year, we are co-located with the USENIX Security Symposium for the first time. Co-locating the two conferences allows for interactions and shared ideas between SOUPS and USENIX Security attendees, and we are excited to see the result.

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. SOUPS is a product of the hard work by all the SOUPS Organizers, the SOUPS Steering Committee, the Technical Papers Committee, the Workshop organizers, the Posters jury, and the USENIX staff. We thank each and every one of you for your contributions to SOUPS 2018.

Mez is serving her final year as General Chair of SOUPS and Chair of the Steering Committee. Next year, Heather, who served as Vice Chair this year, will step into this role for 2019 and 2020. If you are interested in helping with SOUPS 2019 in any way, please contact Heather.

We thank each of our sponsors for their support—NSF, Facebook, Google, Mozilla, and DMTF. SOUPS would not be possible without their generous support. Please visit our web site to view the recipients of the SOUPS 2018 awards—Distinguished Paper, IAPP SOUPS Privacy Award, Distinguished Poster, and the John Karat Usable Privacy and Security Student Research Award. Congratulations to all of the recipients for their outstanding work.

Mary Ellen Zurko, *MIT Lincoln Laboratory*
General Chair

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

Sonia Chiasson, *Carleton University*
Technical Papers Co-Chair

Rob Reeder, *Google*
Technical Papers Co-Chair

SOUPS 2018
Fourteenth Symposium on Usable Privacy and Security

August 12–14, 2018
Baltimore, MD, USA

User Authentication

Behaviors and Practices

Online Privacy

Data Exposure, Compromises, and Access

Developers

Understanding and Mindsets

(continued on next page)

Models, Beliefs, and Perceptions

Replication Study: A Cross-Country Field Observation Study of Real World PIN Usage at ATMs and in Various Electronic Payment Scenarios

Towards Understanding Why People Do, or Do Not, Shield PIN Entry

Melanie Volkamer
Karlsruhe Institute of Technology (KIT)
Technische Universität Darmstadt
melanie.volkamer@kit.edu

Andreas Gutmann
OneSpan Innovation Centre & University College London
andreas.gutmann@onespan.com

Karen Renaud
Abertay University
University of South Africa
University of Glasgow
k.renaud@abertay.ac.uk

Paul Gerber
Technische Universität Darmstadt
gerber@psychologie.tu-darmstadt.de

Peter Mayer
Karlsruhe Institute of Technology (KIT)
Technische Universität Darmstadt
peter.mayer@kit.edu

ABSTRACT

In this paper, we describe the study we carried out to replicate and extend the field observation study of real world ATM use carried out by De Luca *et al.,* published at the SOUPS conference in 2010 [10]. Replicating De Luca *et al.*'s study, we observed PIN shielding rates at ATMs in Germany. We then extended their research by conducting a similar field observation study in Sweden and the United Kingdom. Moreover, in addition to observing ATM users (*withdrawing*), we also observed electronic *payment* scenarios requiring PIN entry. Altogether, we gathered data related to 930 observations. Similar to De Luca *et al.,* we conducted follow-up interviews, the better to interpret our findings. We were able to confirm De Luca *et al.*'s findings with respect to low PIN shielding incidence during ATM cash withdrawals, with no significant differences between shielding rates across the three countries. PIN shielding incidence during electronic payment scenarios was significantly lower than incidence during ATM withdrawal scenarios in both the United Kingdom and Sweden. Shielding levels in Germany were similar during both withdrawal and payment scenarios. We conclude the paper by suggesting a number of explanations for the differences in shielding that our study revealed.

1. INTRODUCTION

People have been drawing cash from automated teller machines (ATM) for at least half a century [4]. The 21st century heralded an increasing use of card-based electronic payments [13]. Most bank cards are Chip & PIN based, allowing people either to withdraw money or pay for goods and services using the same card. To complete a transaction, the customer presents the card and provides a PIN to authenticate themselves. Exceptions are, for instance, Germany, where *Chip & Signature* is a common alternative to Chip & PIN, and the United Kingdom, where contactless payment (*tap only* for amounts less than £30) is gaining market share [39]. PINs are required during withdrawals in all countries, no matter how low the transaction amount.

PIN entry is not without risk, since thieves could observe the PIN (in person, or using a camera) and use the knowledge later, once they have managed to clone or steal the actual card. To prevent this, people are advised to take the precaution of shielding their PINs when they enter it (as well as being advised not to carry a note of their PIN together with the actual card).

In 2010, De Luca *et al.* investigated factors that impacted decisions related to taking security precautions when engaging in PIN-based ATM authentication [10]. The researchers observed how people entered their PINs at ATMs; in particular, whether people acted to protect their PIN entry from possible skimming attacks [3]. They conducted follow-up interviews to gain insights into the contextual factors affecting secure behaviors. We replicated their research study, and extended it as follows:

- *PIN usage scenarios:* Common electronic payment scenarios (i.e. in supermarkets or in restaurants / coffee bars) are very similar to withdrawing money from an ATM in terms of PIN authentication being required. We wanted to explore differences in PIN usage during payment scenarios, too. We also wanted to elicit explanations for shielding differences we observed. Similar research questions were suggested by De Luca *et al.* as a topic of future interesting investigations [10].
- *Countries:* While De Luca *et al.* [10] collected data in Germany and the Netherlands, they only reported overall observations. However, we believed that more detailed comparisons between different countries, particularly when considering different scenarios (payment/withdrawal), would de-

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

liver interesting insights in terms of shielding percentages and factors impacting PIN shielding.

Following the previous researchers' example, we commenced with an observational field study and then conducted interviews once the data from the first study was analysed. We collected data in Germany, similar to De Luca *et al.* We also extended the observation field study to both Sweden and the United Kingdom, and conducted interviews in all three countries.

De Luca *et al.* reported that 67% of the observed ATM users did not take any precautions against PIN skimming attacks. Almost a decade later, we observed the same high percentage of people *not* shielding PIN entry at ATMs (64% in Germany, 71% in the U.K, and 71% in Sweden). We discovered that the activity of either *withdrawing* or *paying*, as well as the observation country, were significant predictors of PIN shielding behavior. Further results are:

- In Germany, there was no significant difference in PIN shielding incidence during withdrawal and payment scenarios.
- In the United Kingdom and Sweden, we observed significantly fewer people shielding their PINs during payment, than during withdrawal transactions.
- Significantly more people shielded their PINs when paying with their cards in Germany, as compared to the United Kingdom and Sweden.
- Significantly more people shielded their PINs when paying with their cards in the United Kingdom, as compared to Sweden.

De Luca *et al.* identified a number of contextual factors from their follow-up interviews to determine why people did, or did not, shield PIN entry. One was that of being accompanied. We also recorded whether or not people were accompanied in our observation field study. However, our study did not reveal significant differences for this factor.

The interviews helped us to explain our findings; particularly with respect to the differences between shielding incidence during withdrawing and paying. Possible explanations are habituation (people engaging in more electronic transactions feel safer doing so, and are less likely to shield their PINs), lack of reminders to shield, the presence of hard cash during withdrawals, different goals (withdrawing means the primary goal is obtaining cash in hand; paying means the primary goal is obtaining desired products or services), and a lack of understanding of the actual attack scenarios. In terms of the latter, the primary threat might not be surrounding people, but rather strategically positioned security cameras which could easily record unshielded PINs.

Thus, we conclude that it seems particularly worthwhile to add opaque hardware shields to Chip&PIN devices which effectively removes the need for people to shield themselves. Just-in-time reminders might also reduce the risk of criminals gaining knowledge of people's PINs, as well as raising awareness of PIN shielding during payment scenarios.

2. METHODOLOGY

We commence by providing details of De Luca *et al.*'s study, and explaining how we went about replicating and extending it. In particular, we explain what precautions we took in order to ensure that the research was carried out in accordance with ethical requirements.

2.1 De Luca et al.'s Study

De Luca *et al.* [10] carried out a PIN observation study in 2010.

Goal: Their goal was better to understand PIN-based ATM authentication both with respect to taking any precautions against PIN skimming attacks and the time needed to authenticate. Furthermore, they wanted to determine how alternative authentication approaches could be evaluated and compared to existing ones.

Methodology: During their research, De Luca *et al.* observed ATM interactions at six locations in two cities in Germany and the Netherlands: a total of 360 observations. The observations (i.e. whether or not to shield the PIN entry and how long authentication takes) were recorded on a tally sheet during multiple sessions, by the same researcher "*to keep the data comparable, since different people might apply different standards during the observation, deliberately or not*" [10, p. 2]. After analyzing the collected data, several problems regarding the timing were identified. Correspondingly, two followup studies were conducted. To gain greater insights into the findings from the field evaluation, they subsequently carried out interviews with other people (not the ones they observed).

Findings: They found that the majority of the people they observed (65%) did not take any precautions against PIN skimming attacks (i.e. less than 65% shielded their PIN entry). In addition. the interviews revealed that contextual factors exerted a strong influence both on security behaviors and to the time required to authenticate. Example factors are distractions, physical hindrance (e.g. due to bags in peoples hands), and trust relations. Based on their findings, they suggested a number of "lessons learned" to inform subsequent field studies into the use of privacy-sensitive technologies, as well as a number of implications for the design of alternative ATM authentication systems. Their lessons learned section emphasised the importance of improving tally sheet designs during trial studies and adherence to strict rules during observations to ensure validity and comparability of the results.

2.2 Achieving Replication

We based the study design on De Luca *et al.*'s [10], and also incorporate design aspects from their lessons learned section.

Similar to De Luca *et al.*'s study, each location was visited at least twice during different time periods. By doing so, we ensured that the collected data was as diverse as possible. Replicating De Luca *et al.*'s study, we observed a variety of different bank ATM machines at different locations. We also observed a variety of scenarios during which PIN-based authentication was required during electronic payment.

We chose the locations similarly to De Luca *et al.* for their study. In effect, we chose locations that enabled us non-intrusively to observe the interactions with the corresponding devices. We identified scenarios where the devices were visible from public seating areas, such as street cafés. By so doing, we ensured that the observer did not arouse suspicion. Similar to De Luca *et al.,* the observation sessions were not prolonged so as to minimize the risk of raising suspicion and concern.

As reported by De Luca *et al.,* all observations were performed and recorded (in written form) by only one researcher. This eliminated inter-observer bias. Following De Luca *et al.*'s protocol, observations were only added to the data set if the observer was 100% sure about whether the subject had shielded their PIN or not. If his view was obscured, the observer did not record the event. The researcher did not observe any fraudulent incidents during the observation ses-

sions.

2.3 Observation Study

We now describe the variations we studied, for each of the two factors (PIN usage scenario, country/locations), and the content of the written protocol.

2.3.1 PIN Usage Scenarios

De Luca *et al.* investigated actions connected with ATM withdrawals. We studied interactions during this scenario and also studied payment scenarios during which PINs were required to authenticate: supermarkets and restaurants/coffee bars. Compared to withdrawing cash, the electronic payment process does not involve actual cash being handled. Furthermore, the subject's main task is to purchase something. Unlike ATM interactions, which is a solo activity, other people are often legitimately involved in payment interactions. For example, a shop assistant might be instructing a customer to insert their card and enter their PIN. We wanted to determine whether these different scenarios (withdrawing vs. paying) would make a difference to PIN shielding rates. We also considered two different types of payment scenarios, so as to reveal differences between payments in supermarkets at the cash register and payments in a restaurant/coffee bar setting.

Our 930 field observations were performed at different locations: 310 in each country. Besides ATMs, we observed people at various electronic payment scenarios involving a PIN authentication. The observation field study took place over a period of two weeks in each country. After the field observation study, follow-up public interviews were conducted in all three countries.

2.3.2 Countries and Locations

We conducted our observation field study in three different European countries, each with different profiles with respect to withdrawing cash and cashless payments. Based on data from the European Central Bank [13] and Eurostat [38], we identified three countries for our study: Germany, the United Kingdom and Sweden. People living in Germany, on average, withdraw money about as frequently as they pay electronically. People living in the United Kingdom use bank cards more frequently for both, to withdraw (smaller amounts of) money and generally pay for things electronically. Furthermore, in the United Kingdom, contactless payment (for payments under £30) is gaining market share [39]. This only requires PIN authentication for amounts over £30. In Sweden, "*cash is used relatively infrequently [...] while cards are used to a great extent*" [33] and also for very small amounts of money. For more details about the differences see Table 1.

We chose locations in each country to collect samples that are broad in range and comparable to each other.

Frankfurt, Germany. We included two ATMs in Germany (45 observations each). Both were located in train stations. Furthermore, observations were conducted in a supermarket (100 observations) and two restaurants (120 observations). A notable distinction between those restaurants was that customers at one restaurant paid before eating, while customers in the other restaurant paid just before departing.

Glasgow, United Kingdom. Our observations in the United Kingdom comprised a supermarket (100 observations), a fast food restaurant, and a coffee bar (both with 120 observations in total) as well as two ATMs in pedestrian precincts (45 observations each). The fast food restaurant provided multiple self-service kiosks, while customers in the coffee bar queued at a single teller.

	U.K.	**Germany**	**Sweden**
Withdrawals *per capita*	43.98	32.43	21.96
Avg. value of withdrawal (Euro)	83.00	128.21	108.88
Card payments *per capita*	178.99	33.21	235.47
Avg. value of card payment (Euro)	59.28	72.09	32.1
Avg. number of PIN entries per capita	222.97	65.64	257.43

Table 1: The number and value of withdrawals and card payments in the United Kingdom, Germany and Sweden in 2014 [13]. The average number of PIN entries *per capita* is based on the population on the 1st of January 2014 [38]. This presents an upper bound for Germany because Chip & Sign is commonly used [12] and for the United Kingdom because of the high usage of contactless payments [14].

Karlstad, Sweden. The observations in Sweden comprised two ATMs inside a building (45 observations each), a supermarket inside a mall (100 observations in total), a restaurant within a department store, and a payment terminal at the exit of the same department store (in total, 120 observations).

2.3.3 Written Protocol

The written protocol comprised the following information: country, scenario (including ATM vs. supermarket vs. restaurant/coffee bar), time of the day/date, shielded (or not), and whether accompanied by other people (or not). The fact that the latter might be important was suggested by De Luca *et al.*'s findings [10]. Their interviewees suggested that being accompanied negatively impacts people's decisions to shield due to social awkwardness.

2.4 Follow-Up: Public Interviews

We conducted public follow-up interviews, in order the better to interpret our observation findings. Interviews took place over a period of several days in the same cities where observations took place (while not necessarily close to the observation locations).

Similar to De Luca *et al.*'s protocol, people were first asked whether they would be available for a short interview. If they consented, they were informed that the interview was being conducted as part of a research project, and assured that no private data would be collected. Subjects were asked to be frank and honest in their responses. They were not interrupted as long as they felt like talking. Notes were taken manually. The interviews were conducted in English in the United Kingdom and Sweden, and in German in Germany.

The interview protocol was slightly different to the one from De Luca *et al.*'s . Because we had extended the observation field study by adding additional scenarios and countries, we wanted to address the differences we identified between these different settings in particular between the payment and the withdrawing scenario. We thus used the following protocol:

1. Describe, in detail, how you use your card to pay when shopping.
2. Describe, in detail, how you use your card to withdraw money at an ATM.
3. If PIN shielding has not been mentioned during the first two responses, ask:

(a) "You probably use only one hand to operate the device. What do you usually do with your other hand in both situations?"

(b) "Do you regularly shield your PIN entry?"

4. If PIN shielding is only mentioned in connection with ATMs:

   (a) What is the difference between withdrawing at an ATM and paying in a shop?

   (b) Why do you shield your PIN at one but not the other?

5. Have you heard about crimes related to PIN entry? If so, what did you hear and where did you hear it?

6. Do you sometimes see other people covering their hands when they enter their PINs? What do you think when you see them do this? Why?

7. Assume you are in a shop, or at an ATM, with a good friend, and he or she shields their PIN as they enter it. What would you think? Why?

Note that we decided to commence the interview with questions about scenarios, whereas De Luca *et al.* asked questions specifically about PIN security. We wanted to make sure we did not bias initial responses by mentioning security.

2.5 Ethical and Legal Considerations

When we investigate security behaviors, self reports often do not reflect actual behaviors, due to the social desirability effect [16, 36]. This makes surveys and interviews less than reliable in delivering insights into security-related behaviors. Observations reveal actual, rather than self-reported, behaviors, which is invaluable in understanding how to improve the design of socio-technical security systems.

Observational studies are a powerful tool for studying social worlds [23], and security behaviors in public places lend themselves to observational studies. Yet observational studies require researchers to take extra special care with respect to ethical and legal aspects of their studies. Before commencing the observations, we thus considered the ethical and legal aspects very carefully.

Ethical requirements and general recommendations provided by the American Psychological Association in their Ethical Principles of Psychologists and Code of Conduct [1] and the British Sociological Society Guidelines [6] were followed in planning this study. Ethics requirements and general recommendations provided by Technische Universität Darmstadt[1] [37] were strictly adhered to. However, two areas of concern merited special consideration and are therefore further discussed in the next paragraphs: (1) informed consent, and (2) deception.

(1) Informed Consent: The first issue was that it was not possible to obtain informed consent from the subjects we observed in our study. To seek consent would likely have changed behavior and compromised the integrity of the investigation [6, 34]. Spicker [34] explains that some studies simply cannot obtain consent. He cites three examples: "*Observing a crowd at a football match, watching drivers in moving cars, or attending a meeting of shareholders*" (p. 3). We believe our context to be similar to these, in the sense that requiring the researcher to obtain consent would have made it impossible for him to carry out the research in an ecologically-valid way.

Murphy and Dingwall [29], reporting on the ethics of ethnographic studies, argue that people in public spaces can expect to be scrutinized by anonymous others. They explain that, in the case of public behavior, people's consent to being observed is implied by their presence in the public place. Yet the researcher has to treat their subjects with respect and decency, which is what we sought to do. We considered that, in our study, consent was unachievable and would have invalidated our findings. Spicker [34] explains that where there is a need to carry out research that is minimally intrusive, in public, it is often not possible to obtain consent from those being observed. We thus did not obtain informed consent from our observed subjects.

(2) Deception: The second potential concern is that subjects in observation studies are often subject to deception. We designed our study to be a *covert non-participant observation* study instead of a *researcher-as-participant* study, which is much more deceptive, and makes it more difficult for researchers to preserve anonymity of subjects. This is harder to justify ethically than the kind of non-intrusive study we carried out [29, 11]. Our subjects were not deliberately deceived at all, so this was not an ethical concern.

However, there are some *limitations* and *challenges* to consider when carrying out non-participant covert observation studies [25, 31]:

(a) **Observer Effect:** the observer's presence could affect the actions of the subject.

(b) **Objectivity:** the observer needs to ensure that he/she maintains objectivity during observation.

(c) **Selectivity:** ensuring that observations are captured in a variety of situations to offset selectivity bias.

(d) **Hearing the subjects' voices:** ensuring that the final account does not only reflect the researcher's voice.

(e) **Unobtrusiveness**: not standing out in the environment when recording observations.

The limitations were addressed by the following precautions, replicating all of those applied by De Luca *et al.* [10] (Table 2 shows the mapping between the limitations and the precautions.)

(1) **Privacy:** PIN entry is a secret and sensitive issue. It was essential to ensure that we did not gain knowledge of anyone's PIN while carrying out the observations. The observation locations were selected so that, in order to respect the privacy and secrecy of our unwitting subjects, we were always able to observe from a vantage point that allowed us to see whether people were shielding PIN entry, but not to be able to observe the PIN itself. This was achieved either by positioning the observer to the side of the device, at an obtuse angle, or to position the observer too far away to be able to observe anything more than the use of a hand or wallet to shield PIN entry.

[1] Relevant for the research reported here (observational study without any interaction with the participants) are the avoidance of damage, stress, fear or other aversive effects on the subjects of the study, i.e. the observed, the avoidance of the collection of personal data, if this is not necessary, and the preservation of subject anonymity, especially in the collection of data related to minorities, which could be deanonymised unintentionally by statistical linking of data.

(2) **Location Accessibility & Variety:** the observation locations were selected in such a way that the observer could not see the device's screen, and were easily accessible. Moreover, observations were carried out at a range of locations.

(3) **Anonymity:** We did not collect any personal data such as names, contact data, photos or videos, so as to grant our subjects full anonymity.

(4) **Respect:** we interviewed *other* Chip&PIN card holders, who were not observed subjects, after we had carried out all the observations, in order to hear their explanations for shielding decisions.

(5) **Inconspicuousness:** the observer acted as required by the environment so that he did not stand out unduly. For example, if he was observing in a coffee shop he ordered a coffee, if he was observing out in the street he sat on a bench and appeared to be resting. He engaged in no interaction with the subjects, so as not to occasion any disquiet.

(6) **Recording Protocol:** the observer manually recorded the data related to the subject's shielding actions.

Limitation	**Precaution**
(a) Observer Effect	(1) Privacy, (3) Anonymity
(b) Objectivity	(6) Reporting Protocol
(c) Selectivity	(2) Location Accessibility & Variety
(d) Hearing the subjects' voices	(4) Respect
(e) Unobtrusiveness	(5) Inconspicuousness

Table 2: The mapping from the aforementioned limitations to the precautions we took in designing our study.

We informally consulted lawyers and experts from data protection authorities in the respective countries. We also asked Karlsruhe Institute of Technology's legal department to provide feedback regarding the legal aspects of our study design. Given the precautions we designed into our study, as detailed above, they could not identify any legal issues with our study design. This included observations carried out in indoor locations, such as restaurants. None of the lawyers we consulted could see that we needed to get in touch with the owner/manager of these locations beforehand, given the precautions we took. In particular, we respected the privacy of the subjects we observed and did not not interact with, or impede, anyone. They also confirmed that, given these precautions, we did not have to obtain signed consent from the subjects. Again, the most important aspects were that subjects were essentially anonymous for research purposes, and that the researcher did not interact with them in any way.

In conclusion, we planned our study activities carefully in order to ensure that we did not harm the safety, dignity, or privacy of the people we observed, as advised by the European Commission [19].

2.6 Methodology Limitations

Following the De Luca *et al.*'s [10] methodology means facing the same limitations. As explained by De Luca *et al.,* it was important not to interview subjects after observing their actions. Instead, an independent set of people was interviewed. That being so, the same limitation holds: the explanations provided by our interviewees were not directly provided by the observed subjects and thus cannot be considered to be reliable causatives.

It is also possible that people falsely represented their usual PIN-related actions during interviews due to social desirability of making a good impression, or to please the interviewer. We have no indication that this happened but this limitation must be acknowledged.

3. FINDINGS

We first present the findings from the observation field study and then those from the follow-up interviews.

3.1 Observation Field Study

The details of the study are provided in Table 3 and summarized in Table 4.

Results from Replication. De Luca *et al.* [10] reported that 120 out of 360 (33.3%) of the people they observed at ATMs did observably shield their input. We recorded that 39% of the people being observed at ATMs in Germany shielded their PINs, with 29% in both Sweden and the United Kingdom shielding.

We compared the shielding behavior at ATMs in all locations with that reported by De Luca *et al.* [10] on pair-wise significance with two-proportion z-tests. This method is appropriate for single characteristics (binary data) of two independent groups sampled at random [7]. The tested hypothesis is that shielding incidence at each of the three locations differ significantly from that reported by De Luca *et al.* . The null hypothesis is that there is no difference. The results of all tests reveal no significant differences at $p<.05$ (see Table 5), therefore the alternative hypothesis is rejected, although this does not mean that the null hypothesis would be accepted.

Regression Modelling. We tested the collected data (see Table 3) with regression modelling techniques and set the shielding behavior as the dependent variable. Categorical variables, e.g. the country of observation, were coded into indicator variables before performing the regression modelling. We identified the person's activity of either withdrawing or paying, as well as the country in which the sample was collected, as significant predictors of shielding behaviors (see Table 6). The linear regression model accounts for about 10% of the variation (R^2 =.100, corrected R^2 =.0956, and standard error =.382). The model provided a significant prediction of the criteria 'shielding behavior' with F=20.636 and $p<.001$. The regression model identified two significant predictors for shielding behavior: the country in which the sample is collected and the activity of either withdrawing or paying. It does not indicate whether the combination of both significant predictors is a significant predictor as well, i.e.

- It is more likely that people shielded their PINs when withdrawing money, as compared to paying.
- It is more likely that people shielded their PINs in Germany, as compared to the United Kingdom and Sweden. It is also more likely for people in the United Kingdom to shield their PINs, as compared to Sweden.

Post-hoc ANOVA comparison. We tested the between-subjects effect of independent variables 'country' and 'scenario' (i.e. paying versus withdrawing at ATMs versus supermarket versus restaurant/coffee bars (labelled 'others' in Table 3)) on the dependent variable 'shielding behavior' with two-way ANOVA. We have applied the Sidak correction to compensate for the accumulation of

	United Kingdom				Germany				Sweden			
	ATM	Pay	*Sup*	*Others*	*ATM*	*Pay*	*Sup*	*Others*	*ATM*	*Pay*	*Sup*	*Others*
Total	*90*	*220*	*100*	*120*	*90*	*220*	*100*	*120*	*90*	*220*	*100*	*120*
Shield	29%(26)	14%(30)	13%(13)	14%(17)	36%(32)	34%(74)	34%(34)	33%(40)	29%(26)	0%	0%	0%
Company	13% 12)	35%(79)	47%(47)	27%(32)	3%(3)	30%(65)	42%(42)	19%(23)	8% (7)	–	–	–
↳ shield	58%(7)	15%(12)	15%(7)	16%(5)	0%	31%(20)	33%(14)	26%(6)	14%(1)	–	–	–

Table 3: The percentages and total amounts for the observations, per scenario, and per country. Note that "Others" refers to restaurants and coffee bars. A long dash denotes irrelevance of the data field due to '0%' in the row above. 'Sup' is used as shortcut for supermarket due to space constraints.

	ATM	Pay	Supermarket	Others
Total	*270*	*660*	*300*	*360*
Shield	31% (84)	16% (104)	16% (47)	16% (57)
Company	8% (22)	31% (206)	45% (129)	23% (77)
↳ shield	36% (8)	16% (32)	16% (21)	14% (11)

Table 4: The percentages and total numbers for the observations per scenarios for all three countries. Note that "Others" refers to restaurants and coffee bars.

	United Kingdom	Germany	Sweden
z-score	-0.81	0.4	-0.81
p value	.42	.69	.42

Table 5: The results of the two-proportion z-test on data reported by De Luca et al. [10] and our ATM samples.

	Standardised beta	T	Significance
Germany	.303	8.409	<.001
ATM	.174	4.961	<.001
United Kingdom	.113	3.124	.002
Company	.010	.289	.773
Supermarket	-.004	-.117	.907

Table 6: The regression model data, with coefficients for dependent variables of whether subjects shielded the PIN entry, or not.

type I error. Both major effects as well as the interaction were significant. Since there were no *a-priori* hypotheses, we calculated post-hoc comparisons, comparing behavior in the three countries across all scenarios. The results are presented in Table 7. The most important findings are:

- For the *withdrawal* scenarios, there were no significant differences between shielding across the three countries.
- For the *payment* scenarios, there are significantly more subjects in Germany who shielded their PINs, as compared to the other two countries.
- For the *payment* scenario, significantly more United Kingdom subjects shielded their PINs, as compared to those in Sweden.
- In Germany, there is no significant difference between shielding while either withdrawing or paying.
- There are significant differences between the three scenarios (withdrawing and supermarket/coffee bar) in the United Kingdom and Sweden (with fewer people shielding their PINs during payment, as compared to withdrawing).
- No differences, in terms of PIN shielding, manifested between the two different payment scenarios: supermarkets and others (restaurants/coffee bars), across all three countries.

We did not find any differences in terms of 'being accompanied during PIN entry', neither for the whole sample nor for the three different country-specific samples.

3.2 Follow-Up: Public Interviews

The focus of our interviews was on explaining the differences between withdrawing and paying in the different countries. We conducted a total of 27 interviews: ten in Sweden, ten in the United Kingdom and seven in Germany. The written notes were coded by two of the authors. We used structural coding [27] for initial segmentation of the data and magnitude coding [28, 42] on the collected segments. A three-level magnitude code was applied: several > some > few. The following categories, as possible explanations for shielding, were identified.

3.2.1 ATM Environments Considered More Risky

Several subjects said that they considered the ATM environment to be less safe. One reason, cited by several interviewees, is that there was little to no media coverage of PIN-related crime elsewhere than at ATMs. During some interviews, it was reported that ATMs were often in less secure environments, especially when they were outside banks. Several participants mentioned that strangers hanging around ATMs were mistrusted more than in other scenarios "*...at an ATM anyone could stand behind you. But people in a supermarket are there to buy something*"). Actually, in payment scenarios, the subjects perceived strangers as a 'protector', and assumed that they would implicitly provide protection by spotting external threats. In particular, the cashier and accompanying friends are perceived to be another person who can 'exercise care'. In Germany, in particular, customers commonly hand over the card to the cashier, who then puts the card into the device, prepares everything and asks the customer to enter their PIN. Few interviewees mentioned that the cashier or waitresses are usually discreet enough to turn their bodies away, or avert their eyes, when a customer is entering their PIN.

Thus, other than the withdrawal scenario, people did not consider co-located people a threat in supermarkets, restaurants and shops. Few subjects were not particularly specific but just commented: "*You're not supposed to get robbed in stores*" or "*Not something you usually think about in a store*"

3.2.2 Reminded by Displayed Advice

During some interviews, subjects mentioned that they shielded their PINs when they were visibly reminded to do so. It was acknowledged that only ATMs display such advice: "*There are warnings*

			Mean diff.	Standard error	Sign.	95% conf. interval for the difference	
						Lower boundary	upper boundary
ATM	Germany	U.K.	,067	,057	,559	-,069	,202
	Germany	Sweden	.067	.057	.559	-.069	.202
	U.K.	Sweden	<0.01	.057	1.000	-.135	.135
Supermarket	Germany	U.K.	.210*	.054	**.000***	.082	.338
	Germany	Sweden	.340*	.054	**.000***	.212	.468
	U.K.	Sweden	.130*	.054	**.046***	.002	.258
Others	Germany	U.K.	.192*	.049	**.000***	.074	.309
	Germany	Sweden	.333*	.049	**.000***	.216	.451
	U.K.	Sweden	.142*	.049	**.012***	.024	.259
Germany	ATM	supermarket	.016	.055	.989	-.116	.147
	ATM	Others	.022	.053	.966	-.104	.149
	Supermarket	Others	.007	.051	.999	-.116	.130
UK	ATM	supermarket	.159*	.055	**.012***	.027	.291
	ATM	Restaurant / Cafe	.147*	.053	**.016***	.021	.274
	Supermarket	Others	-.012	.051	.994	-.135	.111
Sweden	ATM	supermarket	.289*	.055	**.000***	.157	.421
	ATM	Others	.289*	.053	**.000***	.162	.415
	Supermarket	Others	<0.01	.051	1.000	-.123	.123

Table 7: Results of post-hoc comparisons for the three countries, in terms of the scenario, and for the three scenarios for the three countries. Those that are significant are starred.

at ATMs, thus I cover automatically. Else I wouldn't because there is no need". Indeed, in our study only the ATMs displayed such reminders.

3.2.3 Cash Perceptions

Few interviewees expressed their views that ATMs would be more strongly connected to bank accounts and to hard cash ("*Because the ATM is, like, about money*"). In their opinion, this perception would frame actions in the vicinity, implicitly prompting security precautions.

3.2.4 Habitual Protective Actions

Some subjects merely said shielding was a habit, perhaps prompted some time ago because they had observed others doing it (social norm), or because their parents taught them to do it. This type of argumentation was actually used in both ways: some participants said others are doing it (in particular friends or parents), which is why they shield their PIN without really thinking about it: "*This is just normal"*. But few others argued that it is normal to enter the PIN, as "*fast as possible"* as no one else shields. A few also considered that the shopping scenario exerts more time pressure than the ATM scenario: at ATMs people generally stand back and the activity is essentially solo, whereas payment scenarios usually involve at least one other person who is somehow involved in the transaction.

3.2.5 Social Awkwardness

Some people were put off by impressions of social unacceptability. Some participants reported that shielding might signal mistrust to people around you: "*I don't want to look like a freak*", "*Only old people cover*", "*Covering feels stupid*", "*People who cover are paranoid*". While these reasons may hold for both scenarios, it might be worse for paying. These subjects mentioned that they are often accompanied by friends or relatives during payment scenarios. On the other hand, they usually withdrew money on their own. One mentioned situational differences: at the supermarket, friends usually go to the cash register together while someone usually breaks away from the group to withdraw money.

3.2.6 Further Findings

While the sample is clearly not representative, we can conclude the following:

- Very few interviewees specifically mentioned attacks. For example, it is easier to install a skimmer on an ATM. Some mentioned the risk related to strategically-placed surveillance cameras that are able to record unshielded PINs. However, such threats were only mentioned as related to the ATM context. Some subjects only considered shielding necessary at ATMs if strangers were standing too close for comfort. Similar findings were reported by De Luca *et al.* [10]. They, too, reported subjects securing their PINs by entering them as quickly as possible. Others checked the surrounding area before approaching an ATM machine or blocked the ATM with their bodies.
- No interviewees mentioned that the actual behavior is affected by an installed plastic shield over the PIN pad. They did not mention the presence of these, nor whether these were considered helpful and/or effective.
- Physical hindrance was not mentioned by our subjects. This

was identified as factor influencing shielding likelihood by De Luca *et al.* [10] during their observations.

- In Germany, of the seven people we interviewed, six mentioned PIN shielding in their initial descriptions of what they did in the two scenarios. In the United Kingdom, and particularly in Sweden, interviewees explicitly distinguished between ATM withdrawal and payment scenarios in this respect.

4. DISCUSSION

Our study replicated and extended one particular aspect of De Luca *et al.*'s ATM study. We focused primarily on the PIN entry aspects of the original study, and then extended the study to different card usage environments.

4.1 Country Differences for Payment Scenarios

The interesting differences here are *firstly* that there was almost no difference in shielding between withdrawal and payment transactions in Germany. The *second* interesting finding was that no subjects in Sweden shielded during payment transactions. The *third* is the difference in payment shielding between the three countries.

A number of explanations can be advanced for these relative outliers. In the first place, there might be significant differences in the frequency of card use and the amount of money involved in each transaction. The Swedish population uses their cards to pay far more than the German population at large (Table 1). Thus, in Sweden, paying by card seems to be *de rigeur* i.e. nothing out of the ordinary requiring special attentiveness.

Moreover, there is also a difference in amounts paid using cards. In Germany, the average amount is more than twice that of Sweden, while the amount in the United Kingdom is in-between the German and Swedish averages. Hence the risk associated with the transactions is greater in Germany, and subjects might well be behaving in accordance with heightened risk perceptions. The *status quo* might well change over the next few years as Germany, for example, has recently introduced PIN-less payments for amounts less than €30.

These numbers accord with our insights from the follow-up interviews: The number of payment instances (both paying oneself using Chip & PIN, as well as observing others doing so) make people less likely to shield. The extreme observations (no one shielding in the payment scenario) in Sweden might also be due to the high level of trust and transparency in Swedish society [32].

4.2 Differences Between Payment and ATM Withdrawal

Our United Kingdom and Swedish subjects were more likely to shield their PINs when withdrawing money than when paying. The following findings from our follow-up interviews suggest explanations for this:

- In one scenario, people receive **cash in hand**, and for the other the transfer of money happened invisibly. People associate security measures with cash and therefore are more likely to shield in the withdrawing scenarios, as compared to the payment scenarios. Similar findings can be found in the literature: Bijleveld and Aarts [5] explain that "*Money [...] activates knowledge structures that are incompatible with the pursuit of social harmony*" [5, page 16]. Related to this is also the following finding from the literature: There is a substantial difference in terms of goal satisfaction. As opposed to obtaining cash, the primary goal of *buying* something is to obtain an object or experience. The underlying purpose is to maximize happiness [8]. Money becomes a secondary concern, a mere facilitator.

- People say they are more likely to shield during withdrawals because ATMs often display reminders to shield. There is more space to place a sticker or to display the reminder on the screen.

- People perceive ATM environments as being more risky than payment contexts in supermarkets, coffee bars, or restaurants. This explanation suggests the existence of misunderstandings or a lack of awareness of the full range of attack vectors. In both cases, there is a risk of manipulated devices and cameras recording PINs, without a human needing to be anywhere near the person using the card.

- People are more 'alone' at ATMs, thus social awkwardness, which would perhaps prevent them from shielding, is less of an issue.

4.3 Comparing to De Luca's ATM withdrawals

We replicated De Luca *et al.*'s [10] results with respect to the percentages of people shielding their PINs at ATMs. The explanation for the relatively low percentages might still be the same as those advanced by De Luca *et al.* i.e. lack of awareness of actual attacker tactics and, corresponding misconceptions regarding the effectiveness of the security measures that they currently take (e.g. checking that nobody is loitering close by).

This low number also indicates that effective protection can only be assured when the PIN pad has pre-installed shields that prevent PIN leakage. However, it is still important to ensure usability for a wide range of people, including those with disabilities.

4.4 Impact of being Accompanied

We did not uncover any differences in terms of 'being accompanied by other people' (e.g. friends, relatives) during PIN entry. Both the interviews by De Luca *et al.* [10] and ours may create the impression that being accompanied makes a difference. A number of interviewees mentioned the social awkwardness that arises from shielding when accompanied by friends or acquaintances.

It is worth mentioning that in all three countries only very few observations recorded subjects being accompanied at ATMs. The lack of a finding might be a consequence of the low numbers. On the other hand, it looks as if this situation is not very typical because our interviewees suggested that cash withdrawal is generally a solo activity. While friends often accompany each other at the cash register, they do not, as a rule, join each other at the ATM.

It looks as if those Germans who do shield make a habit of shielding: those who have this habit always shield when using their cards, with no differences between withdrawal and payment transactions. They either always shield, or never shield. The context does not seem to influence them, nor does the presence or absence of any other people around them.

In the United Kingdom and Sweden, people shielding their PINs are already in such a minority that the cultural norm of not shielding might well perpetuate not shielding, even when accompanied.

4.5 Limitations

The observation field study, as well as the follow-up interviews, took place in three medium-sized cities in three European countries. Thus, the results have only limited validity with respect to large European cities, small towns, or cities in other countries.

We observed PIN shielding from some distance to guarantee anonymity and privacy. This comes with some limitations. A subject was counted as having shielded the PIN as soon as this person used his/her hand, or some other object to shield the PIN pad. However, even if they did shield their PIN entry, they might not have entirely prevented observation from some other vantage point than the one taken by our observer. What we recorded was shielding *attempts*, not efficacy. Moreover, by only observing whether or not someone acted to shield PIN entry, we did not record other protective activities such as checking for surveillance cameras or ensuring that nobody in the vicinity was trying to observe their PIN entry. It might be that subjects did engage in some situational awareness activities and made a perfectly valid low risk assessment. Even if they did realize that someone was close enough to observe their PIN, they might well have interposed their own body between that person and the PIN pad. These kinds of precautions might have been effective in a pre-surveillance era, but with cameras in inner cities, and especially at ATMs, recording people all the time, such precautions are less than effective.

We compared our results to De Luca *et al.*'s. We were able to replicate their results with respect to people shielding their PINs at ATMs. We are aware that the criteria we used for shielding might be slightly different in the two studies because the observers were different people. Yet we did attempt to replicate the study as exactly as possible, based on the information reported in the paper.

The follow-up interview responses might have elicited social desirability responses, but this is a common issue for any interview situation in the security context. We tried to address this limitation by commencing with an innocuous question asking them to detail their own actions in withdrawing and paying with their Chip & PIN cards. Only after this did we focus on the real issue, i.e. shielding, exploring their perspectives on the need for this action.

5. RELATED WORK

We set out to replicate De Luca *et al.*'s study and another two researchers also recently carried out a non-participant observation study of PIN-related behaviors at ATMs. Ashby and Thorpe [2] observed people entering their PINs at a number of different locations of one bank's ATM machines in London. They focused on "hot spot" areas, those where ATM crimes were highest in the London area. Their study revealed that 47% of subjects attempted to cover the PIN pad when they entered their PINs. Unlike our study, they observed only ATM usage, and only in one country. The higher shielding percentages might well be due to the fact that they focused specifically on crime hotspots. This intuition gains some confirmation from the fact that when they interviewed a subset of their observed subjects, and asked them what kinds of precautions they took, the most common one was to use only ATMs that were in safe areas.

A number of papers exist on usable security ATM research. For example, in [9], the authors studied and discussed the idea of biometric authentication at ATMs. Their research revealed a number of non-trivial issues with the introduction of this type authentication for ATMs. Little [24] examined the influence of external factors on ATM use in general. Privacy was one identified factor that aligns with our findings.

Other observation studies of visibly revealed security-related behaviors appear in the research literature. Von Zezschwitz *et al.* [41] studied real-world behavior related to Android authentication patterns. This helped them to compare the real life usability of these patterns to the more traditional PINs. Machuletz *et al.* [26] observed people working in public, to see how prevalent webcam covering behavior was. Greig *et al.* [18] carried out an observation study of one particular branch of a chain store to monitor security-related behaviors. Despite regular information security training and general awareness, they observed passwords written on blackboards, sharing of credentials and staff taking photos of till screens.

Other researchers left USB sticks lying around to see how many people would plug them in [40]. The visible behavior, here, was plugging the stick in the USB port of a PC, and half of their subjects did so. A number of researchers have proposed sending out fake phishing messages to employees to test actual resilience after phishing awareness training [20, 21, 35]. These kinds of exercises seem to be becoming popular in industry [15, 30]. Finally, Forget *et al.* [17] propose a security behavior observation infrastructure to effect long-term monitoring of user behaviors on client machines and Lévesque *et al.* [22], along the same lines, propose a methodology for a field study of anti-malware software.

6. CONCLUSION & FUTURE WORK

We carried out a field study, during which we observed 930 Chip & PIN card uses, in three countries and in different scenarios, with people either withdrawing or paying. There were significant differences with respect to the scenario, with people shielding their PINs significantly less often when they withdrew cash than when they paid in two of the countries (the United Kingdom and Sweden). In Germany, shielding occurrence was equal in both situations. In addition, we carried out interviews to identify factors that may explain what we observed. These include habituation, lack of reminders, the influence of cash in hand, and a lack of awareness of actual attack scenarios.

In general, the percentage of people shielding is surprisingly low, given that they could lose their hard earned money. We were able to confirm De Luca *et al.*'s findings with respect to the low percentage of people shielding their PINs when withdrawing money at ATMs.

Based on our findings, a number of interesting future research questions emerged:

- What influence does the amount of money that someone withdraws or pays have on the decision to shield? In Germany and the United Kingdom / Sweden the average transaction amounts are different. Is this one possible explanation for the identified differences between these countries?
- Based on the identified explanations, it would also be advisable to study the influence of the type and size of plastic shields over PIN pads, and the actual impact of reminders, as also suggested by [2].
- What is the influence of the actual/perceived liability of cardholders on the decision to shield? Initial investigations suggest that mixed messages are sent by different banks we contacted and information provided on their websites. We informally polled a number of people in our respective countries about their understanding and discovered that people have different ideas about whether, and what types of, consequences they might have to face if their PIN is covertly observed and their card subsequently stolen without their knowledge.

- In the interviews, the scenario of the cashier or waitress putting the card into the device for the payee emerged. This is an additional scenario to study as future work. A similar extension would be to study behavior at ticket machines, which are somehow inbetween pure withdrawal (ATM) and the usual store payment scenarios.

However, the long-term goal must be to replace existing devices with those that have opaque shields pre-installed. In the meanwhile, another area of future work could be awareness raising, because one of the findings that emerged from the interviews was that people were not aware of the surveillance camera attack scenarios. They tend to rely on their innate yet inaccurate sense that humans are the greatest threat in these scenarios.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the Competence Center for Applied Security Technology (KASTEL), the Center for Research in Security and Privacy (CRISP), and has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 675730.

Four of the authors changed institutions between the time this research was started and this publication. When the research was started:

- Melanie Volkamer, Andreas Gutmann, and Peter Mayer were employed by the Technische Universität Darmstadt in Germany.
- Karen Renaud was employed by the University of Glasgow in Scotland.

7. REFERENCES

[1] American Psychological Association. Ethical Principles of Psychologists and Code of Conduct. http://www.apa.org/ethics/code/ (Accessed: 20 May 2018).

[2] M. P. Ashby and A. Thorpe. Self-guardianship at automated teller machines. *Crime Prevention and Community Safety*, 19(1):1–16, 2017.

[3] G. Baltistan. Rising number of ATM-skimming frauds must not be taken lightly, 2018. 17 January http://gbherald.com/index.php/2018/01/17/rising-number-of-atm-skimming-frauds-must-not\-be-taken-lightly/ (Accessed: 15 February 2018).

[4] B. Bátiz-Lazo and R. J. Reid. Evidence from the patent record on the development of cash dispensing technology. In *History of Telecommunications Conference, 2008. HISTELCON 2008. IEEE*, pages 110–114. IEEE, 2008.

[5] E. Bijleveld and H. Aarts. A psychological perspective on money. In *The Psychological Science of Money*, pages 3–19. Springer, 2014.

[6] British Sociological Association. Statement of ethical practice, 2017. www.britsoc.co.uk (Accessed: 20 May 2018).

[7] B. L. Brown, S. B. Hendrix, D. W. Hedges, and T. B. Smith. *Multivariate analysis for the biobehavioral and social sciences: a graphical approach*. John Wiley & Sons, 2011.

[8] T. J. Carter. The psychological science of spending money. In *The Psychological Science of Money*, pages 213–242. Springer, 2014.

[9] L. Coventry, A. De Angeli, and G. Johnson. Usability and Biometric Verification at the ATM Interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 153–160, New York, NY, USA, 2003. ACM.

[10] A. De Luca, M. Langheinrich, and H. Hussmann. Towards understanding ATM security: a field study of real world ATM use. In *Proceedings of the Sixth Symposium on Usable Privacy and Security (SOUPS)*, page 16, San Fransisco, 2010. ACM.

[11] Department of Sustainability and Environment. Effective engagement: building relationships with community and other stakeholders, 2005. The Community Engagement Network Resource and Regional Services Division Victorian Government Department of Sustainability and Environment. Book 3: The Engagement Toolkit. Retrieved from http://www.dse.vic.gov.au/__data/assets/pdf_file/0003/105825/Book_3_-_The_Engagement_Toolkit.pdf (Accessed: 20 May 2018).

[12] Deutsche Bundesbank. Payment behaviour in Germany in 2014. https://www.bundesbank.de/Redaktion/EN/Downloads/Publications/Studies/payment_behaviour_in_germany_in_2014.pdf, 2015. (Accessed: 15 February 2018).

[13] European Central Bank. Payment statistics. http://sdw.ecb.europa.eu/reports.do?node=1000004051, October 2015. (Accessed: 15 February 2018).

[14] D. S. Evans, K. Webster, G. K. Colgan, and S. R. Murray. Paying with cash: A multi-country analysis of the past and future of the use of cash for payments by consumers. *Available at SSRN 2273192*, 2013.

[15] J. Eyers. Banks test staff with cyber security 'fire drills'. http://www.afr.com/technology/banks-test-staff-with-cyber-security-fire-drills-20160914-grg2e8. (Accessed: 15 February 2018).

[16] R. J. Fisher. Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, 20(2):303–315, 1993.

[17] A. Forget, S. Komanduri, A. Acquisti, N. Christin, L. F. Cranor, and R. Telang. Security behavior observatory: Infrastructure for long-term monitoring of client machines. Technical report, Carnegie-Mellon University Pittsburgh PA United States (CMU-CyLab-14-009), 2014.

[18] A. Greig, K. Renaud, and S. Flowerday. An ethnographic study to assess the enactment of information security culture in a retail store. In *2015 World Congress on Internet Security (WorldCIS)*, pages 61–66, Dublin, 2015. IEEE.

[19] R. Iphofen. Research ethics in ethnography/anthropology, 2013. European Commission http://ec.europa.eu/research/participants/data/ref/h2020/other/hi/ethics-guide-ethnog-anthrop_en.pdf.

[20] K. Jansson and R. von Solms. Phishing for phishing awareness. *Behaviour & Information Technology*, 32(6):584–593, 2013.

[21] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M. A. Blair, and T. Pham. School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, page 3. ACM, 2009.

[22] F. L. Lévesque, C. R. Davis, J. M. Fernandez, S. Chiasson, and A. Somayaji. Methodology for a field study of

anti-malware software. In *International Conference on Financial Cryptography and Data Security*, pages 80–85. Springer, 2012.

[23] A. Lindesmith, A. Strauss, and N. Renzin. *Social Psychology*. New York: Holt, 1975.

[24] L. Little. Attitudes Towards Technology Use in Public Zones: The Influence of External Factors on ATM Use. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '03, pages 990–991, New York, NY, USA, 2003. ACM.

[25] F. Liu and S. Maitlis. Nonparticipant observation. In A. J. Mills, G. Durepos, and E. Wiebe, editors, *Encyclopedia of Case Study Research*, pages 610–612. Thousand Oaks, CA: SAGE Publications, 2010.

[26] D. Machuletz, H. Sendt, S. Laube, and R. Böhme. Users protect their privacy if they can: Determinants of webcam covering behavior. In *EuroUSEC*, Darmstadt, Germany, July 2016. Internet Society.

[27] K. M. MacQueen, E. McLellan-Lemal, K. Bartholow, and B. Milstein. Team-based codebook development: structure, process, and agreement. *Handbook for Team-Based Qualitative Research*, pages 119–135, 2008.

[28] M. B. Miles and A. M. Huberman. *Qualitative data analysis: An expanded sourcebook*. Sage, 1994.

[29] E. Murphy and R. Dingwall. Informed consent, anticipatory regulation and ethnographic practice. *Social Science & Medicine*, 65(11):2223–2234, 2007.

[30] D. Pauli. Go phish your own staff: Dev builds open-source fool-testing tool. `http://www.theregister.co.uk/2016/02/04/no_more_excuses_dev_builds_dead_easy_open_source_antiphishing_app/`, 2016. (Accessed: 15 February 2018).

[31] M. Petticrew, S. Semple, S. Hilton, K. S. Creely, D. Eadie, D. Ritchie, C. Ferrell, Y. Christopher, and F. Hurley. Covert observation in practice: lessons from the evaluation of the prohibition of smoking in public places in Scotland. *BMC Public Health*, 7(1):204, 2007.

[32] N. Sanandaji. Trust not taxes have made Sweden a success, 2015. `https://www.thelocal.se/20150711/trust-not-high-taxes-have-made-sweden-a-\success-opinion` 11 July (Accessed: 21 May 2018).

[33] B. Segendorf and A.-L. Wretman. The Swedish retail payment market. *Sveriges Riksbank Economic Review*, (3):48–68, 2015.

[34] P. Spicker. Research without consent. *Social Research Update*, 51:1–4, 2007.

[35] T. Steyn, H. A. Kruger, and L. Drevin. Identity theft. empirical evidence from a phishing exercise. In *IFIP International Information Security Conference*, pages 193–203. Springer, 2007.

[36] R. M. Sutton and S. Farrall. Gender, socially desirable responding and the fear of crime: Are women really more anxious about crime? *British Journal of Criminology*, 45(2):212–224, 2004.

[37] Technische Universität Darmstadt. Webpage of the Ethics Commission of the Technische Universität Darmstadt. `https://www.intern.tu-darmstadt.de/gremien/ethikkommisson/zustndigkeit/zustndigkeit.en.jsp` (Accessed: 20 May 2018).

[38] The Statistical Office of the European Union. Population on 1 January (tps00001). `http://ec.europa.eu/eurostat/tgm/table.do?language=en&pcode=tps00001`, 2016. (Accessed: 15 February 2018).

[39] The UK Cards Association. UK Card Payments 2015. `http://www.theukcardsassociation.org.uk/wm_documents/UK%20Card%20Payments%202015%20taster%20for%20website.pdf`, 2015. (Accessed: 15 February 2018).

[40] M. Tischer, Z. Durumeric, S. Foster, S. Duan, A. Mori, E. Bursztein, and M. Bailey. Users Really Do Plug in USB Drives They Find. In *IEEE Symposium on Security and Privacy*, pages 306–319, Fairmont, San José, CA, MAY 23-25 2016. IEEE.

[41] E. Von Zezschwitz, P. Dunphy, and A. De Luca. Patterns in the wild: a field study of the usability of pattern and pin-based authentication on mobile devices. In *Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 261–270. ACM, 2013.

[42] C. Weston, T. Gandell, J. Beauchamp, L. McAlpine, C. Wiseman, and C. Beauchamp. Analyzing interview data: The development and evolution of a coding system. *Qualitative Sociology*, 24(3):381–400, 2001.

User Behaviors and Attitudes Under Password Expiration Policies

Hana Habib, Pardis Emami-Naeini, Summer Devlin†, Maggie Oates, Chelse Swoopes,
Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor
Carnegie Mellon University University of California, Berkeley (†)
{htq, pemamina, moates, cswoopes, lbauer, nicolasc, lorrie}@andrew.cmu.edu
devlins@berkeley.edu†

ABSTRACT

Policies that require employees to update their passwords regularly have become common at universities and government organizations. However, prior work has suggested that forced password expiration might have limited security benefits, or could even cause harm. For example, users might react to forced password expiration by picking easy-to-guess passwords or reusing passwords from other accounts. We conducted two surveys on Mechanical Turk through which we examined people's self-reported behaviors in using and updating workplace passwords, and their attitudes toward four previously studied password-management behaviors, including periodic password changes. Our findings suggest that forced password expiration might not have some of the negative effects that were feared nor positive ones that were hoped for. In particular, our results indicate that participants forced to change passwords did not resort to behaviors that would significantly decrease password security; on the other hand, their self-reported strategies for creating replacement passwords suggest that those passwords were no stronger than the ones they replaced. We also found that repeating security advice causes users to internalize it, even if evidence supporting the advice is scant. Our participants overwhelmingly reported that periodically changing passwords was important for account security, though not as important as other factors that have been more convincingly shown to influence password strength.

1. INTRODUCTION

Passwords are widely used for authentication, from individual online accounts to organizational access control. It is well known that people create passwords that are easily guessed [22, 37], and engage in insecure practices, such as reusing passwords across different accounts [7, 9, 32, 37]. Prior research has focused on helping users make stronger passwords through password-composition policies (e.g., [20]), which require users to include a defined number of characters and character classes in their passwords, and understanding the impact of password blacklists (e.g., [38]), which prevent users from creating passwords that are too common. The purpose of these password security tools is to help users create passwords that are less vulnerable to automated password guessing.

Historically, password expiration policies have been implemented to help prevent password guessing attacks [31]. At the time these policies were first proposed, computational power was far scarcer than it is now and a successful password cracking attack would have taken several months. Thus, changing passwords every month may have seemed to be a reasonable method for defeating such an attack [31]. Furthermore, password expiration could act as a failsafe mechanism to eventually lock out attackers who may have gained access to a legitimate user's password without their knowledge. As a result of those desirable properties, expiration policies, of varying duration, have become widespread practice, especially for university and government systems [10].

Research has demonstrated that given modern computing capabilities, expiration policies may have limited utility for organizational security, largely due to the predictability of human behavior in password management [3, 39]. Though it is known that people struggle to handle the demands of password management, we question the intuition that expiration policies lead users to choose simpler passwords than their existing ones or reuse passwords from other accounts at a greater rate. Our study complements a survey conducted by the U.S. National Institute of Standards and Technology (NIST) exploring the steps users actually take when they are forced to change their password [4]. We build on this prior work, which analyzed password behaviors of participants from a single U.S. government organization, by surveying participants from numerous and diverse workplaces from across the U.S., who face a variety of different organizational password policies and requirements. Additionally, we analyze how reported coping strategies differ for those who face more frequent expiration. We also contribute additional user perspectives related to expiration, such as how people prioritize password changes among other password-management practices.

Our results are largely consistent with those found in NIST's study [4], and suggest that despite users generally employing harmful password practices, frequent password changes do not lead to some of the negative security effects thought to be introduced by expiration policies. Based on their self-reported behaviors, we found that participants did not create passwords that are simpler than the ones they already

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

use or reuse passwords from other accounts at a higher rate. Though expiration policies do not appear to increase the incidence of account lockouts or lead users to change their password-recall strategies, participants reported relying on coping mechanisms, such as appending digits to their previous password, to update their password. Such coping mechanisms greatly reduce the potential security gains brought by expiration policies.

In general, our participants reported that password expiration had a positive impact on security, with 82% agreeing that it made it less likely that an unauthorized person will log in to their account. However, changing passwords periodically was thought to be less important for account security than creating a complex password, storing the password safely, and avoiding password reuse. This is in line with modern security guidance, such as the recent changes to the NIST authentication guidelines [12], which recommend against password expiration policies. With the additional insights gained in this study, it is evident that users accept and adapt to the security advice they are provided, especially if they hear it repeatedly from a trusted source, such as their employer's IT department. This suggests that, if communicated appropriately, users may be open to more updated recommendations, such as using password managers or enabling two-factor authentication.

In the remainder of this paper we first discuss literature relevant to our study. We then describe the study design and methodology used in analyzing the collected data. Next, we present our findings regarding password usage at work, update behavior, impact of different expiration policies, and security perceptions related to password expiration. Finally, we conclude with a discussion of our results.

2. RELATED WORK

There is a large body of literature pertaining to various aspects of password authentication. We discuss the prior work that is most relevant to our study, such as those examining password management, challenges due to password expiration, or security perceptions related to passwords. Our work builds upon this existing literature by analyzing what strategies people use to cope with password management, including password updates, and how they generally feel toward periodic password changes.

2.1 Password-Management Strategies

Users face considerable burdens in managing passwords. Previous research has found that people use over 20 passwords in their daily lives [9, 27]. A diary study conducted by Grawemeyer and Johnson observed that, on average, their participants logged into various accounts over 45 times in one week [13]. Authentications for work activities accounted for 43% of all logins in their sample, highlighting the importance of studying workplace password management behaviors in particular.

Prior work has also shown that people have varying strategies for selecting passwords [32, 34]. One common strategy for coping with multiple passwords is to reuse passwords across different accounts [7,9,32,34,37]. In a 154-participant empirical study of password usage, Pearman et al. observed that participants exactly reused passwords for 67% of their accounts and had passwords containing a string of at least four characters in common for 79% of their accounts [27]. The more passwords a user has created, the more likely they are to reuse passwords [11]. Previous research has also found that users attempt to match password strength to the relative importance of the account when selecting passwords [25, 34]. Stobert and Biddle further observed in an interview study that their participants rarely changed passwords on their own, and only did so in the case of a breach or forgotten password [32]. This literature motivates our research, which aims to understand how people cope with forced password changes in addition to the normal demands of password management.

Users also differ in how they recall their passwords, typically relying on their memory [11, 13, 32]. However, writing down at least some account passwords is also common practice [32]. Previous research has found the adoption of password managers to be low [16], even though they are widely recommended for password security [29]. Building upon this literature, our work tries to identify whether password recall, a major usability factor related to password use, is impacted by password expiration.

2.2 Password Expiration Challenges

In an empirical study of the password policies of 75 different websites, Florêncio and Herley found that 20% of the websites they examined required participants to update their password regularly [10]. Prior literature has shown that required password changes have negative implications for usability. Shay et al. found that only 30% of their survey participants created an entirely new password when forced to change their university password and 19% had issues recalling their new password [30]. Other user issues related to required password changes include being reminded to change a password too early, difficulty keeping track of updated passwords, struggling to create passwords that meet the institution's password requirements, and fear of being locked out of an account [8, 14].

A major security issue related to password expiration is the tendency for people to make predictable changes when updating their password, which can be exploited to optimize password-cracking attacks [1]. Zhang et al. developed a transform-based password-cracking algorithm, using password history data for 7,700 accounts at their institution. With the knowledge of the accounts' previous passwords, they were successful in guessing 41% of passwords in an offline attack and 17% in an online attack (allowing for a maximum of five guesses). Thus, they demonstrated that password expiration seems to have limited utility for locking out attackers who have already gained knowledge of a user's password [39]. Chiasson and van Oorschot further demonstrated that with modern computing capabilities and taking into account human behavior in password creation, it is no longer feasible to change passwords faster than they can be potentially cracked [3].

Most related to our study, a survey conducted by NIST explored password-management behaviors of 4,573 Department of Commerce (DOC) employees who had, on average, nine work-related passwords [4]. The authors estimated that DOC employees spent 12.4 hours per year changing passwords on a 90-day expiration schedule, or 18.6 hours changing passwords on a 60-day expiration schedule. The study also revealed that most employees coped with the burden of

password changes by making minor changes to their existing password. The authors found that positive attitudes toward password requirements (i.e., composition and expiration requirements) correlated with more secure behaviors and fewer usability problems. We build on this prior work by studying a population that includes users from a wider variety of workplaces with differing password policies. We additionally explore perceptions about expiration independently of other password requirements. Furthermore, we analyze usability patterns more deeply, such as the correlation between creation and update strategies, and whether certain techniques are associated with more frequent account lockouts.

More recent security recommendations have been moving away from password expiration policies [5, 6, 24, 40]. The NIST Special Publication 800-63B, Digital Identity Guidelines, was recently revised and now recommends that "Verifiers should not require memorized secrets to be changed arbitrarily (e.g., periodically)" and that they should only be changed in special circumstances, such as when there is a compromise of passwords [12]. However, the NIST standards are only required for U.S. government systems. Other prominent security standards, including the Payment Card Industry (PCI) Data Security Standards and ISO/IEC 27002, still recommend regular password changes [15,26]. Our work provides insight into the security mindset of workplace password users that can be used to inform future institutional password security recommendations.

2.3 Perceptions of Password Security

Previous research has also studied perceptions related to the security of passwords. In two separate studies, Ur et al. collected users' perceptions of password strength. They discovered that participants had some misunderstandings about what makes a password secure, including thinking that adding digits made their passwords stronger than it really did and that keyboard patterns and common phrases were more random than they actually are [33,34]. This often meant that users created passwords that did not match their desired security level, for example, creating weak passwords for highly valued accounts [33]. We expand on this work by evaluating user perceptions related to several password practices, instead of only password composition.

Furthermore, researchers have discovered that there is a disconnect between what people believe is beneficial to password security and what they actually do. Riley found a number of behaviors, such as changing passwords for accounts or using special characters, that the majority of their participants believed they should engage in, but did not do so [28]. Our study looks into perceptions about similar behaviors, but aims to understand the perceived relative importance of these behaviors.

In a survey comparing the security practices of experts and non-experts, Ion et al. reported that non-experts recommended using anti-virus software, creating strong passwords, visiting only known websites, and changing passwords frequently to stay secure online. Non-experts and experts both perceived using strong and unique passwords as effective security mechanisms and reported that they would be likely to follow those practices. Not writing down passwords was considered somewhat effective, while saving passwords in a file, using a password manager, and writing down passwords were considered ineffective security advice [16]. Through our work, we attempt to gain a deeper understanding of these perceptions in the context of workplace passwords.

Prior work has found that people also have misconceptions about protecting against different threats [32], often overestimating the threat of a targeted attack and underestimating that of automated guessing attacks [33]. For example, Gaw and Felten found that participants viewed password complexity and randomness as means to reduce human gussability and not necessarily as protection against an automated attack. Participants also viewed friends (and others close to them) as the most capable attackers, while hackers were perceived as the most motivated [11]. We evaluate the threats people consider in managing workplace passwords, and the role expiration policies play in these perceptions.

3. METHODOLOGY

In this study, we analyzed data collected from two separate online surveys. The first survey focused on people's workplace password habits, while the second measured perceptions of several password practices, including periodic password changes. We used both quantitative and qualitative methods to analyze data collected from the surveys.

3.1 Data Collection

In this section we describe the procedures for collecting our survey data. Both surveys were approved by our Institutional Review Board (IRB) and were conducted on Amazon's Mechanical Turk[1]. Participants were age 18 or older, residents of the United States, and had a HIT approval rate of over 90%.

3.1.1 Workplace Passwords Survey

The first survey in our study, which will be referred to as the ***workplace passwords survey***, was implemented as a screening survey followed by a full survey about participants' experiences with their workplace passwords. We implemented a screening survey to ensure that only participants who had at least one workplace password were allowed to answer questions in the full survey, as questions about a main workplace password would be irrelevant to those with no workplace passwords.

In the screening survey, participants answered a total of six questions that asked how many workplace passwords they have and their age, gender, ethnicity, education, and occupation. Those who met the qualification criteria of having at least one workplace password were contacted through Mechanical Turk about completing a "bonus survey," which was the full survey about workplace password habits. Questions included in the screening survey are in Appendix A.

The full survey was designed to ask participants about their experiences with their main workplace account and included 31 multiple-choice and five open-ended-response questions. With these questions we explored workplace password habits, such as experiences creating, updating, and recalling passwords, as well as sentiments toward password expiration. In this survey, we confirmed the four demographic attributes participants provided to us in the screening survey. We also included an attention-check question that was a duplicate of

[1] Amazon's Mechanical Turk. `https://www.mturk.com`

the question asking participants how many workplace passwords they have. The full survey is provided in Appendix B.

A total of 618 people submitted the screening survey and 407 finished the full survey. Participants were compensated $0.25 for the screening survey and $2.00 for the full survey. On average, participants finished the screening survey in about two minutes and the full survey in 10 minutes.

3.1.2 Password Perceptions Survey

The second survey we conducted, which will be referred to as the ***password perceptions survey***, explored people's perceptions of the relative importance of four password practices: *using a complex password*, *storing the password in a safe place*, *creating a password that you do not already use somewhere else*, and *periodically changing passwords*. In the survey, participants rated the importance of each of these practices for account security on a five-point Likert scale, and completed open-ended responses explaining their ratings. We also asked participants to rank failure to adhere to each practice (e.g., using a simple password) in order of harm to account security. Participants were then shown pairs of the practices and then were asked to indicate whether one contributes more to account security than the other. Lastly, we asked participants about their anticipated behaviors in a hypothetical scenario in which their workplace implemented or removed an expiration policy (depending on the participant's current workplace policy). The order of the four password practices was randomized in each section to avoid biasing participants based on how the practices were presented. Appendix C contains the questions in this survey.

People who completed the workplace passwords survey were disqualified from taking this survey. The password perceptions survey was completed by 340 eligible participants who were compensated $1.50. On average, participants completed the survey in about 10 minutes.

3.2 Data Analyses

This section describes the statistical tests and qualitative methods used in analyzing the collected data. Data from the two surveys were analyzed separately.

3.2.1 Quantitative Analyses

Prior to running statistical tests, we excluded participants with inconsistent or obviously fraudulent responses to improve the validity of our analyses. For the analyses of data from workplace passwords survey we excluded 49 participants who answered the attention-check question inconsistently, one participant who reported that they did not change their main workplace password (even though they reported that they were required to change all of their workplace passwords), and one participant who selected every answer option for all questions where participants could select multiple options. It is possible that the attention-check question may have led to the exclusion of participants who simply misremembered their workplace passwords, and not just those who truly were not paying attention to the survey. We excluded only one participant from the analyses of data from the passwords perceptions survey as they used the same unintelligible response for each of the open-ended questions. Thus, 356 responses from the workplace passwords survey and 339 from the password perceptions survey were included in our analyses.

We conducted several different statistical tests and used significance level $\alpha = .05$ in our analyses. For categorical data, we used Pearson's chi-squared tests to determine the independence of two nominal variables, or Fisher's exact tests if counts in the contingency table were below five. For tests in which we were examining the impact of expiration frequency, we binned policies into three expiration periods: less than or equal to every 30 days, every 60 days, or greater than or equal to every 90 days. We report the phi coefficient (ϕ) to understand the effect size of the associations found for two binary variables, or Cramer's V (V) if the variables have more than two levels. Both measures are reported on a scale from -1 to 1, such that 1 demonstrates a complete positive association and -1 demonstrates a complete negative association between two variables. We report only statistical results for which we observed at least a small effect (demonstrated by an association of at least .1), which is a recognized threshold for statistical reporting [23].

To analyze data with a categorical independent variable and ordinal dependent variable, such as Likert-scale data, we used Kruskal-Wallis tests. We conducted a Friedman test to test the null hypothesis that password practices were rated as equally important. We also ran one-sample, two-sided Wilcox Signed-Ranked tests to determine whether participants felt one practice contributes more to account security than another by coding the rating options from -3 (left contributes much more) to 3 (right contributes much more), and testing the null hypothesis that the practices have equal contribution (a rating of 0).

In order to evaluate the impact of demographics on the use of password-creation, update, and recall techniques, we ran binomial logistic regressions where the independent variables were age, race, education level, and technical expertise, and the dependent variable was whether or not a certain technique was used. We ran binomial logistic regressions to determine whether password-creation techniques were predictors of update techniques, where the independent variables were one of 17 password-creation techniques (represented as binary variables) and the dependent variable was a password-update technique. For each significant factor found in the regressions, we followed up with chi-squared tests to determine the strength of the association between the factor and the dependent variable.

To analyze whether our participants used certain techniques for password memorability and others to make their password stronger, we ran a multinomial logistic regression. The dependent variable was a nominal variable with four levels: whether the update technique was used for making a stronger password, making the password easier to remember, both security and memorability, or neither security nor memorability. The baseline for the regression was set to neither security nor memorability. The independent variables were the password-update strategies measured in the survey.

3.2.2 Qualitative Analyses

Our surveys collected several open-ended responses which were each systematically analyzed to extract major themes. For each question, one researcher first developed a codebook based on common themes occurring in a sample of 20 responses. Two researchers then coded a random sample of 20% of the responses based on the first iteration of the

codebook. The researchers then reviewed their conflicts and revised the codebook accordingly. If agreement between the two coders, measured by Cohen's kappa, was less than $\kappa = .70$, a recognized acceptable threshold for agreement [36], both researchers recoded the sample and revised the codebook until reaching sufficient agreement. After successfully converging on the samples, one researcher would code the remaining responses for that question. Both researchers coded the full set of data collected for opinions on the impact of password expiration policies and reasons for continuing password changes. For the remaining qualitative data, the two researchers reached $\kappa = .81$ agreement, averaged over all questions. The statistics from qualitative responses reported in this paper are derived from the researchers' coding of the full set of responses.

3.3 Limitations

One of the major limitations of our study is that we recruited participants from Mechanical Turk. Though our participants come from a well-studied convenience sample, they may not reflect the behaviors and attitudes of the general population. Moreover, Mechanical Turk participants have been shown to be more privacy-sensitive than the population at large [17]. However, Mechanical Turk has proven to be a source of high-quality human subjects data [18], and has been successfully used in numerous studies related to passwords (e.g., [16, 20, 33]). Only 6% of our expiration survey population reported that Mechanical Turk was their primary occupation, indicating that the vast majority of participants were reporting on passwords for a different workplace.

Additionally, our study uses self-reported data about participants' past behavior, which participants may not have remembered or reported accurately. The effects of this may have been exaggerated by the privacy paradox, a well-studied observation that people's privacy attitudes often differ from their actual behaviors [19]. It is possible that our participants' reported reactions and attitudes toward password expiration may be different from their actual behaviors when facing their own expiration policies. While our data may be impacted by these limitations, we believe that our study is still a step forward in understanding people's general behaviors and attitudes related to password expiration.

4. RESULTS

Our surveys included questions about how people create, update, and manage their workplace passwords, as well as their attitudes toward password expiration in relation to other password-management practices. Similarly to participants in NIST's study [4], our participants generally reported coping with their expiration policy by modifying their current password, suggesting that updated passwords may not be any stronger or weaker than the ones originally created.

We also found that self-reported behaviors and attitudes related to expiration were largely independent from the presence and frequency of an expiration policy. Participants viewed password changes as important for account security, but felt that other password-management practices, such as using a complex password, storing the password safely, and avoiding password reuse were more vital. Our results indicate that while people may buy into security advice, they are sometimes unable or unwilling to act on the advice in a way that significantly improves password security.

4.1 Password Usage at Work

In this section we describe expiration policies reported by our participants and their password strategies for managing their main workplace password.

4.1.1 Password Expiration Policies

In total we analyzed data collected from 695 participants. The demographics of our participants are described in Table 1. On average, participants in both surveys reported having between three and four workplace passwords. 51% of participants in the workplace password survey reported that they were required to change most or all of their workplace passwords. Figure 1 shows the distribution of our participants' reported password expiration policies. The most common expiration period observed in our samples was expiry every 90 days, reported by 28% of participants in the workplace passwords survey and 19% of participants in the password perceptions survey. A larger percentage of participants in the password perception survey (59%) reported that they did not have an expiration policy for their main workplace password, compared to those in the workplace password survey (26%). It is possible that the wording of the recruitment text and questions in the workplace passwords survey primed participants to think more about expiration and report on their workplace passwords that did expire.

Almost two-thirds of participants (64%) from the password perceptions survey who did not have an expiration policy reported that they changed their workplace password periodically, while a large minority (34%) reported that they never changed it. Those who did change their password primarily mentioned account security in their explanations for doing so, while those who did not change their password most frequently mentioned that they never felt they had a reason to be concerned about the security of their account. In contrast, 53% of participants in a study conducted by Riley did not change their passwords on a regular basis. However, the survey was not specific to workplace passwords [28].

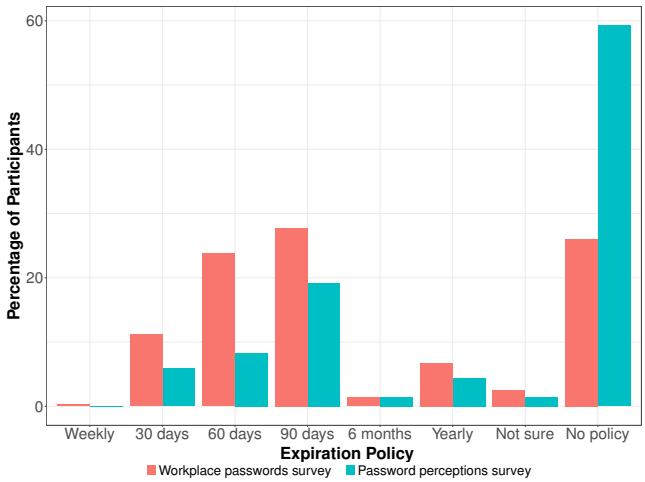


Figure 1: Percentage distribution of participants' workplace password policies. While a larger percentage of participants in the password perceptions survey did not have a workplace expiration policy, note that question wording and recruitment text differed between these two surveys.

Gender	WP	PP	Age	WP	PP	Education	WP	PP	Race	WP	PP
Female	51.7%	45.4%	18-24	6.2%	15.0%	Some high school	.3%	.6%	American/Alaska Native	1.4%	0.0%
Male	47.5%	53.1%	25-34	40.2%	46.3%	High school	7.9%	15.0%	Asian	5.6%	5.6%
Other	.3%	.6%	35-44	29.2%	23.9%	Some college	24.2%	28.3%	Black/African American	7.0%	7.4%
No answer	.6%	.9%	45-54	16.0%	8.3%	Associates	12.9%	13.9%	Hispanic/Latino	5.6%	8.0%
			55-64	7.6%	5.9%	Bachelors	36.0%	33.0%	Non Hispanic	.3%	.3%
			65-74	.6%	.3%	Graduate	17.7%	8.8%	White/Caucasian	77.0%	76.1%
			No answer	.3%	.3%	No answer	1.1%	.3%	Other	1.7%	1.5%
									No answer	1.4%	1.2%

Occupation	WP	PP		WP	PP	Tech Expertise	WP	PP
Business, Management, or Financial	24.2%	9.7%	Medical	6.2%	2.7%	Expert	9.3%	19.2%
Administrative Support	15.4%	10.9%	Mechanical Turk Worker	5.6%	12.7%	Non-Expert	88.5%	80.8%
Education/Science	12.6%	8.3%	Art, Writing, or Journalism	4.5%	7.7%	No answer	2.2%	0.0%
Service	11.5%	11.8%	Other	8.4%	19.8%			
Computer Engineering/IT Professional	9.3%	14.4%	No answer	2.2%	2.1%			

Table 1: Demographic breakdown of our participants from the workplace passwords (WP) survey, and the password perceptions (PP) survey.

4.1.2 Password Creation and Reuse

Our participants reported using common strategies to create their initial passwords, and on average indicated combining three password-creation techniques. The most common self-reported techniques used were using a word in English as part of their password (41% of participants), using a name (37%), and adding numbers (59%) or symbols (32%) to the beginning or end of a word or name. These were also common password-creation strategies observed by Ur et al. [34]. There were some demographic differences in the use of creation strategies. For example, participants ages 45 to 54 years old reported significantly less password reuse (both exact and with modifications) than participants who were 18 to 24 years old (exact: $p = .001, V = .24$, modified: $p = .001, V = .22$). Additionally, those who reported their race as Hispanic or Latino were slightly less likely than white or Caucasian participants to use an English word as their password ($p = .02, \phi = .13$). We also observed that technical participants (those who had ever held a job or received a degree in computer science or any related technology field) were slightly more likely to add symbols to the beginning or end of their password ($p = .05, \phi = .12$) and substitute symbols for letters ($p = .002, \phi = .18$). Participants also reported a moderate amount of password reuse for their main workplace password: 44% said their main workplace password is similar or identical to other work passwords, 57% reported that their main workplace password is very different from their non-workplace passwords.

4.1.3 Password Recall and Lockouts

Our participants had varying strategies for recalling their workplace password. The most common recall technique reported was memorizing the password, which was used by 53% of participants in the workplace passwords survey and 57% of participants in the password perception survey. Other work has also found password memorization to be users' dominant strategy for recalling a password [11, 13, 32]. In both of our surveys, over 85% of password memorizers reported having no backup method for recalling their passwords. Though participants in NIST's study also most frequently recalled their password through memory, over 80% also reported having stored their passwords on paper or electronically [4]. The most common password-storage technique reported by our participants was writing it down on paper, used by 19% of participants in the workplace passwords survey and 10% of participants in the passwords perceptions survey. However, one memorizer reported relying on a "change password" feature as a form of backup for their main workplace account, saying "I have good enough memory and use resetting as a backup." There were some demographic differences observed in the use of recall methods, but none were consistent across the two surveys.

The majority of participants (55%) reported memorizing their main workplace password within the first two times they logged in to their account. However, those who used a password manager ($p < .001, V = .41$) or wrote down their password ($p < .001, V = .22$) were significantly more likely to take more than two logins to memorize their password, on average learning their passwords after three to five logins. A quarter of participants who used a password manager reported that they did not memorize their passwords.

Overall, 45% of participants experienced at least one account lockout in the past year, with 12% reporting three or more lockouts. Participants in NIST's study appeared to face a similar lockout rate, as 48% viewed getting locked out as causing "some" or "a lot" of frustration [4]. Based of their reported lockouts, we found that participants who stored their password in their browser or wrote it down were two to three times more likely to face three or more account lockouts in the past year, while those who memorized their passwords generally faced fewer lockouts. Statistical results for the correlation between recall methods and account lockouts are reported in Table 2.

The most common password-recovery options reported by participants were calling someone on the phone (34%), sending someone an email (31%), and using a website (24%). Ten participants reported using their own method for recovering their main workplace password if they were unable to recall it, such as an encrypted USB drive or a piece of paper that they locked in a safe.

4.2 Password-Update Behavior

In this section we describe the self-reported strategies our participants indicated using during password changes, and

Recall Method	p
Web Browser	.001∗
Encrypted File	.55
Password Manager	.17
Password Protected Computer or Device	.30
Device or Computer Used Only by the Participant	.77
Write Password on Paper	.008∗
Write Reminder for Password	.72
Memorize Password	<.001∗

Table 2: P-values for chi-square tests comparing password-recall methods with the number of account lockouts. All tests had an effect size of $\phi = .24$. Significant results are marked with an asterisk

their reasons for using them. Participants primarily reported modifying their previous password during their last password change, and less than a quarter created one that was completely new.

4.2.1 Most Modify Their Previous Password

Table 3 displays the update strategies participants indicated using during the most recent change of their main workplace password. Participants typically indicated using one or two of these techniques to update their password. 237 (67%) participants reported creating their new password by modifying their previous one. The most common technique reported by our participants to modify the existing password was capitalizing a letter, which was used by 30% of participants. Only 37 (10%) participants reported that they reused passwords from other accounts during their last update, while 162 (46%) of participants reported updating their main workplace password with one that was completely new or using a password generator to create a new one. However, 76 of these participants also selected at least one modification technique with this option. Therefore, we estimate that only 24% of participants updated their password with one that was completely new. Similarly, 68% of participants in NIST's study generated a new password by making a minor change to their old one [4]. However, the NIST study appeared to have a larger degree of password reuse compared to our study population, as 43% of participants generated frequently used passwords by using existing ones.

Some participants described unique approaches for coping with their password expiration policy. For example, one participant reported that they used information from a fast food receipt as their password and used a new receipt to update their password. Demographics also had some impact on the use of update techniques. Most notably, we observed that age had an impact on reusing passwords from other accounts during password updates ($p = .03, V = .20$). On average, only 9% of participants ages 25 to 44 and 4% of those 55 to 64 reported that they reused password from other accounts in their last update, compared to 29% of participants who were 18 to 24 years old.

4.2.2 Creation & Update Strategies Are Consistent

Some initial password-creation strategies were significant predictors for the use of similar update techniques. For example, participants who reported that they used a password generator to *create* passwords were 18 times more likely than those who reported using other password-creation methods to use a generator to *update* their passwords ($p < .001, \phi = .48$). Additionally, those who reported substituting letters with symbols during password creation were seven times more likely than those who did not to report using the same technique to update their password ($p < .001, \phi = .35$). Those who reported using a birthday when creating their main workplace password were four times more likely to report using a date to update it ($p < .001, \phi = .26$). Password reuse was also consistent, as participants who reported exactly reusing a password from another account for their initial password were four times more likely to report reusing a password when updating ($p < .001, \phi = .28$) and those who reported reusing another password with modifications were three times more likely to report reusing a password at update ($p = .002, \phi = .19$). Our participants generally used the same strategy whenever they updated their passwords. In particular, 64% of participants reported using their strategy every time or most of the time when updating their password, while only 4% reported that they use very different techniques each time.

4.2.3 Techniques Associated with More Lockouts

The only update method that had an impact on password memorization was the use of a password generator. Those who reported using a password generator were twice as less likely to report that they memorize their password ($p = .007, \phi = .20$) and seven times more likely to report that they use a password manager to store their password ($p < .001, \phi = .20$), compared to those who did not use one. Two techniques correlated with a higher number of account lockouts. Those who reported that they duplicated characters during their last password update were three times more likely to report that they faced three or more account lockouts in the past year ($p = .005, \phi = .28$) and participants who reported substituting digits or special characters with the same character type were twice as likely to report having three or more lockouts ($p = .04, \phi = .28$), compared to those who did not use these techniques.

4.2.4 Motives & Reminders

Our participants had varying sources and motivations for using their update techniques. 47 of our participants shared where they learned their update strategy. Of these participants, 28% reported learning it from the Internet. Overall, 35% of participants used their strategy because it made their password easier to remember. Reusing a password was largely correlated with using the strategy for memorability ($\beta = 1.13, p = .007$). A quarter of participants reported that they used their strategy because they thought it made their password stronger. Based on the self-reported strategies, we observed that creating a new password ($\beta = 1.32, p < .001$) and using a password generator ($\beta = 2.00, p < .001$) during a password update were correlated with wanting to make the password stronger. Comparatively, memorability was more important to participants in NIST's study, as 81% cited using their password generating strategy so that their password was easy to remember.

We also asked participants about any reminders they receive when their password is about to expire The most common form of password change reminders reported were automated emails and software installed on their computers. We found that the timing of when the last reminder is sent did not have an impact on the effort participants reported

Technique	Example	Responses	%
Modifications			
Capitalizing a character	candy# → candY#	108	30.3%
Incrementing a character	dance#7 → dance#8	61	17.1%
Adding a sequence	dance#7 → dance#789	52	14.6%
Adding a date	raven → raven2016	44	12.4%
Substituting digits/special characters with the same character type	tar!heel1 → tar!heel4	42	11.8%
Moving a letter, digit or special character block	$steve27 → 27$steve	38	10.7%
Duplicating digits/special characters	password1! → password11!	34	9.6%
Substituting letters with matching characters	raven → r@ven	29	8.1%
Deleting digits/special characters	alex28!!! → alex28!!	23	6.5%
Substituting digits/special characters with the "shift" character for the same key	l00py*!2 → l00py*!@	17	4.8%
Changing a small part of the previous password in a way not mentioned		43	12.1%
Other Methods			
Creating a completely new password		139	39.0%
Reusing old passwords from other accounts		37	10.4%
Using a password generator		23	6.5%
Using a different approach		8	2.2%

Table 3: Techniques participants used to update their main workplace password during their most recent password change (which may or may not have been due to an expiration policy). On average, participants used one or two modification techniques for changing their password.

spending in updating their password. However, those who received password change reminders that were not software based were two times more likely to report spending additional effort in updating their password, compared to those who did receive software reminders.

4.3 Expiration Frequency Has Little Impact

We generally observed that the presence and duration of an expiration policy had only a relatively minor impact on password-management behavior. There were some differences in the impact to password recall in the data collected from the workplace passwords survey, but these differences were not observed in the passwords perception survey. For example, we found that 15% of participants with an expiration policy for their main workplace password reported storing their password in their web browser, compared to 5% of participants without a policy, which was found to be significantly different ($p = .02, \phi = .16$). However, the self-reported use of this storage method did not significantly differ between different frequencies. We also found that 40% of participants who stated that they faced a 60-day expiration policy reported memorizing their password, compared to 59% who reported longer expiration periods and 68% who reported facing shorter expiration periods, which was also a significant difference ($p = .003, V = .11$). Furthermore, we found that neither the presence nor duration of an expiration policy impacted the number of reported account lockouts.

We found that different expiration periods did not have an impact on the self-reported strategies participants used to update their main workplace password. Moreover, we found that the presence and frequency of an expiration policy did not impact whether participants reported making their main workplace password similar to passwords they use for other accounts (both workplace and non-workplace related). This suggests that people who face frequent expiration are not more likely to reuse passwords from other accounts.

The majority of our participants reported that they did not find updating their password difficult, but 60% agreed or strongly agreed that it was annoying. The reported frequency of their expiration policies did not impact participants' sentiments toward updating their workplace password. This finding suggests that users adapt to the requirements placed on them, but still find them burdensome.

4.4 Security Perceptions

Participants in both the workplace passwords survey and password perceptions survey considered password changes important to the security of their workplace account. However, periodic password changes were viewed as less important than using a complex password, storing the password safely, and avoiding reusing password. Our results suggest that people accept the security advice provided to them, especially if from a trusted source such as the IT organization of their employer.

4.4.1 Secure But Annoying

In their responses to the workplace passwords survey, 82% of participants agreed or strongly agreed that "frequent password expiration makes it less likely that an unauthorized person will break into my account." Neither the self-reported existence nor duration of an expiration policy significantly impacted participants' agreement with this statement. However, 66% of participants thought that their updated password was about the same strength as their old one, and only 25% thought it was stronger, suggesting that many may not be exerting extra effort into making their password stronger when they change it. This is consistent with the strategies participants typically reported to modify their passwords.

Participants' self-reported update strategies were generally independent from their opinion about the relative strength of their new password, with the exception of capitalizing a letter which was positively correlated with thinking the updated password was stronger ($p = .008, \phi = .12$). Since some update strategies (specifically using a new password and using a password generator) were reportedly used for making the password stronger, participants who reported using these techniques could also have used them in the past and thus felt that their password strength did not change.

4.4.2 Participants' Threat Models

When asked why expiration prevents unauthorized account access, in their open-ended responses a small majority (54%) stated that expiration prevented password compromise. Of those, around a fifth specifically indicated that expiration helps with security by reducing the time window for an attacker to figure out their password. One participant reported, "It takes time to hack or steal a password and if it is changed frequently it is less likely that the hacker will have time to obtain the password." Twenty-eight percent of participants also reported that expiration is beneficial *after* a password, whether new or old, has already been compromised. Around half of these participants reported more specifically that the main benefit of expiration policies is that they reduce the time an illegitimate user has access to the account after they have logged in. To this effect, one participant said that, "There will be less time for a hacker to retrieve your information." Interestingly, of those that disagree with expiration's benefits, most (66%) cited concerns that expiration is insecure or ineffective, while only a small group (10%) cited inconvenience or unproductivity in their text responses.

While discussing potential threats, most participants mentioned concerns about hackers, general unauthorized users (e.g., "people", "attacker,") or guessers; 5% explicitly mention current or former employees as a concern. Around 5% also expressed that expiration would minimize the impact of employees sharing their workplace passwords. It should be noted that participants appear to use the word "hacker" in a broad, colloquial sense beyond the concept of hackers as phishers or computational guessers. It was usually impossible to tease out their conception of "hacker" or "hacking." For example, one participant reported, "I had an exboyfriend hack my Facebook because my password was not strong enough." When asked the open-ended question why a workplace might implement an expiration policy, participants' reasons generally aligned with their responses to the question about the general impact of password expiration.

4.4.3 Desired Policies

Only 10 of the 260 participants who had an expiration policy in the workplace passwords survey expressed the opinion that their passwords should never expire. Otherwise, participants were most likely to recommend their own workplace expiration policy as the appropriate policy ($p < .001, V = .55$). Similarly, only 10% of participants in NIST's study recommended a less frequent change cycle, compared to their current 60- or 90-day policies [4]

In their qualitative responses to *why* the policy they chose was the best, most participants could not really articulate the reason. For example, 13% responded with a sentiment that the time period they selected was "just right." A third of participants said that the expiration period they recommended balanced security with either usability (mainly the ability to remember passwords) or convenience concerns. Most users seemed to be able to reconcile their concerns with the benefits of added security. One participant who recommended a policy of every 60 days explained, "Every 30 days is too frequent. I often forget my password because it's always changing. I do however understand that security is important, so passwords should be changed somewhat frequently." Participants who picked shorter time periods (e.g., 30 days) cited a security reason more often than participants who picked longer ones (e.g., a year), who more often cited a balance of security/usability or security/inconvenience. A small fraction of participants also cited employer or industry norms as part of their recommendation, with responses like "It's the standard we use and it works well."

4.4.4 Relative Importance of Password Changes

Participants in the password perceptions survey generally viewed creating a complex password as the most important practice for account security, followed by storing the password in a safe place and creating a password that is not already used for another account. Changing passwords periodically was reported to be the least important of these practices ($p < .001, V = .14$). Figure 2 shows the distribution of responses for how important each behavior was perceived. Demographics, including technical expertise, did not impact opinions significantly. Additionally, participants' views were found to be independent of whether or not they had a workplace expiration policy.

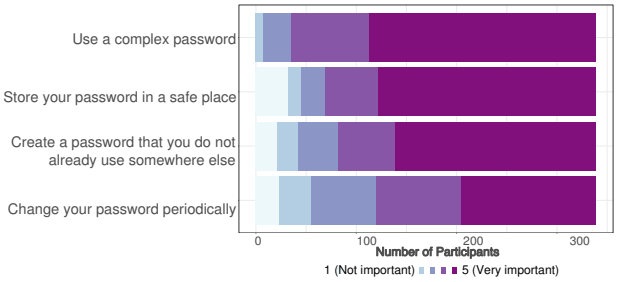


Figure 2: Distribution of ratings for each password-management practice studied. Participants viewed using a complex password the most important of these practices and changing passwords periodically the least important.

In qualitative responses explaining their rankings, participants mentioned usability concerns in roughly equal proportions for changing passwords, creating complex passwords, and avoiding password reuse. In line with their quantitative rankings, participants pointed out more downsides, such as it being inconvenient, insecure, unusable, or ineffective, for periodic password changes than other security practices. For example, one participant explained, "I don't think [periodic password change] is as important as people say...A really strong password doesn't just automatically become weaker simply because you've been using it for a while."

Also in line with the quantitative rankings, 5% of participants reported that a sufficiently complex password renders other practices less important, giving reasons like, "I don't believe it's necessarily important to change your password, if you have a secure one in the first place" or "If the password is good you should be able to [re]use it as much as you want as long as it is good." However, there were indications from the responses that users do not fully understand what comprises a good password, citing that "[...] a long nonsensical sentence works better and is more easily remembered, e.g., securitycomplexitymakesmypasswordssecurebutveryannoying." Lastly, some participants admitted that their attitudes and actions do not always align. Consistent with prior work, 2% volunteered that they believe they *should*

change their password or avoid reuse and that those practices are at least somewhat important, but that they do not do them [28]. When explaining their ranking for avoiding password reuse, one participant said, "It's important, but I do it [reuse passwords] anyway."

4.4.5 Hypothetical: Reversing Expiration Policy

In the hypothetical scenario in which the participant's workplace removed their expiration policy, almost half reported that they would continue changing their passwords periodically, be more likely to use a complex password, and be just as likely to avoid reusing passwords from other accounts. From the qualitative responses, reasons for continuing password changes were centered around it being a habit or beneficial for security. For example, one participant stated, "It's just a natural habit to do now for my own security." Those who stated they would not continue changes generally felt that it was too inconvenient or they would forget to do it if it was not required. As one participant put it, "I would forget as it is not on my high priority list."

From the quantitative data, half of participants who stated that they do not currently have a workplace expiration policy reported that they would be just as likely to use a complex password if their main workplace password expired periodically. Twenty-eight percent of participants who stated that they currently memorize their password reported they would no longer do so if periodic password changes were required. Almost half reported they were just as likely to create a password that is not used somewhere else. These results further highlight that password expiration may not contribute to a larger degree of password reuse, but likely does not encourage people to create more complex passwords during updates.

5. DISCUSSION

Our findings confirm that the strategies people use to adapt to their expiration policy are predictable. The majority of our participants reported coping with password changes by applying a simple modification to their current password. Our results are largely in line with NIST's study of the password-management behaviors of DOC employees [4]. However, our findings related to password reuse and backup recall methods do diverge, and may be attributed to the intense password burdens faced by DOC employees. Our results are also supported by Zhang et al. study, in which they were able to crack a substantial portion of their organization's passwords using the password history for the account [39]. This suggests that people in their organization were also typically using variations of their password during password updates.

Some participants reported using other coping strategies, such as cycling through a dedicated set of passwords for that account. Less than a quarter created a completely new password when it expired, a rate similar to that found by Shay et al. [30]. However, we did not find evidence that the self-reported strategies people use to update their password leads them to have weaker passwords. Furthermore, we observed that more frequent password changes did not lead to more self-reported reuse of passwords from other accounts. These results suggest that the negative security implications related to expiration may be limited to the case where there has already been a breach of an organization's passwords. In such cases, an attacker who knows a user's expired password may be able to easily guess the new password.

Our results also reveal that people generally do not have extremely negative reactions toward workplace password expiration, nor do they report significantly more usability burdens with more frequent password changes. Participants who reported facing more frequent expiration did not report experiencing a higher rate of account lockouts nor were they less likely to report memorizing their passwords than participants who reported facing less frequent password expiration. In addition, participants reported the same level of annoyance with updating their passwords, regardless of their self-reported password expiration frequency. This may be due to the fact that people adapt to expiration policies imposed by workplaces, often employing coping strategies that may reduce security in the event that their password is already known to an attacker.

In other scenarios, the presence of an expiration policy has little impact on security, even though a large percentage of participants held the view that updating their password would prevent hackers from cracking their passwords. Some participants even directly mentioned that they felt people will choose more creative passwords if they have to keep changing them. However, our results indicate that expiration largely does not influence people to create stronger passwords. Thus, password expiration likely provides no additional protection against an attacker with the modern computing resources to launch an automated guessing attack, even if the attacker does not have prior knowledge of the organization's passwords.

A few participants expressed concerns about targeted attacks in which a coworker or former employees of their organization would try to guess their password. Prior work has found targeted attacks to be a prevalent attack scenario that people worry about when managing their passwords [11,32,33]. Some were especially concerned about targeted attacks because they believed that sharing workplace passwords with coworkers was common practice. However, it is likely that expiration provides limited benefits even in the case of targeted attacks, since attackers may already know which modifications are typically used by their target.

Based on our results, we recommend that organizations consider whether the minimal security gains are worth implementing an expiration policy. Expiration policies may be attractive to organizations that have a history of password sharing among employees. Though expiring passwords may solve the immediate problem of system access to former employees, these organizations could be better off considering more secure mechanisms for enabling the collaboration between employees that causes password sharing. The benefits gained by avoiding attacks that are actually preventable by having an expiration policy must be weighed against a number of costs associated with implementing a policy, though our findings suggest that costs due to user burden are minimal considering current password-management demands.

Our second recommendation is that organizations implement enterprise password managers. In their existing implementations, expiration policies have limited benefits as users typically do not make significant changes to their passwords. Companies could enforce rules that require larger

changes and check for certain modifications, but this would have negative usability outcomes. Password expiration policies are most beneficial to account security if passwords are sufficiently random [31]. As people are unlikely to create and maintain random passwords on their own, organizations should consider the use of password managers with built-in generators, especially since some major password managers have enterprise versions of their software [21]. In our study, we found that those who did use a generator to create their password were much more likely to use one to update it and store their password in a password manager. However, it should be noted that many organizations that have implemented password expiration also have other policies which indirectly prevent their employees from using password managers. For example, some organizations in the United States government prevent employees from installing non-approved software, or even storing passwords on their terminals [2,35]. Considering our findings, such policies likely diminish any security benefit of having an expiration policy.

Across both surveys, we observed that participants strongly felt that password changes were important for account security. Some participants revealed that they held this perception because they trust the IT staff at their organizations and that is the advice they are repeatedly told. Overall, we observed that users adapt to the demands placed on them, even if in undesirable ways. This result may bode well for the future as security recommendations and best practices change with technology.

6. ACKNOWLEDGMENTS

This research was supported by the North Atlantic Treaty Organization (NATO) through Carnegie Mellon CyLab. This work was also supported in part by the CyLab Presidential Fellowship. The authors would like to thank Jessica Colnago for her contributions toward our preliminary analysis, and reviewers for their feedback.

7. REFERENCES

[1] S. Bellovin. Unconventional wisdoms. *IEEE Security and Privacy*, 4(1):88, 2006.

[2] Centers for Medicare and Medicaid Services. CMS policy for the acceptable use of CMS desktop/laptop and other it resources. `https://www.cms.gov/Research-Statistics-Data-and-Systems/CMS-Information-Technology/InformationSecurity`, December 2008.

[3] S. Chiasson and P. C. van Oorschot. Quantifying the security advantage of password expiration policies. *Designs, Codes and Cryptography*, 77(2-3):401–408, 2015.

[4] Y.-Y. Choong, M. Theofanos, and H.-K. Liu. NISTIR 7991: United states federal employees' password management behaviors - a department of commerce case study. Technical report, National Institute of Standards and Technology NIST, March 2014.

[5] Communications-Electronics Security Group. The problems with forcing regular password expiry. `https://www.ncsc.gov.uk/articles/problems-forcing-regular-password-expiry`, 2016.

[6] L. Cranor. Time to rethink mandatory password changes. `https://www.ftc.gov/news-events/blogs/techftc/2016/03/time-rethink-mandatory-password-changes`, March 2016.

[7] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. Wang. The tangled web of password reuse. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, volume 14, pages 23–26, 2014.

[8] M. Farcasin and E. Chan-tin. Why we hate IT: Two surveys on pre-generated and expiring passwords in an academic setting. *Security and Communication Networks*, 8(13):2361–2373, 2015.

[9] D. Florêncio and C. Herley. A large-scale study of web password habits. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 657–666, 2007.

[10] D. Florêncio and C. Herley. Where do security policies come from? In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 10:1–10:14, 2010.

[11] S. Gaw and E. W. Felten. Password management strategies for online accounts. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 44–55, 2006.

[12] P. A. Grassi, J. L. Fenton, E. M. Newton, R. A. Perlner, A. R. Regenscheid, W. E. Burr, J. P. Richer, N. B. Lefkovitz, J. M. Danker, Y.-Y. Choong, K. K. Greene, and M. F. Theofanos. NIST Special Publication 800-63b: Digital Identity Guidelines. Technical report, National Institute of Standards and Technology NIST, 2017.

[13] B. Grawemeyer and H. Johnson. Using and managing multiple passwords: A week to a view. *Interacting with Computers*, 23(3):256–267, 2011.

[14] P. G. Inglesant and M. A. Sasse. The true cost of unusable password policies: Password use in the wild. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 383–392, 2010.

[15] International Organization for Standardization (ISO) & International Electrotechnical Commission (IEC). Information technology: Security techniques, code of practice for information security management: ISO-IEC 27002:2013, October 2013.

[16] I. Ion, R. Reeder, and S. Consolvo. "No one can hack my mind": Comparing expert and non-expert security practices. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 327–346, 2015.

[17] R. Kang, S. Brown, L. Dabbish, and S. Kiesler. Privacy attitudes of Mechanical Turk workers and the US public. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 37–49, 2014.

[18] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 453–456, 2008.

[19] S. Kokolakis. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security*, 64:122–134, 2017.

[20] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. 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, 2011.

[21] LastPass. Lastpass enterprise. https://www.lastpass.com/enterprise, February 2018.

[22] D. Malone and K. Maher. Investigating the distribution of password choices. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 301–310, 2012.

[23] K. Muller. Statistical power analysis for the behavioral sciences. *Tehcnometrics*, 31:499–500, 1989.

[24] National Cyber Security Centre. Password guidance: Simplifying your approach. https://www.ncsc.gov.uk/guidance/password-guidance-simplifying-your-approach, August 2016.

[25] G. Notoatmodjo and C. Thomborson. Passwords and perceptions. In *Proceedings of the Australasian Conference on Information Security (ACISP)*, pages 71–78, 2009.

[26] PCI Security Standards Council. Payment card industry (PCI) data security standard.

[27] S. Pearman, J. Thomas, P. Emani Naeini, H. Habib, L. Bauer, N. Christin, L. Faith Cranor, S. Egelman, and A. Forget. Let's go in for a closer look: Observing passwords in their natural habitat. In *Proceedings of the Conference on Computer and Communications Security (CCS)*, 2017.

[28] S. Riley. Password security: What users know and what they actually do. *Usability News*, 8(1):2833–2836, 2006.

[29] B. Schneier. Security of password managers. https://www.schneier.com/blog/archives/2014/09/, September 2014.

[30] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. L. Mazurek, L. Bauer, N. Christin, and L. F. Cranor. Encountering stronger password requirements: User attitudes and behaviors. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, page 2, 2010.

[31] E. Spafford. Security myths and passwords. http://www.cerias.purdue.edu/site/blog/post/password-change-myths/, April 2006.

[32] E. Stobert and R. Biddle. The password life cycle: User behaviour in managing passwords. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2014.

[33] B. Ur, J. Bees, S. M. Segreti, L. Bauer, N. Christin, and L. F. Cranor. Do users' perceptions of password security match reality? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3748–3760, 2016.

[34] B. Ur, F. Noma, J. Bees, S. M. Segreti, R. Shay, L. Bauer, N. Christin, and L. F. Cranor. I Added '!' at the End to Make It Secure: Observing Password Creation in the Lab. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2015.

[35] U.S. Immigration and Customs Enforcement. General rules of behavior for users of DHS systems and IT resources that access, store, receive, or transmit sensitive information. https://www.ice.gov/doclib/sevis/pdf/behavior-rules.pdf, April 2008.

[36] A. J. Viera, J. M. Garrett, et al. Understanding interobserver agreement: The kappa statistic. *Fam Med*, 37(5):360–363, 2005.

[37] E. von Zezschwitz, A. De Luca, and H. Hussmann. Survival of the shortest: A retrospective analysis of influencing factors on password composition. In *Proceedings of the IFIP Conference on Human-Computer Interaction*, pages 460–467. Springer, 2013.

[38] M. Weir, S. Aggarwal, M. Collins, and H. Stern. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *Proceedings of the Conference on Computer and Communications Security (CCS)*, pages 162–175. ACM, 2010.

[39] Y. Zhang, F. Monrose, and M. K. Reiter. The security of modern password expiration: An algorithmic framework and empirical analysis. In *Proceedings of the Conference on Computer and Communications Security (CCS)*, pages 176–186, 2010.

[40] L. Zhang-Kennedy, S. Chiasson, and P. C. van Oorschot. Revisiting password rules: Facilitating human management of passwords. In *Proceedings of the APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–10, 2016.

APPENDIX

A. SCREENING SURVEY

1. How many workplace passwords do you have? *
   - 0
   - 1
   - 2
   - ...
   - 7
   - 8 or more
2. How many of your workplace passwords are you required to regularly change (i.e., they have an expiration policy)?*
   - All of my workplace passwords
   - Most of my workplace passwords
   - Some of my workplace passwords
   - None of my workplace passwords
   - Not sure
3. How old are you? *
   - 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
   - I prefer not to answer
4. What is your gender? *
   - Male
   - Female
   - Other (please specify)
   - I prefer not to answer
5. What is your race/ethnicity? *
   - American Indian or Alaska Native

- ○ Asian
- ○ Black or African American
- ○ White/ Caucasian
- ○ Hispanic or Latino
- ○ Non Hispanic
- ○ Other
- ○ I prefer not to answer

6. Which of the following best describes your highest achieved education level?*
   - ○ Some High School
   - ○ High School Graduate
   - ○ Some college, no degree
   - ○ Associates degree
   - ○ Bachelors degree
   - ○ Graduate degree (Masters, Doctorate, etc.)
   - ○ Other
   - ○ I prefer not to answer
7. Which of the following best describes your primary occupation? *
   - ○ Administrative Support (e.g., secretary, assistant)
   - ○ Art, Writing, or Journalism (e.g., author, reporter, sculptor)
   - ○ Business, Management, or Financial (e.g., manager, accountant, banker)
   - ○ Education or Science (e.g., teacher, professor, scientist)
   - ○ Legal (e.g., lawyer, paralegal)
   - ○ Medical (e.g., doctor, nurse, dentist)
   - ○ Computer Engineering or IT Professional (e.g., programmer, IT consultant)
   - ○ Engineer in other field (e.g., civil or bio engineer)
   - ○ Service (e.g., retail clerk, server)
   - ○ Skilled Labor (e.g., electrician, plumber, carpenter)
   - ○ Unemployed
   - ○ Retired
   - ○ College student
   - ○ Graduate student
   - ○ Mechanical Turk worker
   - ○ I prefer not to answer

B. WORKPLACE PASSWORDS SURVEY

The next few questions will ask you about your main workplace password. Please keep the following in mind:

- If you have more than one workplace, please respond using the workplace you consider to be your main workplace.
- If you have more than one password at your main workplace, respond using the one you consider to be your main password.
- If you are a student you may consider your university to be your main workplace.
- If Mechanical Turk is your main workplace, you should consider your Mechanical Turk password as your main workplace password.

1. How many workplace passwords do you have?
   - ○ none
   - ○ 1
   - ○ 2
   - ○ ...
   - ○ 7
   - ○ 8 or more
2. Thinking back to when you first created your main workplace password, which of the following methods did you use?
   - ☐ Used the first letter of each word in a phrase
   - ☐ Used the name of someone or something
   - ☐ Used a word in English
   - ☐ Used a word in a language other than English
   - ☐ Added numbers to the beginning or end of a word or name
   - ☐ Added symbols to the beginning or end of a word or name
   - ☐ Substituted symbols for some of the letters in a word or name (e.g. '@' instead of 'a')
   - ☐ Substituted numbers for some of the letters in a word or name (e.g. '3' instead of 'e')
   - ☐ Removed letters from a word or name
   - ☐ Used a phone number
   - ☐ Used an address
   - ☐ Used a birthday
   - ☐ Reused a password from another account exactly
   - ☐ Reused a password from another account with some modifications
   - ☐ Used something else (please specify)
   - ☐ I prefer not to answer
3. How many of your workplace passwords are you required to regularly change, i.e. they have an expiration policy?
   - ○ All of my workplace passwords
   - ○ Most of my workplace passwords
   - ○ Some of my workplace passwords
   - ○ None of my workplace passwords
   - ○ Not sure
4. How often are you required to change your main workplace password?
   - ○ Every week
   - ○ Every 30 days
   - ○ Every 60 days
   - ○ Every 90 days
   - ○ Every year
   - ○ Never
   - ○ Not sure
   - ○ Other (please specify)
5. Some organizations require their employees to change their passwords every 60 days. What do you think the impact of this policy is on security compared to organizations that do not require their employees to change their passwords at all?

- ○ It makes it less likely that an unauthorized person will log in to my account
- ○ It makes it more likely that an unauthorized person will log in to my
- ○ account
- ○ It doesn't impact security
- ○ I don't know

6. Why do you think this is the impact?
7. What do you think is the main reason for a workplace to set an expiration date on their employees' main passwords?
8. How often do you think your workplace should require its employees to change their main workplace password?
   - ○ Every week
   - ○ Every 30 days
   - ○ Every 60 days
   - ○ Every 90 days
   - ○ Every year
   - ○ Never
   - ○ Not sure
   - ○ Other (please specify)
9. Why do you think your workplace should require its employees to change their main password with this frequency?
10. The last time you changed your main workplace password, what approaches did you use? (select all that apply)
    - □ Adding a date (e.g. "raven" →"raven2016")
    - □ Adding a sequence (e.g. "dance#7" →"dance#789")
    - □ Capitalizing a character (e.g. "candy#" →"candY#")
    - □ Deleting digits/special characters (e.g. "alex28!!!" →"alex28!!")
    - □ Duplicating digits/special characters (e.g. "1!" →"11!")
    - □ Incrementing a character (e.g. "dance#7" →"dance#8")
    - □ Moving a letter, digit or special character block (e.g. "$steve27" →"27$steve")
    - □ Substituting digits/special characters with the same character type (e.g. "tar!heel1" →"tar!heel4")
    - □ Substituting letters with matching characters (e.g. "raven" →"r@ven")
    - □ Substituting digits or special characters with the "shift" character for the same key (e.g. "l00py*!2" →"l00py*!@")
    - □ Changing a small part of the previous password in a way not mentioned
    - □ Creating a completely new password
    - □ Reusing old passwords from other accounts
    - □ Using a password generator
    - □ Using a different approach (please specify)
    - □ I don't change my workplace password
11. How often have you used your strategy to change your main workplace password when it expired?
    - ○ I only changed my password once
    - ○ a couple of times (not often)
    - ○ most of the time
    - ○ every time
    - ○ I never changed my password
    - ○ other (please specify)
12. Why do you change your password this way? (select all that apply)
    - □ I have always done it this way
    - □ I heard about it from someone
    - □ I read it somewhere
    - □ I think it makes the password easier to remember
    - □ I think it makes the password stronger
    - □ It was the first strategy I thought of
    - □ other (please specify)
13. When changing your workplace password because the old one expired, do you always use the same strategy?
    - ○ I use the same strategy every time
    - ○ I use slightly different strategies at different times
    - ○ I use very different strategies at different times
14. How similar is your main workplace password to a password you use for another account at your workplace?
    - ○ My main workplace password is identical to a password I use for another workplace account
    - ○ My main workplace password is similar to a password I use for another workplace account
    - ○ My main workplace password is very different from any passwords I use for other workplace accounts
    - ○ I only have one workplace password
15. How similar is your main workplace password to a password you use for a non-workplace account?
    - ○ My main workplace password is identical to a password I use for a nonworkplace account
    - ○ My main workplace password is similar to a password I use for a nonworkplace account
    - ○ My main workplace password is very different from any passwords I use for non-workplace accounts
16. Where did you learn about changing your password this way? (select all that apply)
    - □ Boss
    - □ Colleague
    - □ Family member
    - □ Friend
    - □ IT department
    - □ Internet
    - □ Other (please specify)
17. When I last changed my main workplace password because it had expired, my new password was:
    - ○ Much weaker
    - ○ Weaker
    - ○ About the same
    - ○ Stronger
    - ○ Much stronger
    - ○ I don't know
18. How many workplace passwords do you have?
    - ○ none
    - ○ 1

- ○ 2
- ○ ...
- ○ 7
- ○ 8 or more

19. Frequent password expiration makes it less likely that an unauthorized person will break into my account.
    - ○ Strongly Disagree
    - ○ Disagree
    - ○ Neutral
    - ○ Agree
    - ○ Strongly Agree
    - ○ Not applicable
20. I find having to change my password due to my workplace expiration policy **difficult**.
    - ○ Strongly Disagree
    - ○ Disagree
    - ○ Neutral
    - ○ Agree
    - ○ Strongly Agree
    - ○ Not applicable
21. I find having to change my password due to my workplace expiration policy **easy**.
    - ○ Strongly Disagree
    - ○ Disagree
    - ○ Neutral
    - ○ Agree
    - ○ Strongly Agree
    - ○ Not applicable
22. Frequent password expiration makes it less likely that an unauthorized person will break into my account.
    - ○ Strongly Disagree
    - ○ Disagree
    - ○ Neutral
    - ○ Agree
    - ○ Strongly Agree
    - ○ Not applicable
23. I find having to change my password due to my workplace expiration policy **difficult**.
    - ○ Strongly Disagree
    - ○ Disagree
    - ○ Neutral
    - ○ Agree
    - ○ Strongly Agree
    - ○ Not applicable
24. I find having to change my password due to my workplace expiration policy **easy**.
    - ○ Strongly Disagree
    - ○ Disagree
    - ○ Neutral
    - ○ Agree
    - ○ Strongly Agree
    - ○ Not applicable
25. I find having to change my password due to my workplace expiration policy **annoying**.
    - ○ Strongly Disagree
    - ○ Disagree
    - ○ Neutral
    - ○ Agree
    - ○ Strongly Agree
    - ○ Not applicable
26. I find having to change my password due to my workplace expiration policy **fun**.
    - ○ Strongly Disagree
    - ○ Disagree
    - ○ Neutral
    - ○ Agree
    - ○ Strongly Agree
    - ○ Not applicable
27. What do you do to help yourself remember your main workplace password?
    - □ I let my web browser store it
    - □ I store it in an encrypted file
    - □ I store it in a password manager
    - □ I store it on a computer or device protected with another password
    - □ I store it on a computer or device that only I use
    - □ I write down my password on a piece of paper
    - □ I write down a reminder instead of the actual password
    - □ Nothing, I memorize it
    - □ I prefer not to answer
    - □ Other (please specify)
28. Why do you use this strategy to remember your main workplace password?
29. How many logins does it take for you to memorize your main workplace password?
    - ○ 1-2 logins
    - ○ 3-5 logins
    - ○ 6-10 logins
    - ○ More than 10 logins
    - ○ None, I memorize it when I create it or use a password I already memorized
    - ○ I don't memorize my main workplace password
30. How many times have you been unable to log into your main workplace account in the past year due to not having your password? (e.g. you forgot your password, the password was stored in a different device, etc.)
    - ○ Never
    - ○ 1-2 times
    - ○ 3-5 times
    - ○ 6-10 times
    - ○ More than 10 times
31. What do you need to do to change or recover your main workplace password if you forget it? (select all that apply)
    - □ I call someone on the phone
    - □ I send someone an email
    - □ I physically go somewhere or see someone in person
    - □ I mail someone a letter

- ☐ I use a website
- ☐ I don't know
- ☐ Other (please specify)

32. Who or what reminds you in advance of your password expiring to change your main workplace password? (select all that apply)
    - ☐ Boss
    - ☐ Colleague
    - ☐ IT department
    - ☐ Automated e-mails
    - ☐ Software on my computer
    - ☐ I don't get reminders in advance
    - ☐ Other (please specify)
33. When do you get the first reminder to change your main workplace password before it expires?
    - ○ Less than 1 day in advance
    - ○ 1 day in advance
    - ○ Less than a week in advance
    - ○ 1-2 weeks in advance
    - ○ 3-4 weeks in advance
    - ○ 1 month in advance
    - ○ More than 1 month in advance
    - ○ Other (please specify)
34. How does the reminder impact your effort in changing your main workplace password?
    - ○ I put more effort in updating my password
    - ○ I put less effort in updating my password
    - ○ It doesn't, I put the same amount of effort
    - ○ Other (please specify)
35. Has your main workplace password ever been accidentally leaked or otherwise compromised?
    - ○ Yes, I lost the device which had the password stored and the device was not password protected
    - ○ Yes, I lost the paper on which I wrote my password
    - ○ Yes, someone guessed it
    - ○ Yes, someone watched me type it in
    - ○ Yes, the IT infrastructure was breached
    - ○ Yes, other
    - ○ No
    - ○ Not sure
36. What did you do when your password was leaked? (select all that apply)
    - ☐ I changed my password before it expired
    - ☐ I kept my password and waited for it to expire to change it
    - ☐ I learned how to create stronger passwords
    - ☐ I changed where I stored my password
    - ☐ Other (please specify)
37. Do you have any other comments about your workplace password or its expiration policy? (optional)
38. Questions 38-42 are the same as Q3-7 in the screening survey above

C. PASSWORD PERCEPTIONS SURVEY

1. Questions 1-4 are the same as Q3-7 in the screening survey above
5. Have you ever held a job or received a degree in computer science or any related technology field?
    - ○ Yes
    - ○ No
6. Are you either a computer security professional or a student studying computer security?
    - ○ Yes
    - ○ No
7. To keep your account secure, how important is it to use a complex password (e.g., a long password with digits, symbols, and capital letters)? *
    - ○ 1 (Not important)
    - ○ ...
    - ○ 5 (Very important)
8. Please explain your answer to the question above. *
9. To keep your account secure, how important is it to store your password in a safe place (e.g, on a note hidden out of sight of other people) or not store it at all? *
    - ○ 1 (Not important)
    - ○ ...
    - ○ 5 (Very important)
10. Please explain your answer to the question above. *
11. To keep your account secure, how important is it to change your password periodically? *
    - ○ 1 (Not important)
    - ○ ...
    - ○ 5 (Very important)
12. Please explain your answer to the question above. *
13. To keep your account secure, how important is it to create a password that you do not already use somewhere else? *
    - ○ 1 (Not important)
    - ○ ...
    - ○ 5 (Very important)
14. Please explain your answer to the question above. *
15. Please rank the following in their order of their harm to account security, with "1" being the most harmful. (Multiple options may have the same ranking) *
    - Creating a password you have already used somewhere else (either exactly or with small modifications)
    - Storing the password in a place where others can access it
    - Not changing the password periodically
    - Creating a simple password (e.g., with no symbols or digits)
16. For each pair, which do you think contributes more to account security? * [Answered as "Left contributes much more," "Left contributes slightly more," "Both contribute equally," "Right contributes slightly more," "Right contributes much more"]

- Using a complex password | Storing your password in a safe place or not storing it at all
- Using a complex password | Creating a password that you do not already use somewhere else
- Using a complex password | Changing your password periodically
- Changing your password periodically | Creating a password that you do not already use somewhere else
- Storing your password in a safe place or not storing it at all | Changing your password periodically
- Storing your password in a safe place or not storing it at all | Creating a password that you do not already use somewhere else

17. How many workplace passwords do you have in total? *
    - ○ 0
    - ○ 1
    - ○ 2
    - ○ ...
    - ○ 7
    - ○ 8 or more

    Logic: The following two questions are hidden if "How many workplace passwords do you have in total?" is "none"

18. What do you do to help yourself remember your main workplace password? *
    - ☐ Let your web browser store it
    - ☐ Store it in an encrypted file
    - ☐ Store it in a password manager
    - ☐ Store it on a computer or device protected with another password
    - ☐ Store it on a computer or device that only you use
    - ☐ Write it down on a piece of paper
    - ☐ Write down a reminder instead of the actual password
    - ☐ Nothing, you memorize it
    - ☐ Prefer not to answer
    - ☐ Other (please specify)

19. Does your workplace have an expiration policy for your main password? *
    - ○ Yes
    - ○ No

    Logic: The following questions five are hidden if "Does your workplace have an expiration policy for your main password?" is "No" or "Not Sure"

20. How often are you required to change your main workplace password? *
    - ○ Every week
    - ○ Every 30 days
    - ○ Every 60 days
    - ○ Every 90 days
    - ○ Every year
    - ○ Never
    - ○ Not sure
    - ○ Other (please specify)

21. Suppose your workplace's expiration policy changed and your main workplace account password will no longer expire. How likely would you be to continue to periodically change the password of your account anyways? *
    - ○ Very unlikely
    - ○ Unlikely
    - ○ Neither likely or unlikely
    - ○ Likely
    - ○ Very likely

22. Please explain your answer to the question above. *

23. Suppose your workplace's expiration policy changed and your main workplace account password will no longer expire. Going forward, how would you remember your main workplace password? *
    - ☐ Let your web browser store it
    - ☐ Store it in an encrypted file
    - ☐ Store it in a password manager
    - ☐ Store it on a computer or device protected with another password
    - ☐ Store it on a computer or device that only you use
    - ☐ Write it down on a piece of paper
    - ☐ Write down a reminder instead of the actual password
    - ☐ Nothing, you would memorize it
    - ☐ Prefer not to answer
    - ☐ Other (please specify)

24. Suppose your workplace's expiration policy changed and your main workplace account password will no longer expire. Going forward, would you be more or less likely to do the following? * [Answered on a 5-point Likert scale from "Much more likely" to "Much less likely"]
    - Use a complex password
    - Create a password you do not already use somewhere else

    Logic: The following questions five are hidden if "Does your workplace have an expiration policy for your main password?" is "Yes"

25. How often do you change the password of your main workplace account? *
    - ○ Never
    - ○ Every week
    - ○ Every month
    - ○ Every few months
    - ○ Every year
    - ○ Other (Please specify)

26. Please explain why you change your password with the frequency indicated above. *

27. Suppose your workplace implemented an expiration policy and from now on your main workplace account password will expire periodically. Going forward, how would you remember your main workplace password? *
    - ☐ Let your web browser store it
    - ☐ Store it in an encrypted file
    - ☐ Store it in a password manager

- ☐ Store it on a computer or device protected with another password
- ☐ Store it on a computer or device that only you use
- ☐ Write it down on a piece of paper
- ☐ Write down a reminder instead of the actual password
- ☐ Nothing, you would memorize it
- ☐ Prefer not to answer
- ☐ Other (please specify)

28. Suppose your workplace implemented an expiration policy and from now on your main workplace account password will expire periodically. Going forward, would you be more or less likely to do the following? * [Answered on a 5-point Likert scale from "Much more likely" to "Much less likely"]
    - Use a complex password
    - Create a password you do not already use somewhere else

The Effectiveness of Fear Appeals in Increasing Smartphone Locking Behavior among Saudi Arabians

Elham Al Qahtani
College of Computing and Informatics
UNC Charlotte
ealqahta@uncc.edu

Mohamed Shehab
College of Computing and Informatics
UNC Charlotte
mshehab@uncc.edu

Abrar Aljohani
johani.abrar@outlook.com

ABSTRACT

Saudi Arabia has witnessed an exponential growth in smartphone adoption and penetration. This increase has been accompanied with an upward trend in cyber and mobile crimes. This calls to efforts that focus on enhancing the awareness of the public to security-related risks. In this study, we replicated the study performed by Albayram et al. [14] published in SOUPS 2017; however, our study targetted participants in Saudi Arabia. We also investigated different fear appeal video designs that were more suited for this population (customized video, Arabic dubbed, and captions for the original video). The results from the original study, conducted in the United States, showed that 50% of participants in the treatment group and 21% in the control group enabled screen lock. The reason for replicating the original paper was to increase Saudis' awareness regarding the importance of sensitive data, especially with the increasing level of cybercrime. Our results showed that the Saudi-customized video was extremely effective in changing our participants' locking behavior (72.5% of participants enabled the screen lock), based on customized applications and Saudi culture. The dubbed video was the second-most effective (62.5%) locking behavior. Finally, we have illustrated our data comparison analysis in detail.

1. INTRODUCTION

In Saudi Arabia, there has been an exponential grown in the use of smartphone technologies. According to a report by the Saudi Ministry of Communication & Information Technology [5], in 2001 the number of mobile subscriptions in Saudi Arabia was around 2.5 million (12% mobile penetration). By 2017, the number had risen to 43.63 million with a population penetration rate of 137%, which is the highest mobile penetration rate in the region [26]. In Saudi Arabia, the mobile banking penetration is 81%, which is considerably higher than other developed and emerging Asian nations [6]. In addition, Saudis are increasingly using smartphone applications to conduct business and communication.

With this increase, cyber threats are becoming more common as 58% of the Saudi population have experienced some form of online cyber crime in the past year, and one in four users have had their mobile device stolen, potentially exposing sensitive information in their e-mail, social media and banking apps to cyber thieves [3]. Kaspersky Security statistics showed that 53.1% of Saudi users were affected by local threats (malware spread in local networks, by USBs, CDs, DVDs) [7]. In 2018, the Saudi government statistics showed an increase in online blackmail, where extortionists demanded money, sex and many other demands from their victims [4].

In response to these cybersecurity challenges, there have been several efforts lead by both the public and private sectors to provide security tools, education, and awareness to the Saudi population. For example, the organization responsible for awareness of Saudi companies, government organizations, and society in Saudi Arabia is the National Center for Cybersecurity [9]. It aims to educate Saudis about the dangers of using the Internet, the social communication and the loss of personal information through awareness lectures, workshops, and social media. We believe that the security and privacy problems in Saudi Arabia are further exacerbated by the lack of security awareness of the population. Alzahrani et al. [16] stated that 92% of Saudis had never attended security training. Recently [10, 11], the Saudi Federation for Cyber Security has provided educational and awareness programs in Saudi Arabia.

As there are several types of proprietary data storage in mobile phones, screen lock techniques are effective in protecting mobile content and preventing strangers from gaining unauthorized access, which could lead to extortion via the threat of destroying the victim's reputation. Several researchers [37, 29, 36, 23, 45] have noted the important relationship between applying a screen lock mechanism and users' motivation and risk perception. A study of smartwatches by Nguyen et al. [38] used similar concepts to evaluate different locking mechanisms. For effective security and improved user experience, Ohana et al. [40] found that combining biometric identification (e.g., fingerprint, face, or voice recognition) with other security lock mechanisms improved security.

Our paper replicates the study performed by Albayram et al. [14] for two main reasons: the lack of security training leads to 92% of Saudi society to be not aware of the importance of security and its potential consequences [16], and cybercrime (e.g., blackmail) is increasing in Saudi Arabia [4]. For this reason, we decided to investigate if we could improve the efficiency of communicating risk to the Saudi population.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

The original study compared the smartphone locking behavior of a treatment group that watched a fear appeal video to that of a control group. That online study found that 50% of the treatment group enabled the screen lock compared to 21% of the control group. In this paper, we replicated the original study and did so with a Saudi population of mobile users through an in-person study with 200 Saudi participants. We also investigated various video designs that were more suited to the Saudi population (customized video, Arabic-dubbed video, and Arabic-captioned video) in the context of smartphone locking behavior. We found that 72.5% of participants who watched the customized video enabled the screen lock, as compared to 62.5% of those who watched the dubbed video, 42.5% of those who watched the subtitled video, and 20% of those who watched the original video. In contrast, among the control group that was not shown any video, 17.5% enabled the screen lock.

This paper is organized as follows. Section 2 discusses the original study. Section 3 discusses related work. Section 4 describes the methodology. The study results are presented in Section 5. Section 6 presents the study discussion. Section 7 discusses limitations and suggestions for future work. Section 8 concludes our paper.

2. THE ORIGINAL STUDY

In the original study, Albayram et al. [14] discussed how effective risk communication leads to a change in users' risky behavior, rather than simply annoyance among users. The authors designed a video that targeted people who did not use a screen lock on their smartphone. They sought to determine whether this video would affect users' locking behavior and change their attitude about enabling a screen lock. The video explained the risks of having their data stolen if users did not activate a screen lock mechanism on their smartphone. The video was based on guidelines for four appeals to fear from the protection motivation theory [42, 47], as follows: (1) perceived severity (perception of the seriousness of threat), (2) perceived vulnerability (possibility incidence of threat), (3) self-efficacy (confidence about taking the recommended action), and (4) response efficacy (perception of the efficacy of the recommended action).

In that study, the participants were divided into a treatment group that watched the video and a control group that did not watch the video. The study was conducted through Amazon's Mechanical Turk and was divided into two phases, as follows:

1. The main study included three parts with several kinds of questions: both groups were asked questions about demographics, smartphone usage behavior, online security behavior, and reasons for not using a screen lock on their smartphone. Questions about the video evaluation were only asked to the treatment group. The two groups were also asked questions about security and privacy concerns, the perceived value of their data, risk perception, response cost, and response efficacy.
2. The follow-up study included questions for both the control and the treatment group about whether they had enabled a screen lock and what their reasons were for doing so or not doing so.

Albayram et al. [14] found that the informative fear appeal video specified for the treatment group was effective at changing users' locking behavior and risk perception based on fear appeals elements. After conducting the follow-up study, they found that 50% of participants in the treatment group and 21% in the control group had enabled a secure screen lock on their smartphone.

3. RELATED WORK

Two kinds of research are relevant to our study: that dealing with cultural context and that dealing with risk communication techniques.

3.1 Cultural Context

In this section, we clarify how cultural differences play an important role in changing society's behavior, such as national culture, cognitive differences, and laws. The two following studies illustrate significant differences in different populations. A replicated study performed by Mayer et al. [35], conducted a comparison of password composition policy (PCP) samples from U.S. and from German websites. One of their findings was that German websites had lower PCP strengths than did the U.S. samples. Another study performed by Nishi et al. [39] conducted an online experiment using economic games among 337 Indian users, 1,059 American users, and 66 users from other countries. They found that American culture played an important role in users' lives. Based on this, users quickly made decisions involving highly mutual assistance compared to others in online social environments. Regarding national culture, Salter et al. [43] found that the simple national culture effect was important even between societies with similar cultures (e.g., the U.S. and Canada). Using the agency theory of human behavior patterns, for example, American managers had a higher rate of adverse decisions than Canadians. A study performed by Kitayama et al. [33] examined decision making and cultural differences in cognition using a framed-line test that compared Japanese (an East Asian culture) and Americans (a North American culture). They found that Japanese made more accurate decisions than Americans. In addition, laws have a significant influence on a person's behavior; for example, the Kingdom of Saudi Arabia is a conservative society and applies gender segregation [34] in higher education, banks, mosques, and restaurants.

3.2 Risk Communication Techniques

It is important to understand smartphone users' perspective on security and their perception of the seriousness of risk. Egelman et al. [22] examined the security behavior intentions scale (SeBIS), which measures users' behavior in regard to computer security. The SeBIS was used to rate user awareness, password strength, updates on security, and securement mechanisms. Several studies have investigated risk communication methods using videos, graphics, text, symbols, and messages to measure the effects on users' risk perception and behavior [13, 14, 19, 28, 32, 41]. For example, Pattinson and Anderson [41] examined risk communication methods and symbols and graphics included in information security messages. They found no significant improvement in risk communication using this method. Bravo-Lillo et al. [19] performed three experiments using a computer security dialogue with five attractors to attract user attention focusing on the perception of cost. However, they found that interacting with only two inhibitive attractors by swiping over the text and typing in the text box was effective

in reducing the incidence of users' ignoring the appropriate action for granting permission, based on habituation that taught users how to achieve the task and speeded up their responses to a dialogue.

A study performed by Yousra et al. [32] investigated the use of an animated dialogue to attract users' attention toward granting permissions through a non-inhibitive attractor which highlighted personal information in the color red. They found that, compared to the control and to checkbox dialogues, the proposed animated dialogue had a significant impact on the duration of users' looking at the personal information, users' concern about their data, and users' understanding of the purpose for each permission. A study by Harbach et al. [28] investigated risk communication using personalized examples of permissions for Android applications and found that such communication affected users' decisions by making them aware of potential risks. In 2017, two studies conducted by Albayram et al. [13, 14] used effective risk communication by a video to raise user awareness of security and privacy and to motivate users to follow recommended security procedures. Albayram et al. [14] evaluated the effect of a fear appeal video on users' perceptions and behavior in terms of enabling a secure screen lock mechanism. In addition, Albayram et al. [13] investigated the effect of leverage videos about enabling two-step verification (2FA) on users perceptions and attitudes, seeking to motivate them to be familiar with security tools and updated new security advice.

4. METHODOLOGY

The purpose of the present study is to examine the effectiveness of fear appeal videos in changing Saudi Arabians' risk perceptions and raising their awareness. In addition, based on the videos' effectiveness, we want to investigate if there are changes between the users' initial reasons for not using a screen lock and their reasons in the second round follow-up study. The following sections describe the video design for the present study, the hypotheses, and the study design.

Figure 1: A frame from the customized video content

4.1 Video Design

Researchers in the education field have shown that videos are highly effective in attracting users' attention and in making salient features clear [20, 30, 46], and studies have measured the changes videos make in users' perceptions [13, 14, 18, 25]. In order to study the effectiveness of video media on communicating fear appeal to the Saudi population, we designed a customized video targeting the Saudi audience that gave participants an explanation of the expected risks of not using locking mechanisms [31] and a short demonstration of how to add a lock to their phone, based on security advice from Android [12, 17, 44] and iOS [1]. The duration and topic of the customized video were similar to that of the video in the original study (the original video was 3 minutes long). The videos were customized with scenarios and content to ensure relevance to the Saudi audience. The scenarios were applied to the perception of the data value, the perceived vulnerability about Saudis smartphones' security and privacy, the perceived cost that connected to Saudis' decisions of following the recommended tips and self-efficacy that was based on their belief of the suggested tips [42]. The customization focused on the following aspects:

- **Relevant Risks and Fears:** This aspect focuses on ensuring the video targets risks and fears that are related to the target audience and that would clearly communicate threats. For example, the video was customized to present a scenario in Arabic in which a victim is being deceived and blackmailed via stored personal media in WhatsApp messages, which is aligned with the popularity of online blackmail and extortion crimes in Saudi Arabia. In addition, the video demonstrated the risks of exposing media stored in smartphone photo albums and discussed the risk of reputation damage specially in a conservative society. Regarding self-efficacy, the customized video illustrated the steps of enabling a secure lock screen on iPhone and Android phones with Arabic settings on the smartphone. All these scenarios would be familiar to the Saudi population.

- **Relevant applications and attributes:** The video was updated to include smartphone applications and attributes that are common in Saudi Arabia. The video was customized to include information related to the top-ranking applications among Saudis [8] (e.g., WhatsApp, Snapchat, Instagram, and ALRajhiBank (a popular Islamic bank [2])), rather than those used in the original video (Paypal, Netflix, Bank of America app, and Amazon). We presented an ALRajhiBank scenario in which the victim's bank information was stolen by sending a 2FA code to the victim's stolen device and accessing the bank password and username stored on the victim's smartphone. We also included a scenario using Whatsapp. Attributes focus on data attributes that are relevant to the target population, for example the national identity number was used instead of social security number which was used in the original video. The customized video also demonstrated the steps to enable a secure screen locking method using device settings in Arabic, both for Android and for iOS, whereas the original video only included Android's screen lock set up.

- **Cultural:** The video should also be culturally relevant, which was ensured by customizing the video dialog to use culturally relevant vocabulary delivered in Arabic. In addition, the video used photos of men and women in the Saudi dress taking into consideration the ideology of the conservative society.

The main goal was, to customized the video content, to ensure the participants felt that it is relevant their mobile

experience and their security. The narrator of the Saudi-customized video was one of the authors of this paper, and the transcript of the customized video is in the Appendix. Figure 1 shows a frame from the customized video content that is related to Saudi culture.

The study examined five groups–four treatment groups and a control group, as described below:

- Original group (Treatment Group 1): watched the video in English used in the original U.S. study.
- Dubbed group (Treatment Group 2): watched the original video but with Arabic dubbed. Figure 2 (a) shows a frame from the Arabic-dubbed video. [1]
- Subtitled group (Treatment Group 3): watched the original video but with Arabic captions. Figure 2 (b) shows a frame from the Arabic-captioned video. [2]
- Customized group (Treatment Group 4): watched a Saudi-specific video in Arabic. Figure 2 (c) shows a frame from the customized video. [3]
- Control group: did not watch any video.

We assigned the original video to one group to test whether it changed Saudi locking behavior or perceptions based on its visual content, ignoring the English dialogue; our measurements depended on the effectiveness of the video's content (the applications used in the video, the dialogue in the video). As Saudi participants might not understand the original video due to the unfamiliar smartphone applications and the English dialogue, we added Arabic captions and dubbed the original video, assigning these to the subtitled group and the dubbed group, respectively, to test whether these changes would affect Saudis' perceptions.

The customized and dubbed videos significantly affected Saudis' locking behavior on perceived severity, vulnerability, and response efficacy. The rating for the perceived inconvenience of using the secure screen lock was also considerably lower for both groups compared to the original, the subtitled and the control groups.

4.2 Hypotheses

In the present study, we propose the following hypotheses:

Hypothesis 1 (H1): There will be significant differences among groups in their ratings of perceived sensitive data (H1a). The group that watched the Saudi-customized video will have higher ratings on perceived sensitive data than will the other treatment groups, and all the treatment groups will be higher than the control group (H1b).

Hypothesis 2 (H2): There will be significant differences among groups regarding concerns about their smartphones' security and privacy and about their data being used by other people (H2a). The customized group will have a higher level of concerns about their smartphones' security and privacy and their data being used by other people than will the other treatment groups, and all the treatment groups will be higher than the control group (H2b).

[1] https://youtu.be/5OpZ2jVGxRY
[2] https://youtu.be/4P91WgGH-r4
[3] https://youtu.be/g-IlWrvmRF4

Hypothesis 3 (H3): There will be significant differences among groups in their ratings of perceived severity and risk awareness (H3a). The customized group will have higher ratings of perceived severity and risk awareness than will the other treatment groups, and all the treatment groups will be higher than the control group (H3b).

Hypothesis 4 (H4): There will be significant differences among groups in their ratings of the perceived response cost (H4a). The customized group will have lower ratings of perceived response cost than will the other treatment groups, and all the treatment groups will have lower ratings of the control group (H4b).

Hypothesis 5 (H5): There will be significant differences among groups in their ratings of response efficacy (H5a). The customized group will have higher ratings of response efficacy than will the other treatment groups, and all the treatment groups will be higher than the control group (H5b).

Hypothesis 6 (H6): There will be significant differences among groups regarding the number of participants who enabled a screen lock (H6a). The customized group will have a higher number of participants who enabled a screen lock than other treatment groups, compared to the control group (H6b).

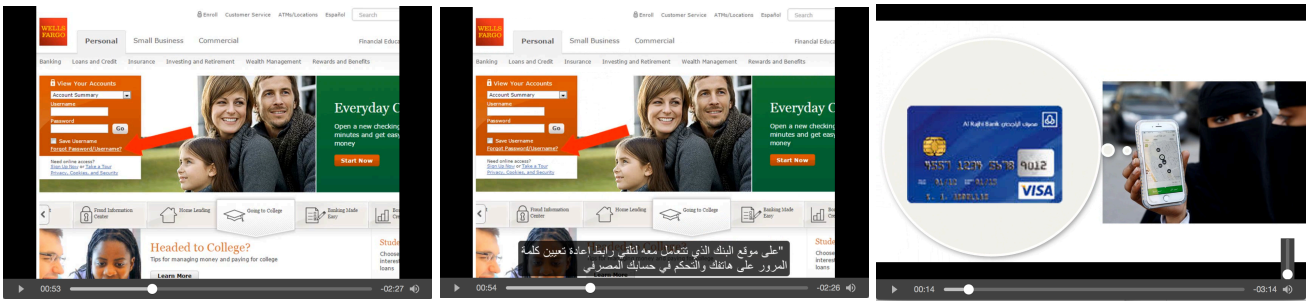
4.3 Study Design

The present study was conducted in person, whereas the original study was an online study. For the present study, we recruited Saudi participants who were at least 18 years old, owned a smartphone that provided a secure screen locking mechanism, and did not activate the screen locking mechanism on their phone. For example, we excluded Saudis who had old cell phones that did not support screen-locking mechanisms. We collected participants' cell phone numbers and used those numbers to call the participants in order to follow up after the study and investigate whether they had enabled screen locking; we then deleted participants' phone numbers after the follow up call was made. The participants were recruited through flyers and through face-to-face recruitment. When administering the study, we did not provide any explanatory information to the participants but instead asked them to watch a video and answer survey questions afterward.

The present study applied the same questions as did the original study, but they were translated into Arabic to fit the Saudi population; the administered translated survey can be seen in the Appendix. The process of the study design, which included a first round of a main study and then a second round with a follow-up study, was as follows.

4.3.1 First Round: The Main Study

We interviewed 200 Saudi participants individually (an in-person study) who met our inclusion criteria; the participants were assigned randomly to one of the five groups to prevent self-selection bias. After obtaining user consent, we explained the purpose of our study and collected participants' phone numbers to use for the second round of the study. Each group included 40 Saudi participants. We met them in public places (e.g., outside prayer areas, shopping malls, schools, and hospital waiting areas). Our study was conducted in different cities in Saudi Arabia. Responses to the questions were recorded in the questionnaire to ensure



(a) Arabic dubbed (b) Arabic captions (c) Customized video

Figure 2: Screen shots of the different videos

that we received accurate responses. We let participants watch a complete video based on their random assignment to groups. All Arabic responses were translated into English.

During the first round of the main study, we asked participants three sections of questions. The first section was background questions, smartphone usage behavior questions, online security behavior questions, reasons for not using a screen lock, and their opinions about people who use a lock screen on their smartphone. The second section was only for those in the treatment groups, who watched a video; we asked them about the video's effects and their evaluation of the video. The control group were not shown a video, and hence they were not asked the questions in the second section. All five groups were asked the third section, which included questions about data value, security and privacy concerns, risk awareness, response cost, and response efficacy. The average total time of our interviews with members of the groups who had watched the video was approximately 20 minutes, including 3 minutes during which participants watched the video, whereas the duration of the interviews with members of the control group was around 15 minutes.

4.3.2 Second Round: The Follow-Up Study

We followed up with the participants a week after their initial interview to evaluate whether they had enabled a screen locking mechanism on their phones and to learn the reason behind their choice. This second interview was performed using a follow-up the questionnaire and was conducted by phone or in a location agreed upon with the participant.

Our study was approved by UNC Charlotte's Institutional Review Board [1] and the Saudi Arabia regulatory committee.

5. EVALUATION

Since our data was ordinal, we used non-parametric tests for the analysis. In comparing all the groups independently, we used the Kruskal-Wallis test (H) in an equal sample size [24]. We also used post hoc multiple comparison to compare groups for each research question. To avoid Type I error (alpha) in testing our significance, we used the adjusted significance for Bonferroni at (0.05) [21]. All analysis was done using the Statistical Package for the Social Sciences (SPSS).

5.1 Sample Statistics

Based on our data analysis of the Saudi population, we assigned an equal number of people of each gender to each group (20 male and 20 female), so that there was no significant difference among five groups in terms of gender. We performed the Kruskal-Wallis test and found no significant differences among all groups regarding the following demographic characteristics: age ($H(4) = 3.881$, $p = .422$), education level ($H(4) = 4.512$, $p = .341$), level of computer knowledge ($H(4) = 4.35$, $p = .365$), and participants' language ($H(4) = 5.191$, $p = .268$).

In terms of smartphone usage behavior, there were no significant differences among the five groups when we asked them five questions about their smartphones' operating system type ($H(4) = 2.576$, $p = .631$), number of times using their smartphones during the day ($H(4) = 4.654$, p = .325), number of applications on their smartphones ($H(4) = 8.921$, $p = .063$), number of times they used these applications ($H(4)= 6.078$, $p = .193$), and the applications they used daily ($H(4) = 2.774$, $p = .596$).

In comparing online security behavior among all five groups, we found no significant differences in concerns about their online accounts being hacked ($H(4) = 5.837$, $p = .212$). All groups had no concerns about online security ($H(4) = 4.019$, $p = .403$) and whether they used antivirus software security ($H(4) = 7.040$, $p = .134$).

5.2 Reasons for Not Employing a Screen Lock

In the questionnaires first section, we asked participants their initial reasons for not employing a lock screen on their smartphones and their opinions about why people use screen locks on their smartphones. For the question related to the initial reasons for not employing lock screen we updated it to a multiple choice question using the coding results that were concluded by the original study as choices, we also added a "other" option for participants that have a reason not included in the listed choices.

In comparing the responses of all five groups, we found no significant differences in the reasons for not employing a screen lock among the treatment groups and the control group ($H(4) = 2.707$, $p = .608$), as 30% of participants in all groups agreed on the top reason, "Annoying to use" [14, 23, 27, 29], 22% chose "Nothing to hide" as their reason, 21.5% chose "No risk", and 17% chose "Forgettable/mental burden." The least common answer was "Don't know how to set up", chosen by 5% of participants in all groups. The ranking of reasons was similar to that found in the original study.

[1] IRB Protocol #17-0426

For instance, a comment from the customized group said, "I share my phone with my mother and sister, especially when using Internet data by connecting through a personal hotspot." A participant from the original group reported not locking the phone "because my children continue to press the secret code by mistake and that hangs up the mobile for a long time when I might need to use my mobile immediately. In other words, I do not use it to avoid suspending the screen." The control group comments included, "Annoying, and there are tools that unlock passwords easily."

We noticed in the interviews when we asked participants why they thought people used a secure lock, participants' reasons were related to their misconceptions about using a screen lock and a failure to recognize the importance of their sensitive data. For example, a Saudi participant over 60 years old from the original group mentioned that lock users "hide inappropriate information, such as forbidden photos inside their phone, and if you are confident you will not hide anything from your family." A participant from the dubbed group said those who locked their phones wanted to "protect their calling balance from anyone using it." One from the customized group commented, "Maybe they have bad photos on their phones and misfortunes to hide from others." In a different vein, someone from the control group replied, "They know how to use this technology."

5.3 Impact of Fear Appeals on Fear of Losing Sensitive Data

We assumed that there were significant differences among the groups in their ratings of perceived sensitive data (H1a), and that the group that watched the Saudi-customized video would have higher ratings on perceived sensitive data than the other treatment groups, which would be higher than the control group (H1b).

To test the first hypothesis (H1), which included (H1a) and (H1b), we asked participants in all the five groups two questions related to their perceived sensitive data. The first question was, "Do you think that data stored in your smartphone is valuable enough to protect?" The second question was, "How much privacy-sensitive data do you think your smartphone stores?" The answers were measured on a scale ranging from (0) "None at all" to (3) "A great deal of privacy-sensitive data."

Performing the Kruskal-Wallis test among all five groups, we found significant differences for both questions at $p \leq .001$ for the first question ($H(4) = 22.58$ with a medium effect size, $\eta^2 = 0.11$), for the second question ($H(4) = 24.88$ with a medium effect size, $\eta^2 = 0.12$).

For the first question, differences were found at an adjusted significant level $p \leq .001$ by performing the Bonferroni multiple comparisons tests between the original and the customized group and between the control and the customized group. We found that 95% of participants from the customized group, 80% from the dubbed group, 67.5% from the subtitled group, and 55% from both the original and the control groups chose "Yes," depending on the priority.

For the second question, we found significant differences when performing the Bonferroni multiple comparisons test with adjusted significance and mean rank, which used to compare the effect of the different of each group, as shown in Table 1. Among participants in the dubbed group (median = 2), 37.5% believed their smartphones had a moderate deal of privacy-sensitive information, whereas 42.5% of those in the customized group (median = 2) thought their smartphones had a great deal of privacy-sensitive information. Among those in the subtitled (median = 1), original (median = 1), and control groups (median = 1), 32.5%, 45%, and 42.5%, respectively, rated the amount of sensitive data they had as "None at all."

"How much privacy-sensitive data do you think your smartphone stores?"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Control (81.6) vs. Customized (131.6)	$\leq$.001
Original (82.1) vs. Customized (131.6)	$\leq$.001
Subtitled (92.2) vs. Customized (131.6)	.016

Table 1: Post hoc test of the second question on sensitive data

Participants' varying ratings of the importance of the data stored on their smartphones, especially for the treatment groups, demonstrated the impact level of the fear appeal of the videos, and these changes reflected participants' perceptions about the importance of the personal data they had on their smartphones. The results of the customized and dubbed groups demonstrated a significant change in behavior compared to the other groups' ratings of their phones as having either "A moderate amount" or "A great amount" of sensitive data. Thus, both H1a and H1b were supported.

5.4 Impact of Fear Appeals on Security and Privacy Concerns

We hypothesized that there would be significant differences among groups regarding concerns about their smartphones' security and privacy and their use by other people (H2a), and that the customized group would have a higher level of concern about their smartphones' security and privacy and use by other people than would other treatment groups, which would be higher than that of the control group (H2b).

To test the hypotheses (H2a, H2b), we asked participants three questions related to their perceived vulnerability: "How much do you worry about your smartphone's security?", "How much do you worry about your smartphone's privacy?", and "How concerned are you about your smartphone use by others?" Answers regarding their worries and concerns were rated on a scale from (0) "Not at all" to (3) "Extremely."

We used the Kruskal-Wallis test to compare the five groups for those three questions, our results indicated that there were significant differences, with a significance level of $p \leq .001$ for the first question ($H(4) = 50.53$ with a large effect size, $\eta^2 = 0.25$), the second question ($H(4) = 55.47$ with a large effect size, $\eta^2 = 0.28$) and for the third question ($H(4) = 42.21$ with a large effect size, $\eta^2 = 0.21$).

As shown in Table 2 for the first question, we found significant differences for the first question when performing the Bonferroni multiple comparisons test with adjusted significance and mean rank. The percentages of participants who rated their concerns about their smartphone security as either "Moderately worried" or "Extremely worried" were as

"How much do you worry about your smartphones security?"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Control (64.9) vs. Dubbed (127.4)	≤.001
Control (64.9) vs. Customized (135.7)	≤.001
Original (75.5) vs. Dubbed (127.4)	≤.001
Original (75.5) vs. Customized (135.7)	≤.001
Subtitled (98.9) vs. Customized (135.75)	.029
"How much do you worry about your smartphone's privacy?"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Control (61.7) vs. Subtitled (99)	.026
Control (61.7) vs. Dubbed (126.3)	≤.001
Control (61.7) vs. Customized (139.2)	≤.001
Original (76.2) vs. Subtitled (99)	≤.001
Original (76.2) vs. Customized (139.2)	≤.001
Subtitled (99) vs. Customized (139.2)	.012
"How concerned are you about your smartphone being used by others?"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Control (67.7) vs. Subtitled (106.6)	.019
Control (67.7) vs. Dubbed (116.3)	≤.001
Control (67.7) vs. Customized (136.6)	≤.001
Original (75.3) vs. Dubbed (116.3)	.010
Original (75.3) vs. Customized (136.6)	≤.001

Table 2: Post hoc test of the three questions on security and privacy concerns

follows: 75% of participants in the customized group (median = 2), 67.5% in the dubbed group (median = 2), 47% in the subtitled group (median = 1), 20% in the original group (median = 0), and 15% in the control group (median = 0).

For the second question about smartphone privacy, as shown in Table 2, we found significant differences between groups, as 47.5% of participants in the subtitled group (median = 1), 67.5% in the dubbed group (median = 2), and 75% in the customized group (median = 2.5) had moderate or extreme worries about their smartphone privacy, whereas 57.5% of participants in the original group (median = 0) and 65% in the control group (median = 0) chose "None at all."

When we asked participants a third question as to their concerns about others using their smartphones, we found significant differences between groups, as shown in Table 2. Percentages of participants who rated themselves as either "Moderately concerned" or "Extremely concerned" for each group were as follows: Customized, 82.5% (where median = 3); Dubbed, 67.5% (where median = 2); Subtitled, 57.5% (where median = 2); Original, 27.5% (where median = 1); and Control, 20% (where median = 1).

Thus, Hypotheses H2a and H2b were supported. The group that watched the customized video was more worried about their smartphones' security, privacy, and use by others than were other groups. Thus the video with customized content (Saudi identities, a Saudi bank, and fake dialogue to lure victims through WhatsApp messages) made participants aware of the risks of not using a screen lock on their smartphones.

5.5 Impact of Fear Appeals on Perceived Severity

For the third hypothesis (H3), we assumed that there would be significant differences among the groups' ratings of the perceived severity and risk awareness (H3a), and that the customized group would have higher ratings of the perceived severity and risk awareness than would the other treatment groups, which would be higher than those of the control group (H3b).

To test Hypothesis 3, we asked participants three questions. The first question was, "If your smartphone is lost or stolen, how disruptive will the loss of your data on your smartphone be to your daily life?" Participants rated their disruption from (0) "Not at all disruptive" to (3) "Highly disruptive." The second question was "How likely is it that you would lose your smartphone?" The third question was "How likely is it that someone else would attempt to access your smartphone?" For both questions, participants rated their likelihood on a scale from (0) "Extremely unlikely" to (3) "Extremely likely."

Performing the Kruskal-Wallis test, we found significant differences among the five groups for the three questions at a significance level of $\leq .001$ for the first question ($H(4) = 38.02$ with a large effect size, $\eta^2 = 0.19$), the second question ($H(4) = 26.31$ with a medium effect size, $\eta^2 = 0.13$) and for the third question ($H(4) = 28.92$ with a large effect size, $\eta^2 = 0.14$).

"If your smartphone is lost or stolen, how disruptive will the loss of your data on your smartphone be to your daily life?"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Original (72.4) vs. Dubbed (121.8)	≤.001
Original (72.4) vs. Customized (134.8)	≤.001
Control (78.2) vs. Dubbed (121.8)	.005
Control (78.2) vs. Customized (134.8)	≤.001
Subtitled (95.2) vs. Customized (134.8)	.015
" How likely is it that you would lose your smartphone?"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Control (75.7) vs. Dubbed (117.1)	.009
Control (75.7) vs. Customized (129)	≤.001
Original (82.2) vs. Customized (129)	.002
" How likely is it that someone else would attempt to access your smartphone?"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Control (74.8) vs. Dubbed (115.6)	.011
Control (74.8) vs. Customized (131.9)	≤.001
Original (81.4) vs. Customized (131.9)	≤.001

Table 3: Post hoc test of the three questions on perceived severity

When we asked participants the first question, as shown in Table 3, we found significant differences between groups by performing the Bonferroni multiple comparisons test with adjusted significance and mean rank. A majority of participants in the dubbed group (median = 3) and the customized group (median = 3) indicated it would be highly disruptive, at 52.5% and 53.5%, respectively. In contrast, 17.5% of participants in the original group (median = 1), 32.5% in the subtitled group (median = 2), and 15% in the control group (median = 1) said it would be highly disruptive, which is significantly lower than in the customized and dubbed groups.

As shown in Table 3, for the second question, there were significant differences between the groups, as 37.5% of participants in the original group (median = 1) and 32.5% in the control group (median = 1) chose "Extremely unlikely", whereas 52.5% in the subtitled group chose "Moderate likely" (median = 2), 47.5% in the customized group (median = 2), and 45% in the dubbed group (median = 2) chose " Extremely likely."

For the third question, we found differences between the groups, as shown in Table 3. We noticed that 30% of the participants in the original group (median = 1.5) and 30% in the control group (median = 1) rated someone else's attempting to acces their phone " Extremely unlikely", whereas 50% in the dubbed group (median = 2.5), 50% in the customized group (median = 2.5) chose "Extremely likely", and 45% of the subtitled group (median = 2) chose "Moderate likely."

Thus, Hypotheses H3a and H3b were supported. The risk perceptions of both the subtitled and the original groups changed only minimally, perhaps because most of them did not pay attention to the Arabic captions or did not realize what the speaker was saying. In contrast, the assigned videos in the dubbed group and the customized group were extremely effective, and the impact was reflected in their perception of the seriousness of potential risks and possible adverse consequences of losing their personal information by not enabling a screen lock on their smartphones.

5.6 Impact of Fear Appeals on Response Cost

We hypothesized that there would be significant differences among the groups in their ratings of perceived response cost (H4a), and that the customized group would have lower ratings of perceived response cost than would other treatment groups, with the control group being the lowest (H4b).

To test Hypothesis 4, we asked participants whether they found using a screen lock to be a hassle, whether they agreed that entering an unlock code several times was inconvenient, and whether they agreed that it was inconvenient because a secure code was easily forgettable. We asked them to rate these on a scale ranging from (1) "Strongly disagree" to (5) "Strongly agree."

Performing the Kruskal-Wallis test, we found significant differences among all the groups for the first question ($H(4) = 24.76$, $p \leq .001$ with a medium effect size, $\eta^2 = 0.12$), the second question ($H(4) = 29.27$, $p \leq .001$ with a large effect size, $\eta^2 = 0.14$), and the third question ($H(4) = 17.55$, $p = .002$ with a medium effect size, $\eta^2 = 0.09$); therefore, H4a was supported.

The groups differed significantly, as shown in Table 4. Fewer participants from the dubbed group (median = 3) and the customized group (median = 3) agreed that the screen lock would be a hassle (25% and 27%, respectively), whereas in the original, subtitled, and control groups, 67.5%, 40%, and 55%, respectively, agreed that it would be a hassle.

For the second question, as shown in Table 4, we found significant differences between the groups, with 77.5% of participants in the original group (median = 5) agreeing it was inconvenient to enter a locking code, 65% in the subtitled group (median = 4), and 82.5% in the control group (median = 5). In contrast, both the dubbed (median = 2) and the customized groups (median = 2.5) had lower levels of agreement (40% and 45%, respectively).

"If I use a secure screen lock on my smartphone, it will be too much of a hassle for me"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Customized (76.9) vs. Control (117.3)	.013
Customized (76.9) vs. Original (128.5)	≤.001
Dubbed (82.2) vs. Original (128.5)	.002
"I feel using a secure screen lock on my smartphone is too inconvenient due to having to enter an unlock code every time I use the phone "	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Dubbed (74.4) vs. Control (120.9)	.002
Dubbed (74.4) vs. Original (126.1)	≤.001
Customized (78.9) vs. Control (120.9)	.007
Customized (78.9) vs. Original (126.1)	≤.001
"I feel using a secure screen lock on my smartphone is too inconvenient because it is hard to remember"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Customized (78.8) vs. Subtitled (114.5)	.043
Customized (78.8) vs. Control (118.9)	.013
Dubbed (82.7) vs. Control (118.9)	.037

Table 4: Post hoc test of the three questions on response cost

Differences were found between the groups in whether they thought it was too hard to remember a secure screen lock, as shown in Table 4. 42.5% of participants in the original group (median = 3) agreed with this idea, 40% in the subtitled group (median = 2.5), and 45% in the control group (median = 3). In contrast, only 22.5% of participants in the dubbed group (median = 2) and 20% in the customized group (median = 2) agreed with this, or about half as many.

Thus, Hypothesis 4 was supported; the evidence shows the major impact of effective risk communication: both dubbed and customized groups changed their perception of inconvenience (see section 5.2) and came to realize the importance of enabling a secure screen lock on their smartphones.

5.7 Impact of Fear Appeals on Response Efficacy

We assumed that there would be significant differences among the five groups in their ratings of response efficacy (H5a), and that the customized group would have higher ratings of response efficacy than would the other treatment groups, with the control group being lowest (H5b).

In this part, we measured participants' confidence in performing the recommended behavior of activating one of the screen lock methods and tested Hypothesis 5 by asking five questions. The first question was whether they thought that using a screen lock was a good idea. The second question was whether they thought it was easy to use it on their smartphones. The third question was whether they thought it secured their smartphones. The fourth question was whether they understood the purpose of using the screen lock. The last question was whether they thought a screen lock protected the data on their smartphones. Answers were rated using 5-point Likert scale from (1) = "Strongly disagree" to (5) = "Strongly agree."

Performing the Kruskal-Wallis test, we found differences among all groups at a significant level, $p \leq .001$, for the five questions (($H(4) = 48.26$ with a large effect size, $\eta^2 = 0.24$), ($H(4) = 47.37$ with a large effect size, $\eta^2 = 0.23$), ($H(4) = 51.66$ with a large effect size, $\eta^2 = 0.26$), ($H(4) = 56.87$ with a large effect size, $\eta^2 = 0.28$), and ($H(4) = 47.40$ with a large effect size, $\eta^2 = 0.24$), respectively).

"Do you think that using a screen lock is a good idea?"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Control (69.6) vs. Dubbed (129.6)	≤.001
Control (69.6) vs. Customized (134.4)	≤.001
Original (71.5) vs. Dubbed (129.6)	≤.001
Original (71.5) vs. Customized (134.4)	≤.001
Subtitled (97.3) vs. Customized (134.4)	.031
"Do you think a screen lock is easy to use on your smartphone?"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Original (69.7) vs. Customized (128.2)	≤.001
Original (69.7) vs. Dubbed (137.1)	≤.001
Control (74.3) vs. Customized (128.2)	≤.001
Control (74.3) vs. Dubbed (137.1)	≤.001
Subtitled (93.2) vs. Dubbed (137.1)	.005
"Do you think a screen lock secures your smartphone?"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Control (66.9) vs. Dubbed (130.9)	≤.001
Control (66.9) vs. Customized (137.4)	≤.001
Original (78.9) vs. Dubbed (130.9)	≤.001
Original (78.9) vs. Customized (137.4)	≤.001
Subtitled (88.3) vs. Dubbed (130.9)	.006
Subtitled (88.3) vs. Customized (137.4)	≤.001
"Do you understand the purpose of using a screen lock?"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Control (61.4) vs. Dubbed (132.2)	≤.001
Control (61.4) vs. Customized (137.1)	≤.001
Original (77.6) vs. Dubbed (132.2)	≤.001
Original (77.6) vs. Customized (137.1)	≤.001
Subtitled (94.2) vs. Dubbed (132.2)	.024
Subtitled (94.2) vs. Customized (137.1)	.006
"Do you think a screen lock protects your personal data in your smartphone?"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Control (71.2) vs. Dubbed (130.3)	≤.001
Control (71.2) vs. Customized (136.2)	≤.001
Original (78.5) vs. Dubbed (130.3)	≤.001
Original (78.5) vs. Customized (136.2)	≤.001
Subtitled (86.3) vs. Dubbed (130.3)	.004
Subtitled (86.3) vs. Customized (136.2)	≤.001

Table 5: Post hoc test of the five questions on response efficacy

As shown in Table 5 for the first question, there were significant differences between the groups by performing the Bonferroni multiple comparisons test with adjusted significance and mean rank, as 77.5% of participants in the dubbed group (median = 5), 57.5% in the subtitled group (median = 4), and 82.5% in the customized group (median = 5) agreed with the first question, whereas only 35% of participants in the original group (median = 2) and 30% in the control group (median = 2) agreed, a level lower than the other groups.

We found significant differences between the groups for the second question, as shown in Table 5. 70% of participants in the original group (median = 2), 42.5% in the subtitled group (median = 3), and 65% in the control group (median = 2) thought that it would not be easy to use a screen lock, whereas 75% of participants in the dubbed group (median = 4) and 67.5% in the customized group (median = 4) thought that it would be easy to use.

As shown in Table 5, for the third question, there were significant differences among the groups. In the original (median = 3) and the control group (median = 3), 47.5% and 32.5%, respectively, agreed that the screen lock secured their smartphones, in contrast to 60% of participants in the subtitled group (median = 4), 82.5% in the dubbed group (median = 5), and 87.5% in the customized group (median = 5).

Significant differences were also found among the groups for the fourth question in Table 5. 82.5% of participants in the dubbed group (median = 5) and 87.5% in the customized group (median = 5) agreed that they understood the purpose of the screen lock, in contrast with 65% in the subtitled group (median = 4), 42.5% in the original group (median = 3), and 30% in the control group (median = 2).

Once again, we found significant differences among the groups for the last question in Table 5. 82.5% of participants in the dubbed group (median = 5) and 85% in the customized group (median = 5) agreed that a screen lock protected their data, in contrast to 60% in the subtitled group (median = 4), 47.5% of participants in the original group (median = 3), and 32% in the control group (median = 3).

These ratings supported the idea that the Saudi-customized video and the Arabic-dubbed video were significantly effective in raising participants' risk awareness and encouraging them to follow recommended security practices that would benefit them and changed their views about activating a secure screen lock.

5.8 Impact of Fear Appeals on Saudis' Behavior (Follow-Up)

A week after the initial interview; during the second round of the follow-up study, we contacted participants to see whether they had enabled the screen lock or not. We hypothesized that there would be significant differences among the groups in terms of the percentage of participants who enabled a screen lock (H6a), and that the customized group would have a higher level of participants who enabled a screen lock than the other treatment groups, with the control group being the lowest (H6b).

If participants answered " Yes", we asked them "What motivated you to enable it?", " When did you activate it?", "What is the type of the screen lock?", and " How was it?" If their answer was "No", we asked them to tell us their reasons for not employing the screen lock.

In this round, we tested our hypotheses to see who among the treatment groups and the control group had enabled the screen lock. As we had predicted in H6a, the KruskalWallis test ($H(4) = 39.46$, $p \leq .001$) indicated significant differences among the five groups in terms of the level of participants who enabled the screen locks.

As shown in Table 6, there were significant differences among the groups regarding the enabling of a screen lock. Table 7 shows the number of participants that did and did not enable a screen lock for all groups.

"Have you enabled the screen lock on your smartphone or not?"	
Comparison of Groups (Mean Rank)	**Adj. Sig.**
Control (75) vs. Dubbed (120)	≤.001
Control (75) vs. Customized (130)	≤.001
Original (77.5) vs. Dubbed (120)	≤.001
Original (77.5) vs. Customized (130)	≤.001

Table 6: Post hoc test of participants who did and did not enable a screen lock

Number of participants	**Enabled**	**Not enabled**
Original (n=40)	8	32
Dubbed (n=40)	25	15
Subtitled (n=40)	17	23
Customized (n=40)	29	11
Control (n=40)	7	33

Table 7: Number of participants who did and did not enable a screen lock for all groups

5.8.1 Comments from Those Enabling the Screen lock

When we asked participants about their motivation for enabling the screen lock, it revealed the thinking behind their responses. A participant from the original group said, " Graphics show the existence or truth of the meaning of not using screen lock, so I activated it in the same day after our interview." A participant from the dubbed group reported, "Now that I know the benefits of having a security code to protect my secrets, and I understand how hackers can steal my personal information from my online accounts. However, if the content in the video is written in Arabic, then it will be really clear, especially how to follow the steps of setting up a screen lock for anyone who does not know English. Of course, I activated the pattern on my phone." A participant from the subtitled group mentioned his motivations, saying, "The privacy examples in this video changed my mind about enabling a secure lock on my phone. I enabled it two days later because I was so busy after our interview." A participant in the customized group commented that "I became convinced of the risk that my data might get stolen if there was no screen lock. Most of my government transactions are managed by my husband. The most important documents are my bank records sent through WhatsApp messages, which includes my Saudi ID, my passport, and my bank information. Previously, I saw no risk from not locking the screen because my mobile was with me all the time, but after watching the video I learned a lot, and I will send this video to my acquaintances and friends. I activated a screen lock immediately after watching the video." A participant from the control group said, "I enabled the passcode again after I responded to the questionnaire. The questions made sense and led me to think about it again. After our interview, I enabled it immediately."

Most of the participants, who enabled screen locking mechanism, enabled it on the same day. Only six of our participants enabled it on the second day.

The control group stated as the first reason for their motivation that the questionnaire led them to change their locking behavior. Among the treatment groups, the main motivation for enabling the screen lock was the videos that they had watched: The numbers in parentheses indicate the number of participants who were motivated by that reason versus the total number of participants who had activated a screen lock for all motivations: Customized (23/29), Dubbed (22/25), Subtitled (13/17), Original (6/8), and Control (4/7). The second most commonly cited motivation was security and privacy concerns: Customized (4/29), Subtitled (3/17), Dubbed (2/25), Original (2/8), and Control (2/7). Only four participants out of all the groups stated having had a bad experience as a reason for their motivation.

Regarding the type of secure lock method used by members of all the groups, the most commonly used was passcode/Touch ID (35 participants), followed by pattern (22), PIN (18), fingerprint (17), and other secure]ity mechanisms (7). Overall, 58 participants said they found the use of a screen lock convenient, in contrast to the 27 who found it inconvenient.

5.8.2 Comments from Those Not Enabling the Screen Lock

Among the treatment groups and the control group, the stated reason for participants' not enabling a screen lock on their smartphones was " Forgettable" (28.9%), "Nothing to hide" (25%), "Annoying to use " (16.7%), "low perceived threat" (15.8%), "Don't know how to set up" (5.3%). The last chosen was "Another reason" (7.9%), meaning that participants stated a reason not listed, such as this one from the dubbed group: "I and my family use my phone as a personal hotspot for sharing Internet data, and it is annoying to put a screen lock on, especially when someone who trusts you shares your phone." A participant from the control group noted, "As there are some advanced tools that break the screen lock mechanisms, I am not motivated to use any of them."

5.9 Ratings for Treatment Groups

The present study investigated the effects of different video designs incorporating fear appeal on four treatment groups (160 participants). The results showed that communicating risk had a positive effect on peoples' perceptions which led them to change their screen locking behavior and increased their awareness of new security recommendations. The following section evaluates the video used for each treatment group (Customized, Dubbed, Subtitled, and Original).

As shown in Table 8, each treatment group of participants was shown their assigned video and were asked to rate the persuasion, believability, and effectiveness of the video on a scale from (0) "Not at all" to (3) "Very."

We noticed that the percentages of participants from the customized and the dubbed groups that found the video persuasive, believable, and effective, were higher than the percentages of participants who enabled the screen lock. The main reasons for the different percentages despite their high level of video rating, especially Customized and Dubbed, referred to their reasons for not employing the screen lock,

	Original	Dubbed	Subtitled	Customized
Enabled	20%	62.5%	42.5%	72.2%
Persuasion	17.5%	72.5%	35%	87.5%
Believability	10%	72%	47.5%	90%
Effectiveness	15%	82.5%	40%	92.5%

Table 8: Percentage of participants who enabled a screen lock and treatment groups' evaluations of videos

which were "Forgettable" (Customized : 36.4%, Dubbed: 40%) and "Nothing to hide" (Customized: 27.3%, Dubbed: 33.3%).

We asked participants in the treatment groups what aspects of the video they saw that they liked and did not like. Table 9 shows the good aspect and the bad aspect most chosen by participants in each group.

Spearman's coefficients were used to verify the correlation of a mutual relationship between participants' conviction that lock screen was a good idea and those who enabled it among treatment groups. Based on their responses, we found a significant correlation at $p \leq 0.001$, as shown in Table 10 (first row). We also verified the correlation of a mutual relationship of persuasive and effectiveness levels on the video with participants who enabled a screen lock on their smartphones at $p \leq 0.001$, as shown in Table 10 (second and third row, respectively). This showed the extent to which the video affects the participants, based on their assessment of the level of effectiveness and their conviction, that lead to change their locking behavior.

6. DISCUSSION

Through interviewing participants face to face, we were able to record their answers accurately, and we saw their reactions to the video reflected in their responses, especially for those in the treatment groups. It was interesting during our interview to listen to participants' questions related to our study that went beyond those in our questionnaire. For example, one of the participants from the customized group commented, "If the steps of setting up a screen lock were only printed on paper, it would be easy for people to follow the steps in case they did not watch the video."

The results of our study showed us that the Saudi customized-video had the most effect on participants' perceptions, and led them to change their phone-locking behavior. This can be attributed to the video customization, which employed banks that are heavily used in Saudi Arabia, an Arabic scenario in which the victim is deceived via the use of WhatsApp, and Arabic descriptions of how to enable a screen lock for both iOS and Android systems. The Arabic-dubbed video had the second highest level impact. These findings were based on the four axes set forth in the protection motivation theory [42]. In the second round, the follow-up study, depending on the impact of each video, the percentage of Saudi participants who employed the screen lock increased, to 72.5% for the customized group, to 62.5% for the dubbed group, to 42.5% for the subtitled group, and to 20% for the original group. This significant impact is reflected in participants' answers, especially those of the customized group and the dubbed group. Despite the impact of the videos on participants' locking behavior, however, 7 participants from the customized group and 11 participants from the dubbed group did not enable the screen lock, stating that the main reason they did not do so was either that the phone locking process was "Forgettable" or that they had "Nothing to hide."

Among the subtitled group, participants' responses to fear appeal questions varied. Our findings showed that this video was effective in a simple proportion. It was proven that the percentage of Saudi participants in the subtitled group who did not enable the screen lock was 57.5%, which was higher than the percentage of those who enabled it (42.5%). The extent of this simple effect was reflected in participants' answers. Those who did not enable the screen lock chose "Don't know how to set up a screen lock for iPhone" and "I did not understand the video's content." Participants who did not like the video's content complained that they did not understand the activation process because they focused on the video's graphics instead of on the Arabic captions.

Moreover, we found that the original video was only minimally effective in changing Saudis' locking behavior, as only 20% of our participants enabled their phone locks, compared 50% of the participants in the original study. It was clear from their responses that our treatment group who watched the original video had difficulties with the English dialogue, even though the graphics were simple. We noticed that responses from our original treatment group were very close to those of the control group, and we believe the reason was a lack of understanding of the video's content. Their comments also bear witness to the video's ineffectiveness; for example, "I trust all people around me who use my phone and I do not expect they will steal my personal information," and "Because of my age, I always forget what the password is and I do not know how to set up a screen lock." These can comments can be compared to those of participants from the control group, such as "I spend all my time using my phone, and it makes me so nervous each time I unlock my screen and the number of times I get confused especially in public places that I will not use it" and "As there are some advanced tools that break the screen lock, I am not motivated to use any screen lock mechanism."

Factors that participants identified as contributing to the effectiveness of a video and as making a positive impression were related to language, customized applications, and clearing up misconceptions about the purpose of phone locks. The factors are related to the perceptions that hindered them from enabling a screen lock on their smartphones. The last factor we noticed was misconceptions about the purpose of phone locks, which appeared clearly in participants' reasons for not locking the screen and their view of people using the locked screen before they had watched a video (Section 5.2).

7. LIMITATIONS AND FUTURE WORK

During our research before the first round of the main study, we faced several limitations. The hardest challenge we faced was the time required to search for participants who met the criteria in this study and to interview them individually. Additionally, the researcher recorded responses to the Arabic questionnaires that were given to participants, especially to the elderly.

Based on the sample numbers of the Saudi population, it is important to test for the age factor within the sample. For

Groups	Good aspect	Bad aspect
Original ($n = 40$)	Graphics (17/40)	Language (32/40)
Dubbed ($n = 40$)	Explanation of risks (21/40)	None (23/40)
Subtitled ($n = 40$)	Explanation of risks (13/40)	Language (22/40)
Customized ($n = 40$)	Explanation of risks (22/40)	None (37/40)

Table 9: Good and bad aspects of videos watched by treatment groups

	Original	Dubbed	Subtitled	Customized
Correlation between a screen lock as a good idea and those who enabled	.628 *	.701 *	.614 *	.803 *
Correlation between those who enabled and video persuasiveness	.665 *	.402 *	.714 *	.625 *
Correlation between those who enabled and video effectiveness	.662 *	.555 *	.617 *	.545 *

* Correlation is significant at .001

Table 10: Correlation of mutual relationship with behavior change among treatment groups

example, Alkhunaizan et al. [15] investigated the effects of mobile commerce acceptance among the Saudi population based on three factors: gender, age, and education. They found that the age factor significantly impacted mobile usage, indicating that further study of a larger sample in Saudi Arabia is needed to test the impact of age on the effectiveness of communicating risk to change behaviors.

Moreover, the native language of Saudis is Arabic; for example, when the original group and the subtitled group watched the video, most participants expressed that they did not understand the dialogue. However, some of them were able to understand via the graphics the consequences of not enabling screen lock. It is important to study social, cultural, and linguistic factors that motivate participants to change their behavior.

This study has proven the effectiveness of the customized video and the dubbed video in changing users' locking behavior andleading them to follow the recommended procedures to reduce risks. We found that, when we asked Saudi participants their initial reasons why people use a screen lock, the majority of their responses indicated they held a misconception about the reasons. After they watched the video, they changed their locking behavior. It is be good to conduct a similar study among the Saudi population dealing with cybercrime (e.g., blackmail) and to monitor the most important factors extracted from the data.

8. CONCLUSION

With the increase of cybercrime in Saudi Arabia, people have to be conscious of possible threats to their personal data if they do not follow security advice. We presented a replication of the study by Albayram et al. [14] on the Saudi population of smartphone users, and we extended the investigation of the effectiveness of several fear appeal video designs that fit Saudis' perceptions of locking behavior.

As a result of comparisons among the four treatment groups and the control group, we found that the most effective video among the treatment groups was the Saudi-customized video, as 72.5% of that group's participants enabled screen lock, and 92.5% rated this video as effective. The customized video included the Saudi-specific factors as described above in Section 4.1. The condition having the second-highest level of the effectiveness was the dubbed video, as 62.5% of that group's participants enabled their screen locks, and 82.5% rated this video as effective. After that came the original video with Arabic captions, as 42.5% of that group's participants enabled their screen locks, and 40% of them rated this video as effective. The least effective video was the original video for our Saudi original treatment group, as only 20% of that group's participants enabled their screen locks, and only 15% rated the video as effective, compared to the treatment group in the original study conducted in the U.S., where 50% of participants enabled their screen locks, In contrast, among the control group that was not shown any video, 17.5% enabled screen their locks, which was similar to the 21% who did so among the control group in the original U.S. study.

The participants' initial highest reason for not using their screen locks in all five groups was "Annoying to use" (30%); however, in the second round of the follow-up study, the highest-ranking reason changed to "Forgettable/Mental burden" (28.9%), but only for those who did not enable the screen lock on their smartphones. Finally, based on the impact of the fear appeal videos, the effectiveness of the Saudi-customized video showed that communicating risk had a significant effect on Saudis' perceptions that led them to change their locking behavior and to increase their awareness of the importance of following security recommendations.

9. ACKNOWLEDGMENTS

The authors express deepest gratitude to their families support. Special thanks to Dr. Heather Lipford for reviewing the study and providing feedback. Finally we would like to thank the Saudi participants for participating in this study.

10. REFERENCES

[1] 5 ways to protect your iphone from hackers. http://www.marketers-voice.com/2017/11/protect-your-iPhone-from-spyware.html. Accessed: 2018-02-08.

[2] Al rajhi bank: About us.

http://www.alrajhibank.com.sa/en/investor-relations/about-us/pages/about-us.aspx. Accessed: 2018-01-29.

[3] Cybercrime hit 6.5m in kingdom last year. http://www.arabnews.com/node/967966/saudi-arabia. Accessed: 2017-09-16.

[4] The head of the committees reveals and warns against responding to the demands of extortionists: most of them are sexual. https://sabq.org/GVKc5Y. Accessed: 2018-01-30.

[5] ICT Indicators in K.S.A by end Q2-2017. http://www.citc.gov.sa/en/reportsandstudies/indicators/Pages/CITCICTIndicators.aspx. Accessed: 2018-05-23.

[6] Interest in digital banking offers major opportunities for gulf banks. https://www.consultancy.uk/news/13440/interest-in-digital-banking-offers-major-opportunities-for-gulf-banks. Accessed: 2018-05-18.

[7] Kaspersky lab presents cybersecurity trends in the meta region. https://me-en.kaspersky.com/about/press-releases/2017_kaspersky-lab-presents-cybersecurity-trends-in-the-meta-region. Accessed: 2018-05-23.

[8] Mobile app ranking in saudi arabia. https://www.similarweb.com/apps/top/google/store-rank/sa/all/top-free. Accessed: 2018-01-24.

[9] The national center for cybersecurity. https://www.amen.sa/index.html. Accessed: 2018-05-24.

[10] Saudi arabia aims to develop cyber security, programming skills of students. http://english.alarabiya.net/en/variety/2018/02/04/Saudi-Arabia-aims-to-develop-cyber-security-skills-of-students-.html. Accessed: 2018-05-23.

[11] Saudi cyber security college signs mou for us training. http://www.arabnews.com/node/1291616/saudi-arabia. Accessed: 2018-05-23.

[12] Set the screen lock. https://support.google.com/nexus/answer/2819522?hl=ar. Accessed: 2018-02-08.

[13] Y. Albayram, M. M. H. Khan, and M. Fagan. A study on designing video tutorials for promoting security features: A case study in the context of two-factor authentication (2fa). *International Journal of Human–Computer Interaction*, pages 1–16, 2017.

[14] Y. Albayram, M. M. H. Khan, T. Jensen, and N. 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. In *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[15] A. Alkhunaizan and S. Love. Effect of demography on mobile commerce frequency of actual use in saudi arabia. In *Advances in Information Systems and Technologies*, pages 125–131. Springer, 2013.

[16] A. Alzahrani and K. Alomar. Information security issues and threats in saudi arabia: A research survey. *International Journal of Computer Science Issues (IJCSI)*, 13(6):129, 2016.

[17] Anas. Five steps to protect your data from spyware, theft and loss on android. https://ardroid.com/5-steps-to-secure-your-android-phone/. Accessed: 2018-02-07.

[18] J. Blythe, J. Camp, and V. Garg. Targeted risk communication for computer security. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, pages 295–298. ACM, 2011.

[19] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs, and S. Schechter. Your attention please: designing security-decision uis to make genuine risks harder to ignore. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, page 6. ACM, 2013.

[20] J. M. Clark and A. Paivio. Dual coding theory and education. *Educational Psychology Review*, 3(3):149–210, 1991.

[21] O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[22] S. Egelman, M. Harbach, and E. Peer. Behavior ever follows intention: A validation of the security behavior intentions scale (sebis). In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5257–5261. ACM, 2016.

[23] S. Egelman, S. Jain, R. S. Portnoff, K. Liao, S. Consolvo, and D. Wagner. Are you ready to lock? In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 750–761. ACM, 2014.

[24] A. Field. *Discovering statistics using IBM SPSS statistics*. Sage, 2013.

[25] V. Garg, L. J. Camp, K. Connelly, and L. Lorenzen-Huber. Risk communication design: Video vs. text. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 279–298. Springer, 2012.

[26] S. Gazette. Mobile banking, not traditional banking, is what saudi customers want: Infographic. https://www.albawaba.com/business/mobile-banking-saudi-arabia-infographic-1021314. Accessed: 2017-09-13.

[27] M. Harbach, A. De Luca, N. Malkin, and S. Egelman. Keep on lockin'in the free world: A multi-national comparison of smartphone locking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4823–4827. ACM, 2016.

[28] M. Harbach, M. Hettig, S. Weber, and M. Smith. Using personal examples to improve risk communication for security & privacy decisions. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, pages 2647–2656. ACM, 2014.

[29] M. Harbach, E. Von Zezschwitz, A. Fichtner, A. De Luca, and M. Smith. Itsa hard lock life: A field study of smartphone (un)locking behavior and risk perception. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 9–11, 2014.

[30] C. Herron, H. York, C. Corrie, and S. P. Cole. A comparison study of the effects of a story-based video instructional package versus a text-based instructional package in the intermediate-level foreign language classroom. *Calico Journal*, pages 281–307, 2006.

[31] A. Jabir. What are the most significant risks faced by users? https://goo.gl/ti7y8L. Accessed: 2018-02-08.

[32] Y. Javed and M. Shehab. Investigating the animation of application permission dialogs: A case study of facebook. In *Data Privacy Management and Security Assurance*, pages 146–162. Springer, 2016.

[33] S. Kitayama, S. Duffy, T. Kawamura, and J. T. Larsen. Perceiving an object and its context in different cultures: A cultural look at new look. *Psychological Science*, 14(3):201–206, 2003.

[34] A. A. Madini and J. De Nooy. Cross-gender communication in a Saudi Arabian Internet discussion forum: Opportunities, attitudes, and reactions. *Convergence*, 22(1):54–70, 2016.

[35] P. Mayer, J. Kirchner, and M. Volkamer. A second look at password composition policies in the wild: Comparing samples from 2010 and 2016. In *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[36] I. Muslukhov, Y. Boshmaf, C. Kuo, J. Lester, and K. Beznosov. Understanding users' requirements for data protection in smartphones. In *Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference*, pages 228–235. IEEE, 2012.

[37] I. Muslukhov, Y. Boshmaf, C. Kuo, J. Lester, and K. Beznosov. Know your enemy: the risk of unauthorized access in smartphones by insiders. In *Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 271–280. ACM, 2013.

[38] T. Nguyen and N. Memon. Smartwatches locking methods: A comparative study. In *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[39] A. Nishi, N. A. Christakis, and D. G. Rand. Cooperation, decision time, and culture: Online experiments with American and Indian participants. *PLOS ONE*, 12(2):e0171252, 2017.

[40] D. J. Ohana, L. Phillips, and L. Chen. Preventing cell phone intrusion and theft using biometrics. In *Security and Privacy Workshops (SPW), 2013 IEEE*, pages 173–180. IEEE, 2013.

[41] M. R. Pattinson and G. Anderson. How well are information risks being communicated to your computer end-users? *Information Management & Computer Security*, 15(5):362–371, 2007.

[42] R. W. Rogers. A protection motivation theory of fear appeals and attitude change1. *The Journal of Psychology*, 91(1):93–114, 1975.

[43] S. B. Salter and D. J. Sharp. Agency effects and escalation of commitment: do small national culture differences matter? *The International Journal of Accounting*, 36(1):33–45, 2001.

[44] A. Samer. 14 tips for protecting and securing mobile or tablet. https://mobilesgate.com/secure-android-mobile-top-ways/17956.php. Accessed: 2018-02-07.

[45] D. Van Bruggen, S. Liu, M. Kajzer, A. Striegel, C. R. Crowell, and J. D'Arcy. Modifying smartphone user locking behavior. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, page 10. ACM, 2013.

[46] C. D. Wetzel, P. H. Radtke, and H. W. Stern. *Instructional effectiveness of video media*. Lawrence Erlbaum Associates, Inc, 1994.

[47] K. Witte. Putting the fear back into fear appeals: The extended parallel process model. *Communications Monographs*, 59(4):329–349, 1992.

APPENDIX

A. CUSTOMIZED-VIDEO TRANSCRIPT

تم تصميم هذا الفيديو لشرح بعض المخاطر الرئيسية لعدم حماية جوالك وكيف تحمي نفسك من تلك المخاطر

كما تعلمون، الجوالات تخزن قدرا كبيرا من المعلومات الخاصة بك مثل بطاقة المصرف، محادثات الواتساب، رسائل البريد الإلكتروني، الفيديوهات والصور والمواقع، وبعض المعلومات الخاصة بك.

في حال تم سرقه او فقدان جوالك الذي لا يحتوي على أي رمز للقفل، يصبح من السهل جدا لشخص ما الوصول إلى البيانات المخزنة في جوالك.
صحيح! مجرد تخيل كم هو سهل بالنسبة لشخص ما التقاط الجوال الخاص بك والوصول إلى جميع المعلومات المخزنة في الجوال.

على سبيل المثال، إذا وقع جوالك في أيدي خاطئة، يمكن للمهاجم البحث بسهولة من خلال البريد الإلكتروني الخاص بك او الرسائل النصية لكلمة " بنك الراجحي " في حال اذا كنت تستخدم الخدمات المصرفية عبر الإنترنت او التطبيق المحمل في إجوالك فبتالي المهاجم يمكن ان ينقر زر الدخول لتلقى رساله نصيه من البنك في جوالك ويتم ادخال كلمه السر المؤقتة والتحكم في حسابك المصرفي.

إذا كان حساب بريدك الإلكتروني مرتبطا بالعديد من الحسابات الأخرى عبر الإنترنت، فيمكن للمهاجم استخدام نفس التقنية للتحكم في حساباتك الأخرى أيضا

بالإضافة، المهاجم يستطيع استكشاف المعلومات الحساسة الخاصة بك من خلال رسائل البريد الإلكتروني، مثل رقم الهوية ،جواز السفر والملفات المرفقة كسيره الذاتية ومعلومات بطاقات الائتمان للبنك وتاريخ الميلاد وكلمه المرور

إذا كان المهاجم يستطيع الحصول على هذه المعلومات كالهوية او جواز السفر فباستطاعته بيعها لصوص الهوية، أو انتحال شخصيتك بتقديم طلب للحصول على بطاقة ائتمان جديدة

إذا كان الهاتف يحتوي على صور لك أو لأهلك، يمكن للمهاجم استخدامها لابتزازك مقابل المال أو تدمير سمعتك عن طريق نشر .الصور على الانترنت أو إرسالها إلى جميع جهات الاتصال الخاصة بك

علاوة على ذالك، المهاجم بإمكانه استخدام بعض التطبيقات مثل الفيسبوك أو الواتساب لإرسال رسائل إلى أصدقائك أو غيرهم، يتظاهر بكونك انت ويسأل عن المال. بل يمكن أن يطلب منهم أن يأتوا إلى أماكن معينة كحالة طارئة

هذه الاحداث تسلط الضوء فقط على عدد قليل من المخاطر المشتركة الناتجة عن عدم قفل الجوال الخاص بك. فبتالي نستطيع تجنبها بسهولة باستخدام أي من آليات قفل الشاشة الأمنة المتاحة على الجوال الخاص بك مثل رمز الدخول، الرسم، كلمة المرور بشدة بها من قبل خبراء الأمن. هذه هي تدابير أمنية بسيطة لضمان عدم وصول أي شخص إلى أو بصمه الأصابع التي يوصى المحتويات المخزنة في جوالك دون إذنك

من السهل إعداد قفل الشاشة وعادة ما يستغرق أقل من دقيقة. ، على سبيل المثال لجوال الأفيون، يمكنك الانتقال إلى \ الاعدادات ثم اختيار البصمه ورمز الدخول |وبعدها يتم تفعيل قفل الشاشه

من ناحية انظمه الاندرويد الانتقال الى الاعدادات ثم اختيار خيار الامن ثم اختيار قفل الشاشة وبعدها يمكنك اختيار أي اليه لقفل الشاشة
أتمنى ان هذا الفيديو ساعدكم بادراك اهميه استخدام قفل الشاشة وتشجيعكم باستخدام اليه قفل الشاشة المتوفرة في جوالك
نشكركم على مشاهده الفيديو

B. ARABIC QUESTIONER

First Round of the Main Study
الدراسة الأساسية

Saudi background Questions (5 questions)

Q1) نوع الجنس
أ- ذكر
ب- انثى

Q2) كم عمرك بين هذه القيم
أ- اقل من ٢٠
ب- ٢٠-٢٩
ت- ٣٠-٣٩
ث- ٤٠-٤٩
ج- ٥٠-٥٩
ح- ٦٠-٦٩
خ- فوق ٧٠

Q3) مستوى التعليم
أ- امي (لا يقراء ولا يكتب)
ب- ابتدائية
ت- المتوسطة
ث- الثانوية
ج- جامعي
ح- ماجستير او دكتوراه

Q4) خلفيتك في استخدام الكومبيوتر
أ- لاشي
ب- منخفض
ت- وسط
ث- عالي

Q5) ماهي اللغة التي لديك
أ- العربي
ب- الانجليزي
ت- كلاهما

Saudi smartphone usage behavior Questions (5 questions)

Q1) ما هو نظام التشغيل المستخدم في جوالك؟
أ- IOS (الأيفون)
ب- Android (سامسونج)
ت- نظام اخر

Q2) يرجى تقدير عدد الساعات التي تستخدم الهاتف الذكي خلال اليوم
أ- 1-2 ساعتان
ب- 3-4 ساعات
ت- 5-6 ساعات
ث- أكثر من ست ساعات

Q3) كم عدد التطبيقات تم تنزيلها في جوالك
أ- 1-3 تطبيقات
ب- 4-6 تطبيقات
ت- أكثر من 6 تطبيقات
ث- لا يوجد

Q4) يرجى تقدير عدد الساعات التي يتم فيها استخدام تطبيقات الجوال (سناب شات، واتساب .....)
أ- 1-3 مرات
ب- 4-6 مرات
ت- أكثر من 6 مرات
ث- لا يوجد

Q5) ما تطبيقات التي تستخدمها يوميا
أ- واتساب
ب- فيسبوك، توتير
ت- سناب شات
ث- جميع التطبيقات
ج- تطبيقات اخرى

Online security behavior Questions (3 questions)

Q1) هل تشعر بالقلق في حاله تعرض حساباتك على الإنترنت للخطر أو الاستيلاء عليها
أ- نعم
ب- لا

Q2) هل تشعر بالقلق حول الامن باستخدامك الانترنت
أ- نعم
ب- لا

Q3) هل تستخدم برامج الحماية ضد الفايروسات في جوالك؟
أ- نعم
ب- لا

Reasons for Not Using Lock Screen (2 questions)

Q1) لماذا لا تستخدم احدى تقنيات قفل الشاشة لجوالك؟
أ- ليس لدي معرفه للإعدادات تفعيل قفل الشاشة
ب- لا توجد أي مخاطر
ت- لا يوجد شيء لإخفائه
ث- أنسى
ج- استخدام قفل الشاشة مزعج
ح- سبب اخر

Q2) برائيك، لماذا بعض الاشخاص يستخدمون احدى تقنيات قفل الشاشة لجوالاتهم؟

Video Evaluation (7 questions)

Q1) ما هو مستوى اقتناعك في هذا الفيديو؟
أ- غير مقنع
ب- نوعا ما مقنع
ت- مقنع بشكل متوسط
ث- مقنع بشده

Q2) ما هو مستوى المنطق في هذا الفيديو؟
أ- غير منطقي
ب- نوعا ما منطقي
ت- منطقي بشكل متوسط
ث- منطقي بشده

Q3) ما هو مستوى الفعالية في هذا الفيديو؟
أ- غير فعال
ب- نوعا ما فعال
ت- فعال بشكل متوسط
ث- فعال بشده

Q4) هل الفيديو جعلك مدرك للبيانات المخزنة في جوالك؟
أ- نعم
ب- لا

Q5) هل الفيديو جعلك قلق بخصوص امن وحماية جوالك؟
أ- نعم
ب- لا

Q6) ما لسمات التي عجبتك في هذا الفيديو؟
ت- اللغة
ث- شرح المخاطر
ج- المحتوى
ح- البساطة
خ- الصور
د- عرض طريقه استخدام قفل الشاشة
ذ- سبب اخر

Q7) ما لسمات التي لم تعجبك في هذا الفيديو؟
أ- لا شيء
ب- اللغه
ت- قله المعلومات
ث- المحتوى ممل
ج- المحتوى طويل
ح- الصور
خ- سبب اخر

Effect of Fear Appeal on Response Efficacy (5 questions)

Q1) استخدام شاشه القفل، سوف تكون فكره جيده
أ- (1) غير موافق بشده
ب- (2) غير موافق
ت- (3) عادي
ث- (4) موافق
ج- (5) موافق بشده

Q2) اعتقد ان شاشه القفل سهله الاستخدام
أ- (1) غير موافق بشده
ب- (2) غير موافق
ت- (3) عادي
ث- (4) موافق
ج- (5) موافق بشده

Q3) اعتقد ان تفعيل شاشه القفل، سوف يوفر الأمان لجوالي
أ- (1) غير موافق بشده
ب- (2) غير موافق
ت- (3) عادي
ث- (4) موافق
ج- (5) موافق بشده

Q4) فاهم ماهي فوائد استخدام شاشه القفل
أ- (1) غير موافق بشده
ب- (2) غير موافق
ت- (3) عادي
ث- (4) موافق
ج- (5) موافق بشده

Q5) اعتقد ان تفعيل شاشه القفل سوف يحمي بياناتي في جوالي
أ- (1) غير موافق بشده
ب- (2) غير موافق
ت- (3) عادي
ث- (4) موافق
ج- (5) موافق بشده

Second Round of the Follow up Study
الدراسة التاليه

Q1) هل فعلت رمز شاشه القفل لجوالك
أ- نعم
ب- لا

إذا الجواب نعم
Q2) ما لذي حمسك لتفعيل شاشه القفل على جوالك؟

أ- اهميه الأمان والخصوصية
ب- الفيديو
ت- تجربه سيئة
ث- أخرى

Q3) ما لذي حمسك لتفعيل شاشه القفل على جوالك ومتى فعلته ؟

Q4) ما هو نوع قفل الشاشه الذي تم تفعيله ؟

أ- البصمه
ب- الباسكود / البصمه للايفون
ت- الرمز السري لسامسونج
ث- النمط
ج- نوع اخر

Q5) كيف كان قفل الشاشة الذي تم تفعيله؟

أ- مريح
ب- غير مريح

إذا الإجابة لا
Q2) 2 لماذا لم تفعل شاشه القفل على جوالك؟

أ- انخفاض الادراك لتهديد
ب- استخدامه مزعج
ت- لا يوجد شيء لإخفائه
ث- ينسى
ج- عدم معرفه اعدادات تفعيل شاشة القفل
ح- سبب اخر

Effect of Fear Appeal on Perceived Data Value (2 questions)

Q1) هل تعتقد أن البيانات المخزنة على جوالك ذو قيمه لحمايته؟

أ- نعم
ب- لا

Q2) كم من البينات الخاصة والمهمة تخزنها في جوالك
أ- لا شيء على الإطلاق
ب- كميه منخفضة من المعلومات الخاصة
ت- كميه معتدلة من المعلومات الخاصة
ث- قدرا كبيرا من المعلومات الخاصة

Effect of Fear Appeal on Security and Privacy Concerns (3 questions)

Q1) ما هو مدى قلقك بشأن أمن الجوال الخاص بك؟
أ- (1) لست قلق على الإطلاق
ب- (2) نوعا ما قلق
ت- (3) بعض الأحيان قلق
ث- (4) قلق بشكل كبير

Q2) ما هو مدى قلقك بشأن خصوصيه الجوال الخاص بك؟
أ- (1) لست قلق على الإطلاق
ب- (2) نوعا ما قلق
ت- (3) بعض الأحيان قلق
ث- (4) قلق بشكل كبير

Q3) ما هو مدى قلقك بشأن استخدام الغرباء الجوال الخاص بك؟
أ- (1) لست قلق على الإطلاق
ب- (2) نوعا ما قلق
ت- (3) بعض الأحيان قلق
ث- (4) قلق بشكل كبير

Effect of Fear Appeal on perceived severity and risk awareness (3 questions)

Q1) ما هو مدى تأثير التدمير لفقدان او ضياع جوالك على حياتك اليومية؟
أ- (1) لا يوجد أي تأثير على الإطلاق
ب- (2) نوعا ما يوثر
ت- (3) بعض الأحيان يؤثر
ث- (4) يؤثر بشكل كبير

Q2) ما هو مدى احتمال فقدان جوالك؟
أ- (1) غير محتمل بشكل كبير
ب- (2) نوعا ما غير محتمل
ت- (3) نوعا ما محتمل
ث- (4) محتمل بشكل كبير

Q3) ما هو مدى احتمال وصول أي شخص لجوالك؟
أ- (1) غير محتمل بشكل كبير
ب- (2) نوعا ما غير محتمل
ت- (3) نوعا ما محتمل
ث- (4) محتمل بشكل كبير

Effect of Fear Appeal on Response Cost (3 questions)

Q1) في حال لو استخدمت شاشه القفل، سوف يكون صعب بنسبه لي
أ- (1) غير موافق بشده
ب- (2) غير موافق
ت- (3) عادي
ث- (4) موافق
ج- (5) موافق بشده

Q2) في حال لو استخدمت شاشه القفل، سوف يكون غير مريح في كل مره ادخل رمز القفل للجوال
أ- (1) غير موافق بشده
ب- (2) غير موافق
ت- (3) عادي
ث- (4) موافق
ج- (5) موافق بشده
من فضلك اشرح في جمل قصيره ما سبب اختيارك

Q3) في حال لو استخدمت شاشه القفل، سوف يكون غير مريح لصعوبة تذكر رمز القفل للجوال
أ- (1) غير موافق بشده
ب- (2) غير موافق
ت- (3) عادي
ث- (4) موافق
ج- (5) موافق بشده

Action Needed! Helping Users Find and Complete the Authentication Ceremony in Signal

Elham Vaziripour, Justin Wu, Mark O'Neill, Daniel Metro, Josh Cockrell,
Timothy Moffett, Jordan Whitehead, Nick Bonner, Kent Seamons, Daniel Zappala
Brigham Young University
elhamvaziripour@byu.edu, justinwu@byu.edu, mto@byu.edu, joshuakcockrell@gmail.com,
danieltmetro@gmail.com, timothytmoffett@gmail.com, jordan9001@gmail.com, j.nick.bonner@gmail.com,
seamons@cs.byu.edu, zappala@cs.byu.edu

ABSTRACT

The security guarantees of secure messaging applications are contingent upon users performing an authentication ceremony, which typically involves verifying the fingerprints of encryption keys. However, recent lab studies have shown that users are unable to do this without being told in advance about the ceremony and its importance. A recent study showed that even with this instruction, the time it takes users to find and complete the ceremony is excessively long—about 11 minutes. To remedy these problems, we modified Signal to include prompts for the ceremony and also simplified the ceremony itself. To gauge the effect of these changes, we conducted a between-subject user study involving 30 pairs of participants. Our study methodology includes no user training and only a small performance bonus to encourage the secure behavior. Our results show that users are able to both find and complete the ceremony more quickly in our new version of Signal. Despite these improvements, many users are still unsure or confused about the purpose of the authentication ceremony. We discuss the need for better risk communication and methods to promote trust.

1. INTRODUCTION

Numerous secure messaging applications [18] have been developed to provide end-to-end encryption for personal communication. These applications typically automate the encryption process as much as possible, in order to provide a simpler experience for their users. However, the confidentiality provided by these applications relies on the integrity of its central servers, which exchange users' public keys automatically. To protect against a man-in-the-middle attack, either through compromise of the server or other means, users need to verify the exchanged keys with their conversation partners. This is typically done by comparing a fingerprint of the public keys. We refer to this verification process as the *authentication ceremony*, and variations of it have been adopted widely in secure messaging applications.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

Research using lab studies has reported that users have difficulty performing the authentication ceremony within secure messaging applications [4], and this makes them susceptible to attack [14]. Two recent papers demonstrated that with some instruction about the ceremony itself [8] or the importance of comparing keys [20], users can successfully find and use the authentication ceremony. However, users still took an inordinate amount of time—over 11 minutes on average—to find and complete the ceremony [20].

In this paper we examine whether *opinionated design* can make it easier for users to find and perform the authentication ceremony, without relying on instruction about the importance of the ceremony or providing any details about how the ceremony works. Our use of opinionated design is inspired by work on the security indicators for the Chrome browser [6], which led to greater adherence to SSL warnings, but not necessarily greater comprehension. We apply opinionated design to the Signal messaging application, seeking to make the minimal set of changes needed to encourage users to find and perform the ceremony. Our design principles follow recommendations from Schröder et al. [14] in their study of the Signal application. We seek to improve both adherence and performance with respect to finding and using the authentication ceremony, with comprehension a secondary goal. We use Signal because it is open source and because it has been at the forefront of this space, having pioneered the Signal protocol that is also used in WhatsApp, Facebook Messenger, Allo, and Skype.

To test the effectiveness of our design, we created two modifications of Signal, which we label Modification 1 and Modification 2. Modification 1 focuses only on helping users find the authentication ceremony, and the ceremony itself is unchanged. Comparing this version to the original version of Signal enables us to test whether it leads to greater adherence, while also providing a baseline for performance with the original ceremony. Modification 2 incorporates all the changes from the first, and also updates the authentication ceremony to make it easier to use. Comparing Modification 2 to Modification 1 enables us to test for differences in performance among the two authentication ceremonies. We used a between-subject lab study to evaluate the impact of these modifications. We encouraged participants to be security minded by promising them a small monetary bonus. We then observed participant actions, measured their accuracy and time to complete the task, and conducted interviews to

understand their comprehension of the ceremony and their opinions regarding the ceremony.

Our findings include:

- Our modifications of the Signal use interface led to 90% of participants finding the authentication ceremony on their own, combining results for Modification 1 and 2. These modifications included visual cues in the Signal conversation screens to indicate the authentication status of users' contacts, with accompanying actions to initiate the authentication ceremony. Most participants found the authentication ceremony in less than a minute, often within a few seconds. This is compared to a 25% discovery rate for the authentication ceremony among those who used the original version of Signal.
- Our redesigned authentication ceremony was successfully completed by 90% of participants who used Modification 2, as compared to 30% for the original ceremony in Modification 1. The new ceremony clearly separates a QR-code method (for in-person authentication) from a phone call method (when contacts are not in the same location), and uses an in-app phone call modeled after Viber's ceremony. The median time to complete the new authentication ceremony was 2 minutes, as compared to 7 minutes for the few who actually completed the original authentication ceremony.
- Our use of opinionated design, combined with an incentive to be security-minded, resulted in equal or better results than the study by Vaziripour et. al [20], which relied on directly instructing users about the importance of comparing keys. The success rate of 90% is better than the 78% who were successful across all participants and applications in their work, and comparable to the 96% success rate they saw with Viber. Moreover, the time to find the ceremony and complete the ceremony in our modifications (less than a minute, median of 2 minutes) is much lower than in their work (3.5 minutes and 7.8 minutes, respectively).
- Comprehension of the purpose of the ceremony is mixed. Many users associate the ceremony with authentication and confidentiality, but express doubts about their answers. Others clearly do not know what the purpose of the ceremony is. Likewise, while many users express trust in Signal, with further probing many indicate a lack of knowledge or experience to really know if they should trust it. When the purpose of the authentication ceremony is explained to participants, they mostly express a desire to use it, though one third would only use it for some content or with some contacts. This leaves room for future work to further improve the authentication ceremony.

Artifacts: We have created a companion website at `https://action.internet.byu.edu` that provides the source code, study materials, and data.

2. RELATED WORK

The usability of the authentication ceremony for secure messaging applications is a relatively new topic in the field usable security. To the best of our knowledge, there are currently only five papers focused on this topic [20, 14, 4, 8, 1]. The common conclusion of these works is that users are vulnerable to attacks and cannot locate or perform the authentication ceremony without sufficient instruction. This is largely due to users' incomplete mental model of threats and usability problems within secure messaging applications.

Our work has been inspired by one of the most recent studies on the usability of the authentication ceremony in secure messaging applications by Vaziripour et al. [20]. In this work, the authors studied users' ability to locate and perform the authentication ceremony in WhatsApp, Facebook Messenger, and Viber. The first phase of this work instructed participants about potential threats, while the second phase added instruction concerning the necessity of the authentication ceremony. From the first to the second phase, the average ceremony success rate increased from 14% to 79%. It took users, on average, over 3 minutes to find the authentication ceremony and over 7.5 minutes to complete it when they succeeded in the second phase. We borrow some of the methodology from this work.

Our Signal modifications are informed by recommendations from a paper by Schröder et al. that studied the usability of Signal under attack conditions. This study revealed that security experts also are susceptible to man-in-the-middle attacks due to usability problems and incomplete mental models of security. Only seven out of 28 (25%) expert participants successfully authenticated their conversation partners [14]. Asal et al. asked 20 participants to complete authentication by available methods (fingerprint, shared secret, and QR code) in ChatSecure. Herzberg and Leibowitz showed in their study that the majority of users fail to perform the authentication ceremony, and that successes were difficult and time-consuming, even when participants were taught how to authenticate [8]. Abu-Salma et al. conducted a usability study on Telegram to show that the UI was a source of confusion when performing the authentication ceremony [1].

There are several works on the usability of the verification mechanism itself. Shirvanian et al. studied key verification performance by users performing authentication on remote and local conversation partners. They showed that users perform poorly under most key verification methods, especially in the remote case [15]. Independent of a particular application, Tan et al. compared eight representations of authentication material, including textual and graphical representations, with varying degrees of structure, in a simulated attack scenario [17]. They showed that graphical representations were relatively more susceptible to attack but were easy to use, and comparison of graphical forms was quick. Dechand et al. studied textual key verification methods, finding that users are more resistant to attacks when using sentence-based encoding as compared to hexadecimal, alphanumeric, or numeric representations [5]. Sentence-based encoding rated high on usability but low on trustworthiness on a post-study Likert scale.

Another important aspect of our work is the qualitative analysis of users' comments and thinking process to inspect their decision-making processes. A study by Google shows that redesign of Chrome's SSL warnings to promote safe decisions resulted in 30% more users making correct decisions, but found that user comprehension of threats remained low. The authors hypothesized that if users understood the risks better, they would not ignore warnings. [6]. Cormac Herley calculated that the economic cost of time users spend on

Design Principle	Modification 1	Modification 2
Awareness of security status of conversations	Added verification status in conversation list and view (Figures 2a, 2b)	Same as Modification 1
Comprehensible instructions for recommended actions	Added instruction to visit verification screen via button (Figure 2b)	Same as Modification 1 + Separate in-person and remote authentication walkthroughs (Figures 3, 4)
Clear risk communication	None	Inform users of additional actions needed to secure conversations (Figures 3a, 4d)
Easily accessible verification	Clickable action bar in conversations (Figure 2b)	Same as Modification 1 + Clickable action bar in conversations (Figure 3a) and walkthrough (Figures 3, 4)

Table 1: Description of our application of Schröder's design principle recommendations

standard security is substantially higher than the benefits they incur. He argues that users' rejection of security advice is therefore rational economically [7]. Implications for nudging users toward more beneficial and secure choices have been considered recently [3]. Angela Sasse argues that security mechanisms with a high false-positive rate undermine the credibility of security and train users to ignore them [13].

3. MODIFYING SIGNAL

Schröder et al. found several problems with the usability of Signal under attack conditions [14]. They recommend four design principles to overcome these obstacles: awareness of conversation security status, comprehensible instructions for recommended actions, clear risk communication, and easily accessible verification. We applied these principles to redesigning the Signal application and evaluated their effect with a user study. Table 1 outlines our modifications and how they correspond to Schröder's recommendations. We created custom implementations of both the iOS and Android versions of Signal with these changes

We began by creating visual mockups of our modifications to Signal's interface that would employ three of our target design principles. In particular, we provided visual cues to the Signal conversation screens to indicate the verification status of users' contacts, with accompanying actions to initiate the authentication ceremony. Signal already employs a rudimentary indication of verification status in the form of a (hardly noticeable) checkmark under the names of verified contacts, but this is not easily associated with verification status and nothing is shown in the case where a user has not yet verified a contact. We were also careful not to overstate vulnerabilities in our visual cues, in line with recommendations from Sasse [13]. We showed these mockups to 40 university students to gather feedback for various designs, which varied in their use of icons, colors, phrasing, and position of verification status cues and options. We settled on the design as shown because it performed best in our mockups and provided clear warnings. We also used the Signal color scheme and terminology (e.g., *safety number*) for consistency with the original version. Next, we performed a cognitive walkthrough on the modified application to make sure the language used in the interface was clear. Once we were confident in our design, we made the necessary modifications to Signal to implement it. These changes comprise our first modification of Signal (Modification 1).

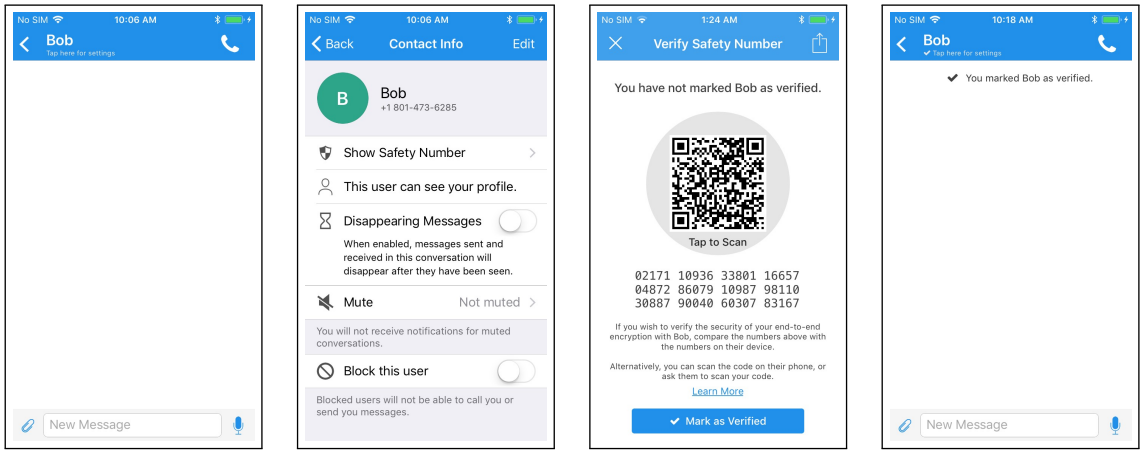
Our second modification of Signal (Modification 2) incorporated all of the changes of the first, but added a set of instructions for users to follow that streamline the authentication ceremony process. In a study by Vaziripour et al. [20], users were more successful performing the authentication ceremony in Viber, and did so in less time compared to other apps. We hypothesize that this was due to Viber providing an in-app phone call that presented encryption keys to users for verification on the same screen. Accordingly, we separate the QR code and phone call verification options in Signal, provide in-app functionality for verification phone calls, and incorporate guiding dialogue to successfully perform verification in each scenario. To develop this second variant of Signal, we conducted a set of pilot studies. We learned that users expect to be able to scan the QR code on each other's phone simultaneously, which could not be done using the original version of Signal. As a result, we modified the application to use a new dual camera/QR code screen.

3.1 Original Signal

Signal [16] uses a Double Ratchet algorithm [12] to update session keys with each exchanged message, which provides forward secrecy for the conversation. Before initiating the ratchet, it uses a triple Diffie-Hellman (3-DH) handshake to exchange public keys. This exchange is automated using a central server. To avoid a man-in-the-middle attack, users must verify the authenticity of the public keys that have been exchanged by the central server. Under Signal, the authentication ceremony is performed using fingerprints from a combination of a user's public key and his/her contact's public key. This fingerprint is called a *safety number*.

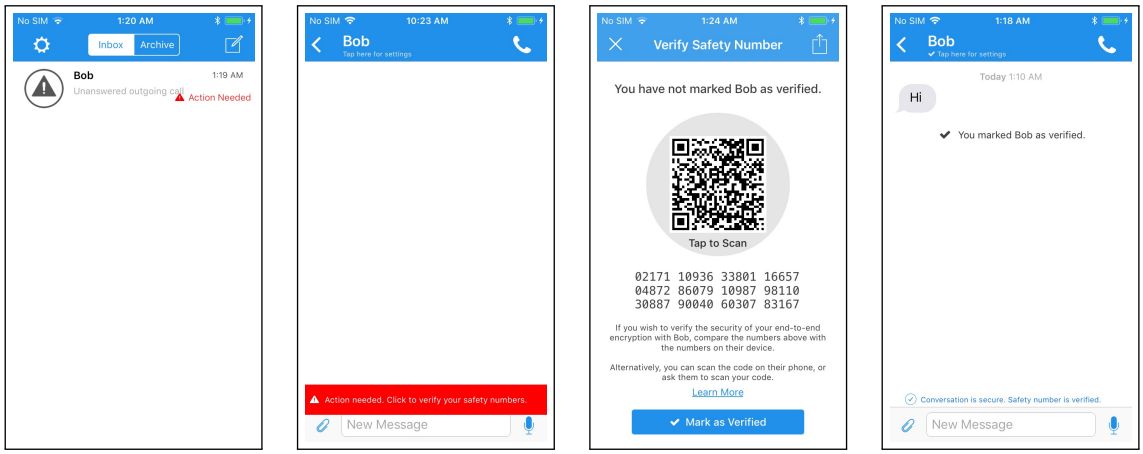
Figure 1 shows the workflow for the authentication ceremony in the current version of Signal. In the conversation screen, after a conversation is initiated with any contact, users can tap on a contact's name in the conversation screen shown in Figure 1a. At this point an option labeled *Show Safety Number* is found, shown in Figure 1b. By selecting this option, users will be transferred to the screen shown in Figure 1c, wherein two options are given to perform the authentication.

Users can either compare their safety numbers directly using their numeric representations, or by scanning an equivalent QR code displayed on their contact's device. After users verify that their safety numbers are equivalent, they are ex-



(a) (b) (c) (d)

Figure 1: Authentication ceremony within the current Signal application



(a) (b) (c) (d)

Figure 2: User interface for finding the authentication ceremony and showing successful verification (Modification 1)

pected to toggle a UI switch captioned *verified*, also shown in Figure 1c, to indicate that they manually verified the numbers to be identical. If users choose to scan the QR code and the result is a successful match, the *verified* switch is changed automatically. Next to the name of verified contacts, the interface places a check mark, shown in Figure 1d, which confirms that the contact has been verified and can be trusted to have a secure conversation with, through the Signal application.

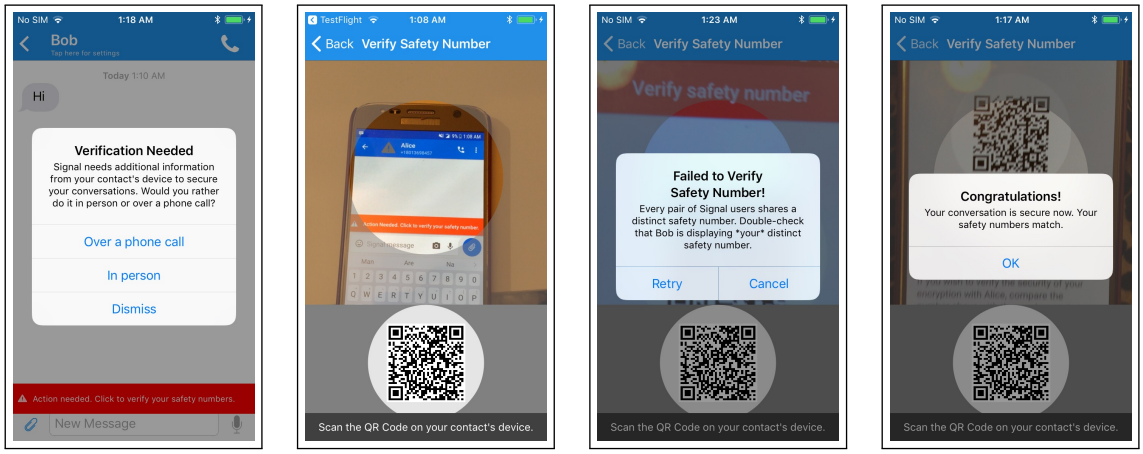
If the encryption keys change for this contact, due to reinstalling the application or a man-in-the-middle attack, users will be prompted to redo the verification process.

3.2 Modification 1

Modification 1 was designed to facilitate the process of finding the authentication ceremony. Users are prompted to perform the authentication ceremony in two locations, as shown in Figure 2. First, in the list of contacts, shown in Figure 2a, any unverified contact has a warning tag indicating *Action Needed*. We also replaced the profile image of unverified contacts with a warning icon until they are verified. Second, in the conversation view depicted in Figure 2b, if the contact is not verified, the bottom of the screen contains a red warning banner with the text *Action needed! Click to verify your safety numbers*. If users notice the red warnings and press either of them, they are directed to the original authentication ceremony screen, shown in Figure 2c. After successful verification, a check mark appears next to the contact name, the red warning band disappears, and each are replaced by a blue message which indicates that the contact has been verified. This text for a conversation window is shown in Figure 2d. The profile image of this contact also shown in favor of the alert icon. Note that in this version users still use the original authentication ceremony.

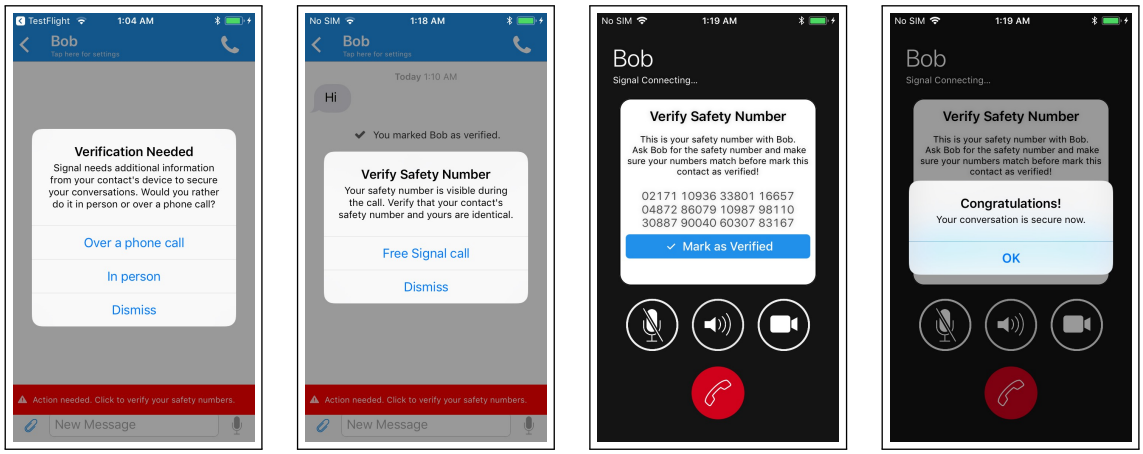
3.3 Modification 2

Our second variant of Signal, Modification 2, was designed to reduce the time required to perform the authentication ceremony. We also attempted to enhance participant under-



(a) (b) (c) (d)

Figure 3: Authentication ceremony for scanning the QR code (Modification 2)



(a) (b) (c) (d)

Figure 4: Authentication ceremony for comparing safety numbers using a phone call (Modification 2)

standing of the purpose of the ceremony, while not necessarily understanding the details of its inner workings.

We separated the two options of scanning the QR code and verifying the safety numbers. Figure 3 and 4 show these modifications. When users press the red warning within the conversation windows, a small dialog appears, shown in Figure 3a, informing users that the verification is necessary for the security of their conversation. They are given two choices of performing the authentication: over a free phone call (via Signal) or in person (QR code scan).

If users choose to verify the safety number in person, as shown in Figure 3, they will be directed to the screen shown in Figure 3b, with the camera activated. In this screen, the local QR code is also shown, allowing the user and his/her contact to scan and verify the safety numbers simultaneously. If the authentication fails, users are given another chance to scan the correct QR code, shown in Figure 3c.

If users instead choose to verify the safety number over a phone call, they will be informed that the call will be free, shown in Figure 4b. We modified the call screen such that immediately after initiating the phone call, users see their safety number with a very brief instruction, shown in Figure 4c. Users are expected to read their safety numbers and ensure they have an identical sequence of numbers. We use a phone call from within Signal because this allows users to see the safety numbers while making a call. Afterward, users press the *Mark as verified* button (iOS) or flip the toggle (Android).

We noticed during pilot studies that users lacked feedback after a successful verification. As a result, contacts who have been verified by the user have a *Verified* tag next to their names in the conversation list, instead of an *Action needed* tag. In addition, the profile image is loaded. During the pilot studies we noticed that users also need feedback to make sure they completed the ceremony correctly, so we created a short congratulation message, shown in Figures 3d and 4d.

4. METHODOLOGY

We conducted an IRB-approved, between-subject user study, examining how participant pairs locate and complete the authentication ceremony across three versions of the Signal secure messaging application. These three versions are the current version of Signal, Modification 1 (with changes to prompt the user to find the authentication ceremony), and Modification 2 (with additional changes to improve the usability of the authentication ceremony). Our study materials are shown in Appendix B.

In the study, we asked participants to complete a scenario wherein one participant needed to send a credit card number to the other participant. The base pay for the study was $7 per participant, with a $3 bonus if they performed the task safely. To avoid any hurt feelings, all participants were given the bonus, but in our observations the bonus served to sufficiently motivate participants to act securely.

We also wanted to test whether we followed Krug's first law of usability—*Don't make me think!* [10]. Thus, we did not provide participants with any instructions on the necessity of performing the authentication (in contrast to [20]), nor did we give them instructions on how to find or complete the authentication ceremony. We gave each participant a time limit of 10 minutes to complete the task, though they were not aware of this limit in advance.

To test each version of Signal equally, we assigned each pair of participants to one of the versions in a round robin manner. Prior to conducting the study a power analysis (described in Appendix A) indicated we needed 10 pairs of participants for each version. During the study subjects installed and used the Signal version to be evaluated on their own mobile devices. The original version of Signal was retrieved from the relevant official app stores for iOS and Android. We uploaded the Android versions of Modification 1 and Modification 2 to Google Play, and we used TestFlight for evaluating our Signal modifications on iOS.

4.1 Task design

In each experiment, the task provided to participants was as follows:

> *You left your credit card at home! You are going to be using the Signal app to ask your friend to send you the credit card number.*
>
> *This is the message you should send to your friend:*
>
> *"Hi! Can you send me my credit card number? I left my card on my desk at home."*
>
> *You can both earn a bonus of $3 for this study if you make sure that nobody can steal this information while your friend is sending it.*

Participant B was instructed similarly:

> *Your friend is going to use the Signal app to ask you for their credit card number. Use the credit card given to you by the study coordinator.*
>
> *You can both earn a bonus of $3 for this study if you make sure that nobody can steal this information while you're sending it.*

Despite a difference in roles, our intention was for both participants to complete the authentication ceremony. Participants were instructed to "talk aloud" as they performed the task, explaining their observations, actions, and reasoning.

Participants failed the task if they sent the credit card number before performing the authentication ceremony correctly, or ten minutes elapsed before completion of the task. In failure cases, participants still performed post-task duties such as responding to questionnaires and interview questions.

During the study, the coordinators checked whether participants had performed the authentication ceremony correctly. If the participants were successful, the coordinators recorded the method used (QR code or comparing the fingerprints verbally in a phone call). If the participants were not successful, the coordinators recorded the reason why.

4.2 Study questionnaire

Participants used a web-based Qualtrics survey on a laptop during the study. This survey both recorded participant answers to various questions both before and after the task, and also briefed them on the task itself. The survey contained:

- A standard set of demographic questions.
- A description of the primary study task, involving the exchange of a credit card number.
- A question asking if the participant believed they had exchanged the credit card number safely, followed by a free-response question to explain the answer.
- A question asking if the participant had seen the authentication ceremony screen (depicted by a screenshot in the survey) during the task. If so, the survey asked the participant several followup questions.
- A question asking if the participant had previously used secure messaging applications to send sensitive information, and the nature of that information.
- A question asking if the participant trusted Signal to be secure, followed by an open-response question to explain the answer.
- A question to rank participant knowledge of computer security.

4.3 Post-study interview

At the conclusion of each study, the coordinators verbally asked each individual participant the following questions:

- We asked participants what features they were looking for to aid in accomplishing the task. This provided us with insight into reasons for success and failure.
- We showed participants how to find the authentication ceremony and asked them to explain how they thought this ceremony helped them (or would have helped them) accomplish the task.
- We asked participants whether they were willing to perform the authentication ceremony before exchanging information with their friends in the future.

We recorded the audio of each study and transcribed the post-study interviews. To analyze the data for open-response questions in the survey and interviews, two authors coded the data together using conventional content analysis. Any disagreements were resolved via discussion. First, we reviewed qualitative comments phrase-by-phrase and word-by-word to assign codes that classified users' comments with regards to a particular topic. Then, we used the constant comparative method to group codes into concepts and organized related categories by merging related codes.

4.4 Study recruitment and design

We recruited pairs of participants on our campus, telling them that each person needed to bring a friend, and that both participants needed to have smartphones. Recruitment proceeded from November 14, 2017, to January 28, 2018, with 41 unique participant pairs recruited in total: 10 pairs for testing each version, eight pairs for pilot studies, and three pairs for replacement. We had to replace the data for three studies, two because the participants had participated in similar studies recently and one because a participant's device had security software that warned them against using our modified version of Signal.

When participants arrived for their scheduled appointment, we presented them with the requisite forms for consent and compensation. We instructed them to download and install the Signal application being tested. We then read them a brief introduction describing the study conditions and their rights as study participants. We informed them that they would be placed in separate rooms. We also informed participants that a study coordinator would be with them at all times and would answer any questions they might have. We let participants choose the study coordinator they would be comfortable working with.

We led the participants to their respective rooms, initiated audio recording, and instructed them to begin the survey. Throughout the study, coordinators were available to answer general questions but were careful not to provide any instructions that would aid in the use of the applications. Sometimes, participants asked if they could meet, and we told them they could. The nature of the scenario led most participants to assume they would not meet.

4.5 Limitations

The scenario we gave participants to exchange a credit card number included telling participants to make sure that no one could steal their information. This caused confusion in one case, when the participants made a phone call through the app in order to perform the authentication ceremony, when they noticed that they could use the same phone call to exchange the credit card number. It may be better to create a scenario where users first validate the safety numbers, then are given a task to exchange the credit card number.

The iOS and Android versions are slightly different. The Android version in Modifications 1 and 2 tells the user that they need to send a message in Signal before they can verify safety numbers. This message appears because the safety number is generated from a combination of local identities and remote identity public keys, and on Android the remote identity key is only received after exchanging the first message. For iOS, this is not the case, and safety number is available before any message exchange.

Due to our method of recruitment, our participants were largely students and their acquaintances, and subsequently exhibited some degree of homogeneity. All participants were between 18 and 34 years of age and had received at least some college education. This could cause absolute success rates or usability scores to be higher than in a broader population, though it should not affect comparisons among different versions of the application.

4.6 Demographics

Our participants were not balanced with respect to gender—50.0% (10) of our participants for the original Signal, 70.0% (14) of participants for Modification 1, and 35.0% (7) of participants for Modification 2 were male.

Since we distributed recruitment flyers on the university campus, most of our participants were undergrads, between 18 and 24—90.0% (18), 100.0% (20), and 90.0% (18) for each of the three versions. Most participants had some college but not yet earned a diploma—90.0% (18), 75.0% (15), and 65.0% (13) for the three versions.

Participants had a variety of backgrounds, skewed toward fields with non-technical backgrounds and less explicitly IT-related. Participants were asked to place themselves into categories of "beginner," "intermediate," and "advanced" regarding their security expertise. Most participants regarded themselves as beginners—85.0%(17), 85.0%(17), and 70.0% (14) for the three versions. None of our participants classified themselves as advanced, including the four participants from computer science or computer engineering.

5. RESULTS

In this section, we discuss the quantitative and qualitative results regarding the use of the authentication ceremony by participants. Details of our statistical methods are given in Appendix A.

5.1 Adherence and Completion

Participants who completed the ceremony compared their safety numbers by either scanning the QR code or by comparing the numbers over a phone call. We recorded a failure when participants transmitted sensitive data before verifying safety numbers, or if they failed to locate and validate safety numbers within ten minutes of launching the application. We also asked participants whether they felt they had safely exchanged the credit card number. Success and failure reports from both participants and the study coordinators are shown in Table 2.

Half of the participants who used the original Signal, and the majority of participants who used the modified versions, believed that they completed the task safely. However, none of the participants who used the original Signal version successfully performed the authentication ceremony. Only five participants even located the screen where safety numbers were displayed. In one of these cases, the participant ignored the instructions on the screen and simply pressed the *Mark as verified* button. In the other cases, participants ignored the screen entirely and immediately dismissed it. Participants tried several methods to deliver the message securely, including using various forms of primitive coding (e.g. developing their own substitution cipher), or enabling Signal's message impermanence feature.

	Participant self-report			*Study coordinator report*				
				Yes		**No**		
Application	**Yes**	**No**	**Not sure**	**QR code**	**Phone call**	**Not found**	**Ignored**	**Toggled**
Original	10	3	7	0	0	15	4	1
Modification 1	18	0	2	4	2	0	2	12
Modification 2	12	1	7	0	18	1	1	0

Table 2: Did the participants safely exchange the credit card number?

Application	**Time to locate authentication ceremony**	**Time to complete authentication**	**Time to complete the task**
Original	3.5	N/A	5
Modification 1	<1	7	8
Modification 2	<1	2	4

Table 3: Median time, in minutes, for finding and using the authentication ceremony.

All of the participants who used Modification 1 located the authentication ceremony screen, a large increase over the original Signal. However, while six participants correctly verified their safety numbers, the remaining 14 did not. Two of these latter participants ignored the screen and dismissed it, and the other 12 simply toggled the *Mark as verified* switch without comparing numbers. In successful cases, participants met to scan the QR code on each other's phone and in one case they wrote the safety numbers on paper and then made a phone call to verify them. We also notice that under this version of Signal, nearly all of the participants (18) believed they had safely performed the task. Only one of the participants who toggled the switch claimed to be unsure about the safety of the exchange.

Participant performance with Modification 2 was drastically better when compared to both the original Signal and Modification 1. Under Modification 2, 18 (90%) participants successfully performed the authentication ceremony, all of whom elected to do it over a phone call. The two failures were from the same pair of participants. In this case, Participant A erroneously informed his partner that the information had been transmitted safely, which caused Participant B to abandon his viewing of the authentication ceremony. However, Participant B did note that he was unsure the information was transferred safely in the post-task survey.

To test whether there are any differences between the versions of Signal, we used Cochran's Q test. We found that the success rate was statistically different for the applications ($\chi^2(2) = 27.11$, $p<.0005$). We then ran Barnard's exact test to find the significant differences among the pairs of applications. This test shows the differences among all the pairs are significant (Signal vs. Modification 1, $p = 0.0165$; Signal vs. Modification 2, $p = 1.15E-05$; Modification 1 vs. Modification 2, $p = 0.0163$).

5.2 Timing

The study coordinators timed each of participants and obtained three metrics, all with a granularity of minutes. First, the time required to locate the authentication ceremony was measured from the time that participants launch the application to the time where they first find the screen wherein the safety numbers reside. Second, the time for authentication completion was measured from the time users find the safety number screen to the time they verify their partner's safety number matches their own. Third, task completion time was measured from the time participants launch the application to the time they send (or receive) the credit card number from their partner.

Table 3 shows median times for each of the discussed metrics. For studies involving Modification 1, no one performed the authentication; thus we did not include this data in the table. In all of the studies with Modification 1 and Modification 2, all participants except for two discovered the authentication ceremony in less than 1 minute, with many taking just a few seconds. For the original version, only 5 (25%) of the participants found the screen, with a median of 3.5 minutes. Note the average discovery time in [20] was 3.2 minutes.

Participants correctly performed the authentication ceremony in 3 out of the 10 experiments with Modification 1, taking a median of 7 minutes. Participants correctly performed the authentication ceremony in 9 out of the 10 experiments with Modification 2, finishing in a median of 2 minutes. Note that the average time to complete the ceremony in [20] was 7.8 minutes.

For finding the ceremony, a two-tailed, two-sample t-test with equal variance shows there is no significant difference between Modification 1 and Modification 2 (p=0.484, 95% CI: [-0.37, 0.759] minutes). This is expected since the interfaces for finding the ceremony are identical in these two versions. For completing the authentication ceremony, a two-tailed, two-sample t-test with equal variance shows there is a significant difference between Modification 1 and Modification 2 (p=7.849E-05, 95% CI: [1.937, 6.16] minutes).

5.3 Usability

We asked participants who found the authentication ceremony to rank the usability of the ceremony on a five-point Likert scale, from *Extremely easy* to *Extremely difficult*. Table 4 shows the participant responses to this question. No one reported the task as extremely difficult and the majority of participants found it easy or somewhat easy to work with the authentication ceremony.

Note that the one who ranked the ceremony in the original Signal as *Extremely easy* to use simply toggled the *Mark as verified* switch. Of the nine participants using Modification 1 who reported it was extremely easy for them to use the ceremony, 5 either ignored it or toggled the *Mark as verified*

Application	Extremely easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Extremely difficult
Original	1	0	2	1	0
Modification 1	9	5	5	1	0
Modification 2	7	10	1	1	0

Table 4: Responses to: "How difficult or easy was it to use this screen to verify the safety number?"

switch, with the rest successfully completing the ceremony. All of the participants using Modification 2 saw the authentication ceremony screen and the majority believed it was easy to use. Because many of the participants either didn't use the ceremony or didn't complete it properly, we didn't run any statistical comparisons among the different versions.

We also asked these same participants what they liked or disliked about verifying their safety number, in an open-response question. Interestingly, some users felt the length of the safety numbers improved the security of the task, while others felt they were too long or hard to keep track of. This is well illustrated by the comment from one participant who used Modification 2:

> *"I liked that it came up on the middle of the phone call screen rather than being sent through a text message that I would have to pull up during the conversation. There were a lot of numbers, which could be hard to keep track of if you were reading them over the phone, but the amount of numbers ensures greater safety."*

The confusion regarding the original authentication ceremony is well illustrated by this comment from a participant who used Modification 1:

> *"I was a little confused at first and I wondered if we needed to be in the same room to scan the QR code to make sure our conversation was secure. At the bottom it just asked if I could switch the conversation to verified and so I did."*

Another participant who used Modification 2 stated: *"I liked how the numbers were large and visible but I didn't like that the numbers had to be read on speaker phone so everyone could have heard them.* This indicated some confusion about the role that safety numbers play in securing the conversation.

Note that we didn't make any statistical comparisons for this or other qualitative data in the paper. Our qualitative data is noisy, meaning some users may not have offered all their reasoning in a particular answer, while other users gave multiple reasons and were coded into multiple categories. In addition, because of a large number of categories, the values of many cells in the tables are small. These factors make statistical comparisons problematic.

5.4 Comprehension

During the post-task survey we also asked participants who found the authentication ceremony what they thought the screen did. Overall, 43 out of 60 participants answered this question. We also showed the screen to participants during the interview portion of the study, asking them how they thought the screen helped them with the task. We coded this data, with the results shown in Table 5.

Code	Original	M1	M2
(A) Survey			
Authentication	3	4	6
Confidentiality	2	2	3
Security	0	6	7
Trust	0	2	1
Didn't know	0	7	5
(B) Interview			
Authentication	7	7	6
Confidentiality	3	6	7
Security	2	1	2
Trust	2	0	0
Didn't know	5	7	5

Table 5: Coded responses to: (A) "What does this screen do?" (shown if they saw the ceremony during the study), (B) "How does this screen help you to accomplish the task?"

Many participants believed the authentication ceremony was involved with either authentication or confidentiality. Typically when mentioning authentication they discussed making sure they were talking to the right person and not an impostor. Some participants indicated a good level of understanding. For example, one participant who used Modification 2 said:

> *"That made sure that you weren't talking to someone pretending to be your friend or someone who had hacked her number and was answering the phone for her. Because there was not picture of her, no live stream video. So it could have been someone that sounded like her really closely. So I think that's what the numbers did...If numbers didn't match.It would mean that I would send him a message, and then his phone would try to unencrypt it, and it would just get garbage information."*

A participant who scanned QR codes with Modification 1 said:

> *"I think it helps, that, there were so many of them that its hard to replicate so I don't think that it would be easy for someone to just steal them or come up with them and so, cause there were enough of them that when I saw that (A) had all the same ones I was like 'Cool I am definitely talking to the person I think I am.' "*

When mentioning confidentiality, participants discussed making sure nobody else could read their conversation. A participant who used Modification 2 said: *"I was thinking for safety reasons. To make sure that the information we're telling to each other is just between the two of us."*

Participants also often mentioned security, generally, without any additional clarification about what it meant to have a

Code	Original	M1	M2
Positive impressions			
Use of primitive cipher	6	3	0
Trust in the application	4	9	1
Message impermanence	3	3	0
Use of other security features	2	0	0
Successful message delivery	1	0	0
Contact list synchronization	1	0	1
Trust voice call	1	3	2
Absence of physical threats	1	2	1
Authentication ceremony	0	7	11
Negative impressions			
Lack trust in the application	7	1	1
Lack of knowledge	4	2	4
Time cutoff reached	1	0	0
Lack of transparency	1	0	1
Lack of trust in mobile apps	0	1	0
Lack of trust in text	0	0	1
Possible physical threat	0	0	1

Table 6: Coded responses to: "Do you think you have safely exchanged the credit card number with your friend? Explain your answer."

"secure connection". There are also significant numbers who didn't know and, by their own admission, could not make a guess, or who were clearly making up an answer on the fly.

Note that many participants, across all codings, expressed doubt about their answers, as is typical in lab studies involving technical topics. The vast majority of participants were not entirely sure about the role that safety numbers played.

5.5 Participant Report on Success or Failure

During the post-study survey, participants were asked: "*Do you think you have safely exchanged the credit card number with your friend? Explain your answer.*" The available discrete responses were *Yes*, *No*, and *Not sure* (reported previously in Table 2), and an adjacent free response field required respondents to explain in their own words. We coded the free response portion of these answers into two groups: positive impressions and negative impressions. Positive impressions were used to support claims of success and negative impressions support claims of possible failure. Note that these categories are not mutually exclusive. For example, a persons unsure of their success in the task sometimes provided both positive and negative impressions. The number of responses in each identified category across all variants of the Signal application tested are shown in Table 6.

The use of a primitive cipher, such as writing the credit card number backwards, sending a screenshot instead of textual data, or mapping numbers to letters in the recipients name, was the most popular positive impressions for tasks under the original Signal. This was followed by trust in the application and the use of message impermanence settings. We see an increase in mentions (from 0 to 7) of the authentication ceremony from study participants who used Modification 1, which is then further increased by participants using Modification 2, with over half of participants mentioning this in their response. However, we also note that lack of knowledge seemed relatively unaffected by Modification 2 with respect to the original Signal.

In cases of failure to find the authentication ceremony or perform it correctly, participants were asked: "*What were you looking for to accomplish the task?*" By this time we had already showed them the authentication ceremony screen and its purpose, so this question allowed them to provide us with insight to what information they lacked during the study that would have helped them find and use the ceremony.

For users of the original Signal, this responses were largely aimed at explaining why they did not locate the ceremony. These reasons varied wildly, which in itself became the overall theme: participants lacked sufficient direction under the original Signal. One participant said *"I had no idea what I needed to do"* and another said he was *"just looking for any sort of security setting or application."* Some explained their method for ad-hoc cipher use, implying that they didn't look for built-in functionality to provide safety and instead resorted to their own means. Others explained that they got caught up experimenting with other security features of the application, such as message destruction, or blamed their own laziness for not finding the ceremony.

Since modified Signal versions effectively led the participants to the ceremony screen, responses provided insight into why the authentication ceremony may not have been performed properly. The primary difficulty in using Modification 2 was that participants had difficulty knowing what to do with the authentication ceremony screen and its "Mark as verified" button. For example, one participant remarked,

> *"I hit [the button] and then I was like, 'well that did nothing' and so I hit it again and nothing happened...I hit verify and then it says that I just unhit it immediately afterwards...I was just like, 'verify what? What am I verifying?' It didn't really tell me...Honestly it meant nothing."*

Our second modification was designed to deal with these problems by guiding users through the authentication ceremony. Only one pair was unsuccessful in properly performing the ceremony. The response to this question from that pair explained that the participants felt comfortable once they identified each other on the call and thus didn't proceed.

5.6 Trust

Participants were presented with the statement *I trust that Signal is secure*, and asked to rank their agreement with the statement on a five-point Likert scale ranging from *Strongly agree* to *Strongly disagree*. Table 7 shows responses to this question for the different versions of Signal. A one-way ANOVA shows that there are no significant differences between the different versions (p=0.143). We also asked participants to explain their answer in an open response question. We coded their positive and negative impressions, and this data is shown in Table 8.

The majority of participants somewhat or strongly agreed with the statement, with more users of Modification 1 and 2 expressing these sentiments as compared to the original version. Note that many of the people expressing trust in the application had no specific reason other than that the application seemed secure, or that it seemed more secure than other applications they had used. A number of people pointed to the authentication ceremony as a reason to trust the application, but this could be because they sensed this

Application	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
Original	4	9	6	1	0
Modification 1	5	13	2	0	0
Modification 2	6	12	2	0	0

Table 7: Responses to: "I trust that Signal is secure."

Code	Original	M1	M2
Positive impressions			
Seem secure	6	3	7
Relatively secure	1	2	0
Authentication ceremony	2	4	6
Security settings	1	1	0
No evidence to contrary	2	0	2
Trust university	0	3	0
Message impermanence	2	0	0
Few people using it so far	0	0	1
User interface	0	1	0
Negative impressions			
Lack of knowledge/experience	9	7	8
Lack of reputation	1	1	4
Lack of transparency	1	1	0
Lack of trust	1	1	0
Confusing user interface	0	2	0

Table 8: Coded responses to: "Please explain your answer" (regarding whether they trust Signal to be secure)

was the purpose of the study. One participant who used Modification 2 stated: *"I think this app is strongly agree because once you are verified with others you can actually trust the person on the call and exchange your information."* Several participants indicated they trusted the application because they assumed it had been made by developers at our university.

The only person who chose *Somewhat disagree*, for the original version of Signal, couldn't find the authentication ceremony and referred to lack of transparency as the reason for this choice. This participant said: *"There is no proof of this at all. It says it is secure but does not give me any information."* The main negative impressions expressed by participants were lack of lack of knowledge about the application or experience using it, and lack of reputation.

5.7 Adoption

During the interview portion of the study, we read participants the following statement:

> *"It is possible for someone to intercept your messages. These screens we have been showing you are called an authentication ceremony. Using the authentication ceremony ensures that nobody, not Signal, not hackers, and not even the government, is able to intercept your messages. You only need to do this once (or if your friend reinstalls the app). Now that you know this, are you willing to use the authentication ceremony before you exchange messages with a friend the first time?"*

We then asked participants if they would be willing to use the authentication ceremony in secure messaging applications in the future. Of the participants who answered this question, 32 said yes, 4 said no, 14 said only if they were exchanging confidential information, and 6 said only with certain contacts. We emphasize that participants were likely to say yes to this question, due to the nature of the study.

As an example of how we rated someone who would use the ceremony only when sending certain content, one participant who used Modification 1 said:

> *"Am I willing? Yes. Will I? No. Because here is the thing, I don't really care if my messages get intercepted because most of the time I am not sending my credit card number or social security numbers. Will I use it for things that are really important? For sure."*

6. DISCUSSION

In this section we discuss the significance and shortcomings of our results.

6.1 Adherence, Timing, and Comprehension

One of the primary contributions of this work is that the modifications that we made to Signal result in a higher success rate and lower task completion time in comparison to the original version. With Modification 1 and Modification 2 combined, 97.5% of participants found the ceremony, compared to only 25% for the original version of Signal. In addition, the changes made to the authentication ceremony in Modification 2 resulted in a success rate of 90% for completion of the ceremony, as compared to 30% for Modification 1. Numerous participants were confused by the *Mark as Verified* toggle in the ceremony for Modification 1 (the same as Signal's current ceremony), and assumed that flipping this switch would activate some kind of automatic verification.

Our results improve on prior work by Vaziripour et al. [20]. Participants found the authentication ceremony in an average of less than a minute (and often seconds), as compared to 3.5 minutes for [20]. Likewise, the average time to complete the authentication ceremony was 2.11 minutes, as compared to 7.8 minutes across the three applications [20] studied.

These advances were made using opinionated design to encourage participants to use the authentication ceremony, combined with a small monetary incentive to be security-minded. Our methodology included no instruction on finding or completing the ceremony, as in prior work [8, 20]. This indicates that the interface changes were enough to lead to the desired behavior, once participants had a security mindset.

Despite these results, participants did not demonstrate a strong comprehension of the purpose of the authentication ceremony. Although some participants believed the ceremony had something to do with authentication or confidentiality, many expressed doubts about their opinions. Still others either directly or indirectly admitted they didn't know what

it was for. As one participant who used Modification 2 stated, *"I don't know. I'm not really sure, actually, how it helped."*

Overall, these results are similar to a recent Google study on SSL warnings [6]. This study found that design of the warnings enhanced secure behavior from users and boosted threat understanding, but did not necessarily improve user comprehension of the warnings. This indicates that more work is needed to help users understand what they are doing in the authentication ceremony, and why they are doing it.

6.2 Adoption, Risk Communication, and Trust

Our interviews with participants indicate more work is needed within secure messaging applications to explain the purpose of the ceremony and to help users make choices about when it is necessary. Once the purpose of the authentication ceremony was explained to users, they readily understood it. However, a third of participants indicated that they would only want to use it certain in cases when they were sending sensitive information, and their responses indicated that they viewed the risk as acceptable when sending ordinary information. Others indicated they would never see the need, or said they would have trouble convincing their contacts to adopt secure messaging apps or use the ceremony.

A review of terminology used in Signal and in our modifications illustrates the difficulty. Our warning message to users reads "*Action needed! Click to verify your safety numbers.*" There is no indication of what comparing these numbers will do for users, nor what risks occur if they don't. Likewise, in the current Signal ceremony, it tells users that:

> *"If you wish to verify the security of your end-to-end encryption with Bob, compare the numbers above with the numbers on their device."*

Many users may not know what end-to-end encryption is, why comparing these numbers helps, nor what risks occur if they do not do this. Similar criticisms are valid for our modified ceremony.

In addition, users make rational tradeoffs between security and convenience [7]. Even if the ceremony is highly usable, users may still not adopt it, since usability is not the primary obstacle to adoption of secure messaging applications [2]. Rather, users may perceive the ceremony as "geeky" [9], they may not be convinced there is a need for it, or they may not be able to convince their contacts to use it.

Finally, many users readily admitted that they lacked the knowledge and experience necessary to know whether to trust Signal. The difficulty this poses for users was expressed well by one participant who used Modification 2:

> *"I don't know that there is anything that would make be sure that no one else is listening in. I don't know if whoever has developed Signal has someone set it up so that they can listen in. I would assume that they don't because it seems like their purpose is security. But I guess it might be possible for someone to be listening in. I don't know how I would know that that isn't happening."*

It's not clear how to give users a sense of trust in secure applications, especially when there are regular breaches of security that they hear about in the news.

6.3 Generality

Our results on finding and using the authentication ceremony should generalize to other secure messaging applications. We examined several major messaging applications to identify how our research would apply to them.

- *Finding the Authentication Ceremony:* WhatsApp, Telegram, Facebook Messenger, and Viber all require multiple clicks to find the authentication ceremony within the menu system, similar to Signal. With both Telegram and Facebook Messenger, encrypted chats are optional, so additional steps are needed to initiate a secure chat. We expect our improvements for finding the ceremony would be applicable to all of these applications.
- *Using the Authentication Ceremony:* The ceremonies in WhatsApp, Telegram, and Facebook Messenger differ in varying degrees from Signal. WhatsApp is nearly identical, with options for scanning a QR code or comparing an alphanumeric fingerprint, and no integrated phone call. Telegram allows the user to compare either a graphical or alphanumeric fingerprint, with no integrated QR scanning or phone call and few instructions. Facebook Messenger only offers the option to compare an alphanumeric fingerprint, and there are separate keys for each device, again with no integrated phone call. We expect our improvements for using the ceremony will be applicable to all of these applications. Viber is unique in that it integrates a phone call into their application to make the ceremony easier to use. Thus it is likely that Viber's ceremony would have similar success as our design. In prior work [20] Viber had the highest success rate for the authentication ceremony once people were directed to find it.

7. CONCLUSION

Our study indicates that users can find and complete the authentication ceremony in secure messaging applications, provided they have a security mindset and the application is designed to help them easily accomplish these tasks. This raises numerous open questions for further study. First, comprehension is still somewhat low, and additional design is needed to help users understand why they should perform the ceremony and when it is necessary. Second, it is not clear whether users will be security-minded without encouragement, such as a small monetary reward in the case of our study. More work is needed to determine if user interface changes alone can encourage use of the ceremony. Third, work is needed to determine if these advances can be applied to helping users cope with an attack scenario or when a contact re-installs Signal. Both of these situations will cause the security numbers to change, alerting users to a possible attack, and evidence to date shows that users do not cope well. Fourth, it may be possible to fully automate the authentication ceremony, using social authentication [19] or CONIKS [11]. Finally, work is needed to help users make good choices about which secure messaging applications are safe to use.

8. ACKNOWLEDGMENTS

The authors thank the anonymous reviewers and our shepherd, Apu Kapadia, for their helpful feedback. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1528022.

9. REFERENCES

[1] R. Abu-Salma, K. Krol, S. Parkin, V. Koh, K. Kwan, J. Mahboob, Z. Traboulsi, and M. A. Sasse. The security blanket of the chat world: An analytic evaluation and a user study of telegram. In *European Workshop on Usable Security (EuroUSEC)*. Internet Society, 2017.

[2] R. Abu-Salma, M. A. Sasse, J. Bonneau, A. Danilova, A. Naiakshina, and M. Smith. Obstacles to the adoption of secure communication tools. In *IEEE Symposium on Security and Privacy (SP 2017)*, pages 137–153. IEEE, 2017.

[3] A. Acquisti, I. Adjerid, R. Balebako, L. Brandimarte, L. F. Cranor, S. Komanduri, P. G. Leon, N. Sadeh, F. Schaub, M. Sleeper, et al. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)*, 50(3):44, 2017.

[4] H. Assal, S. Hurtado, A. Imran, and S. Chiasson. What's the deal with privacy apps?: A comprehensive exploration of user perception and usability. In *International Conference on Mobile and Ubiquitous Multimedia (MUM)*. ACM, 2015.

[5] S. Dechand, D. Schürmann, T. IBR, K. Busse, Y. Acar, S. Fahl, and M. Smith. An empirical study of textual key-fingerprint representations. In *USENIX Security Symposium*. USENIX Association, 2016.

[6] A. P. Felt, A. Ainslie, R. W. Reeder, S. Consolvo, S. Thyagaraja, A. Bettes, H. Harris, and J. Grimes. Improving SSL warnings: Comprehension and adherence. In *Conference on Human Factors in Computing Systems (CHI)*, pages 2893–2902. ACM, 2015.

[7] C. Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *New Security Paradigms Workshop*, pages 133–144. ACM, 2009.

[8] A. Herzberg and H. Leibowitz. Can Johnny finally encrypt? Evaluating E2E-encryption in popular IM applications. In *Workshop on Socio-Technical Aspects in Security and Trust (STAST)*, Los Angeles, California, USA, 2016.

[9] A. Kapadia. A case (study) for usability in secure email communication. *IEEE Security & Privacy*, 5(2), 2007.

[10] S. Krug. *Don't make me think!: a common sense approach to web usability*. Pearson Education India, 2000.

[11] M. S. Melara, A. Blankstein, J. Bonneau, E. W. Felten, and M. J. Freedman. CONIKS: Bringing key transparency to end users. In *USENIX Security Symposium*, pages 383–398. USENIX Association, 2015.

[12] T. Perrin and M. Marlinspike. The double ratchet algorithm. `https://signal.org/docs/specifications/doubleratchet/doubleratchet.pdf`, 2016.

[13] A. Sasse. Scaring and bullying people into security won't work. *IEEE Symposium on Security and Privacy (S&P)*, 13(3):80–83, 2015.

[14] S. Schröder, M. Huber, D. Wind, and C. Rottermanner. When SIGNAL hits the fan: On the usability and security of state-of-the-art secure mobile messaging. In *European Workshop on Usable Security (EuroUSEC)*, 2016.

[15] M. Shirvanian, N. Saxena, and J. J. George. On the pitfalls of end-to-end encrypted communications: A study of remote key-fingerprint verification. In *Annual Computer Security Applications Conference (ACSAC)*, pages 499–511. ACM, 2017.

[16] O. W. Systems. Signal. `https://signal.org/`. Accessed: 2018-02-10.

[17] J. Tan, L. Bauer, J. Bonneau, L. F. Cranor, J. Thomas, and B. Ur. Can unicorns help users compare crypto key fingerprints? In *Conference on Human Factors and Computing Systems (CHI)*, pages 3787–3798. ACM, 2017.

[18] N. Unger, S. Dechand, J. Bonneau, S. Fahl, H. Perl, I. Goldberg, and M. Smith. SoK: secure messaging. In *IEEE Symposium on Security and Privacy (S&P)*, pages 232–249. IEEE, 2015.

[19] E. Vaziripour, M. O'Neill, J. Wu, S. Heidbrink, K. Seamons, and D. Zappala. Social authentication for end-to-end encryption. In *Who Are You?! Adventures in Authentication Workshop (WAY)*. USENIX Association, 2016.

[20] E. Vaziripour, J. Wu, M. O'Neill, R. Clinton, J. Whitehead, S. Heidbrink, K. Seamons, and D. Zappala. Is that you, Alice? a usability study of the authentication ceremony of secure messaging applications. In *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

APPENDIX

A. STATISTICAL TESTS

This section contains the details of the statistical tests we ran.

A.1 Sample Size

We calculated the necessary sample size to compare two sample proportions (for comparing success rates) and two sample means (for comparing task times). With a 95% confidence interval, 80% power, and an expected success rate for the two samples (15% and 80%, based on our previous work [20]), the required sample size is 6. With a 95% confidence interval and 80% power, the hypothesized difference in timing completing the ceremony (4 minutes), and our previous measurements of variance for the task (9 minutes), the required sample size is 9. We rounded up to 10.

A.2 Success and Failure Rates

This data measures whether the participants were successful in using the authentication ceremony for the original Signal and each of the modifications. We want to test whether there are any differences in the success rate between the three versions of the Signal application.

Because the data is dichotomous we used Cochran's Q Test and found that the success rate was statistically different for the applications ($\chi^2(2) = 27.11$, $p<.0005$).

Since we used a between-subject study design, we performed Barnard's exact test to find the significant differences among the pairs of applications. This test shows the differences among all the pairs are significant (Signal vs. Modification 1, $p = 0.0165$; Signal vs. Modification 2, $p = 1.15E-05$; Modification 1 vs. Modification 2, $p = 0.0163$).

A.3 Task Completion Times

This data measures the time taken by participants to (a) find the authentication ceremony and (b) complete the authentication ceremony. We want to test whether there are any differences between the three versions of the Signal application, in finding and task completion time.

We did not perform a multiple samples comparison test because of the high failure rate with the original version of Signal. Since the studies are between subject, we ran a two-tailed two-sample t-test between Modification 1 and Modification 2.

For finding the authentication ceremony, a two-tailed, two-sample t-test with equal variance shows there is no significant difference between Modification 1 and Modification 2 (p=0.484, 95% CI: [-0.37, 0.759] minutes). This is expected since the interfaces for finding the ceremony are identical in these two versions. For completing the authentication ceremony, a two-tailed, two-sample t-test with equal variance shows there is a significant difference between Modification 1 and Modification 2 (p=7.849E-05, 95% CI: [1.937, 6.16] minutes). For the total time to find and complete the ceremony, a two-tailed, two-sample t-test with equal variance shows there is a significant difference between Modification 1 and Modification 2 (p=1.05E-05, 95% CI: [2.982, 6.518] minutes).

A.4 Trust Scores

A one-way ANOVA shows that there are no significant differences between the different versions (p=0.143).

B. STUDY MATERIALS

This section contains the study materials we used. The study coordinators used the interview guide to ensure that each pair of participants experienced an identical study. The study participants used the questionnaire to guide them through the study.

B.1 Interview Guide

Make sure to complete the following steps:

- When the two users arrive, read them the following:

  Welcome to our secure messaging application study. We are the study coordinators and are here to assist you as needed.

  Before we start the study, we need you to let us install an application called Signal on your phone. You will use this application during the study, and then we will delete it for you when we are done.

- Install the Signal application on their phone.
- Now read the following:

  In this study, the two of you will be in different rooms and will use the Signal app to communicate with each other.

  You will be asked to *think aloud* during the study. This means that you should explain everything you are thinking and feeling during the study so we can understand how you interact with the Signal application.

  During the course of this study we will be making an audio recording of what you say. We will transcribe these recordings and may publish them as part of our study, but we will not identify you in any way. We will destroy the audio recordings and will publish only transcripts so that you will be anonymous. We will not collect any personally identifying information about you.

  You will also take a survey during the study, and we will publish your answers, but without any information that can identify you.

  You will each receive $7 cash as compensation for your participation in this study. You will also have an opportunity during the study to earn a bonus of $3 cash, based on your performance. The expected time commitment is approximately 30 minutes.

  If you have any questions or concerns, feel free to ask us. You can end participation in this survey at any time and we will delete all data collected at your request. A study coordinator will be with you at all times to observe the study and also to answer any questions you may have.

- Before going to the study rooms, make sure the participants sign the audio recording consent form.
- Flip a coin and choose one participant to be Person A and one person to be Person B. Take the participant with whom you will work to the study room. Ask the participant to sit down.
- Start the audio recording using the equipment in the study room.
- Read the following instructions to your participant:

  We are going to ask you to do a series of tasks. Once you are done with each step, let the study coordinator know you have finished the task. You will then fill out a questionnaire and go to the next step. We need you to think out loud while you are doing the tasks in this study, meaning you are supposed to talk about how you are accomplishing the task and express any feelings you have. If you have any questions about the study ask the study coordinator. Remember you are allowed to talk to or meet your friend during the study.

  Please do not forget to think out loud.

- On the chromebook, load the survey from Qualtrics.
- Before using Signal, the survey will instruct the participant to tell you they are ready to begin the next task.

  During the course of the task pay attention to what user is doing and fill out one of the attached sheets.

  The user is supposed to think aloud while doing the tasks. If she forgets, gently remind her.

  Do not answer any questions from the participants.

  The participants have 10 minutes to complete the primary task, which is using Signal to exchange credit card information. If they do not finish the task on time, guide them to the next part of the survey. If you end

the task, inform the other study coordinator that you have done so, so that he catches up with you.

If it takes the pair too long to complete authentication or if they sent a credit card number before performing the authentication, then record that as a failure.

- When the survey is finished, ask the participant about their experience.

  Use the situations you noted while they took the study or interesting things they said on the survey. If they had any problems during the study, ask them to use their own words to describe the problem. Ask them how they would like to see it resolved.

- When the participant is finished, ask his/her opinion on the following questions:
  - Ask user if they trust the voice or text messaging for secure conversation?
  - If they did not use the authentication ceremony:
    - Ask them what they were looking for.
    - Show them how to find the application ceremony. Why did they not find it?
    - How do you think this screen would have helped you accomplish the task?
  - If they did use the authentication ceremony, show them the screen(s).
    - How do you think this screen helped you accomplish the task?
  - Explain the following:
    It is possible for someone to intercept your messages. These screens we have been showing you are called an authentication ceremony. Using the authentication ceremony ensures that nobody, not Signal, not hackers, and not even the government, is able to intercept your messages. You only need to do this once (or if your friend reinstalls the app). Now that you know this, are you willing to use the authentication ceremony before you exchange messages with a friend the first time?
- Stop the audio recording.
- Return to the study room. Thank the participants for their time. Ask them not to invite their friends to participate. Help them fill out the compensation forms and give them compensation.

B.2 Study Questionnaire

Signal study

1. Please enter whether you are Participant A or B.
   - A
   - B
2. What is your gender?
   - Male
   - Female
   - I prefer not to answer
3. What is your age?
   - 18-24
   - 25-34
   - 35-45
   - 46-64
   - 65 and over
   - I prefer not to answer
4. What is the highest degree or level of schooling you have completed?
   - None
   - Primary/grade school
   - Some high school, no diploma
   - High school graduate: diploma or equivalent (e.g., GED)
   - Some college, no diploma
   - Associate's or technical degree
   - Bachelor's degree
   - Graduate/professional degree
   - I prefer not to answer
5. What is your major, or if employed, your occupation?
6. Tell the study coordinator that you are ready for the next task to begin.
7. **For Person A**

   You left your credit card at home! You are going to be using the Signal app to ask your friend to send you the credit card number.

   This is the message you should send to your friend:

   "Hi! Can you send me my credit card number? I left my card on my desk at home."

   You can both earn a bonus of $3 for this study if you make sure that nobody can steal this information while your friend is sending it.

   Talk out loud as you do this task.

   For Person B

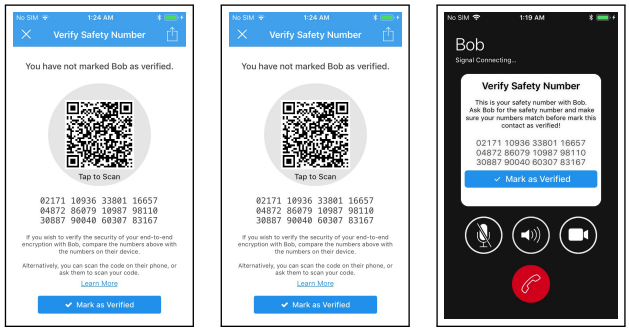
   Your friend is going to use the Signal app to ask you for their credit card number. Use the credit card given to you by the study coordinator.

   You can both earn a bonus of $3 for this study if you make sure that nobody can steal this information while you're sending it.

   Talk out loud as you do this task.
8. You will now be asked several questions concerning your experience with Signal.
9. Do you think you have safely exchanged the credit card number with your friend?
   - No
   - Yes
   - Not sure
10. Please explain your answer:
11. Did you see this screen during the study?
    (showed Figure 5)
    - No
    - Yes

    If (Yes), ask the following three questions.
12. What do you think this screen does?
13. Overall, how difficult or easy was it to use this screen to verify the safety number?

    (Extremely easy to extremely difficult)



(a) Original Signal (b) Modification 1 (c) Modification 2

Figure 5: Authentication ceremony screen

14. When you used this screen during the study to verify the safety number, what did you like or dislike about this? Please explain why.
(showed Figure 5)
15. Before this study, have you ever tried to send sensitive information when you use a secure messaging application like Signal?
    - Yes
    - No
16. (If Yes), Explain what kind of sensitive information you have sent.
17. I trust that Signal is secure.
    - Strongly agree
    - Somewhat agree
    - Neither agree nor disagree
    - Somewhat disagree
    - Strongly disagree
18. Please explain your answer to the above question.
19. How would you rate your knowledge of computer security?
    - Beginner
    - Intermediate
    - Advanced
20. Which of the following applications have you ever used? Select as many options that applies to you.
    - ☐ WhatsApp
    - ☐ Signal
    - ☐ Telegram
    - ☐ Line
    - ☐ Allo
    - ☐ Facebook Messenger
    - ☐ iMessage
    - ☐ Skype
    - ☐ Viber
    - ☐ Other

Informal Support Networks: an investigation into Home Data Security Practices

Norbert Nthala
Department of Computer Science
University of Oxford
norbert.nthala@cs.ox.ac.uk

Ivan Flechais
Department of Computer Science
University of Oxford
ivan.flechais@cs.ox.ac.uk

ABSTRACT

The widespread and rising adoption of information and communication technology in homes is happening at a time when data security breaches are commonplace. This has resulted in a wave of security awareness campaigns targeting the home computer user. Despite the prevalence of these campaigns, studies have shown poor adoption rates of security measures. This has resulted in proposals for securing data in the home built on interdisciplinary theories and models, but more empirical research needs to be done to understand the practical context, characteristics, and needs of home users in order to rigorously evaluate and inform solutions to home data security.

To address this, we employ a two-part study to explore issues that influence or affect security practices in the home. In the first part, we conduct a qualitative Grounded Theory analysis of 65 semi-structured interviews aimed at uncovering the key factors in home user security practices, and in the second part we conduct a quantitative survey of 1128 participants to validate and generalise our initial findings. We found evidence that security practices in the home are affected by survival/outcome bias; social relationships serve as informal support networks for security in the home; and that people look for continuity of care when they seek or accept security support.

1. INTRODUCTION

Securing home devices, services, and data is increasingly difficult and necessary. While home users are not as attractive a target as many organisations, they are both commonplace and vulnerable to several attacks. Initial work in exploring the security of home computer users [1, 22, 25] has highlighted the importance of this domain, and yet much more needs to be done to be able to address the scale and complexity of the security challenge.

According to the 2013 census, 74.4 percent of [U.S.] households use the Internet [42]. Similarly in 2015, 86 percent of households in Great Britain (22.5 million) had Internet access, up from 57 percent in 2006 [14]. Worldwide, Internet Live Stats reveals that over 46 percent of the world's population (3.4 billion) had Internet access in their homes by July 2016, up from 29 percent in 2010 [40]. And as the number of connected homes increases worldwide, so too do the threats.

In 2012, Rao and Pati [36] conducted a study in India revealing common threats and attacks facing home users: viruses, malware, identity theft and privacy violation, and phishing. Large organisations generally mitigate these types of threat well, however this is not the case for typical home computer users. Best practice in mitigating viruses in a home context seems to focus on running antivirus software, patching, and warnings to avoid untrusted or malicious websites (from web browsers and awareness campaigns). In contrast, in addition to antivirus software and patching solutions, larger organisations also have acceptable usage policies to manage risky behaviour from employees; segmented network architectures to avoid the spread of viruses; active firewalls, intrusion detection and prevention systems to identify problems before they cause significant damage; backup strategies to recover from incidents; and, perhaps most critically, an IT support function that can deal with problems should they arise. In comparison, home users have very few resources, capabilities, knowledge, skills, or tools to protect themselves from the multitude of threats that harm them directly.

But threats that directly harm the home are not the only concern. In today's highly interconnected world, the security of cyberspace depends on the security of all the different devices connected to the Internet. Ng and Rahim state that home users play a crucial role in securing cyberspace: if not well-protected, home systems can be compromised and used to attack critical infrastructure (such as telecommunication and banking) that heavily depends on the secure functioning of cyberspace [29]. While security breaches affecting organisations receive much attention, breaches involving home users usually come to light only when home devices or users themselves are involved in an attack affecting critical infrastructure. The October 2016 attack on Dyn, for instance, which is thought to have been enabled by insecure IoT devices in homes [5], triggered a number of reactions from different stakeholders, with some device manufacturers reportedly recalling their devices. Users at home face many different kinds of threats and mitigation requires interventions both within and outside the home.

A key strategy for improving home security practices so far has focussed on increasing awareness [33, 21, 24, 38, 30].

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

Despite the effort put in such approaches, studies [3, 33, 17, 27] and recent events [5, 12, 26] show that home users remain vulnerable as evidenced by insecure practices and choices to ignore security advice, leading the research community to explore alternatives to increasing awareness.

Dong et al. [10] propose an economics approach to designing security solutions for communities rather than individuals. They argue that incentivizing people to improve the security of a community (from which they benefit) through a shared venture would motivate personal security investment. While maintaining user-centeredness, Gutmann [20] proposes the application of problem structuring methods (PSMs), a technique from social planning, to help analyse security problems. The intent is to ensure the most appropriate solution is applied to a problem, and Gutmann claims to tackle a common problem where developers and service providers impose their favourite technology on people, without considering the environmental, social, political, and legal aspects of the overall problem. Wash and Rader [44] propose security story-sharing to help shape the mental models which inform home security decisions. Through sharing the right stories, and with expert involvement, the authors foresee changing home user security behaviour. Adding to the body of proposed approaches, Rowe et al. [39] put forward an approach modelled on public health systems for a shared secure cyberspace. They argue for a population-centred approach in dealing with cybersecurity issues. This is a departure from the typical practices in information security which take an individual focus in trying to understand how systems are compromised, and how they can be protected. The authors outline the technical requirements of a public cyber-health system, with specific focus on how the system would achieve monitoring, prevention, and incident response.

Building on this work, we believe that secure (and security) systems in the home need to be designed from an empirical and grounded understanding of home users, the context of use in which they operate, and how they make data security decisions. We report on the qualitative and quantitative research we have undertaken to explore the security practices of home computer users. We conducted 15 scoping semi-structured interviews, followed by a further 50 targeted semi-structured interviews lasting approximately 60 minutes each. We analysed the data systematically using Grounded Theory and used this to design and run a quantitative survey of 1128 home users to explore how widely shared the qualitative findings are. Our key findings are:

- *Social relationships* play a vital role in information security in the home. They serve as informal support networks of security practices.
- *Perceived competence* is an important factor in security decision-making in the home. It is used to assess the quality of a security source, and the support offered in the home. The participants use different metrics to evaluate competence, including the profession of the source, the educational standing of the source, the level of usage of technical devices of the source, and negative experiences of the source.
- *Continuity of care* is an important characteristic of security support in the home. Participants report seeking or accepting support from a source that is constantly available when needed.
- Participants look for *evidence of a security problem or need* for them to practice security. Typical evidence is direct harm to an individual, or their social relation, resulting from the individual's insecure behaviour.
- *Confidence* of the participants in an implemented *security control* can increase insecure practices.

The remainder of this paper will review the related work in this domain in section 2, describe our research methodology in section 3, and present our results in section 4. We finally discuss the implications of our findings and highlight areas of interest for future work in sections 5 and 6.

2. RELATED WORK

In this section, we review prior work investigating home user security practices, structuring the concept of security practices into: (i) security behaviours; and (ii) the factors that influence the security decisions that precede the behaviours.

2.1 Security Behaviours

Studies have been conducted to understand and improve security behaviours in the home. AOL and the National Cyber Security Alliance conducted a study of online safety of home computer users [2] where 329 home users were interviewed and their computers were analysed. Researchers asked and checked for the availability of virus and spyware protection software, firewalls, parental controls, and the use of encryption for wireless network users. The study concluded that the majority of those studied lacked core protection. Similarly, Furnell et. al assessed the security perceptions of UK home users [17]. They surveyed 415 home users about their awareness of security threats, usage of system safeguards (firewall, antivirus, anti-spyware, and anti-spam software), and their awareness and understanding of security-specific tools found in contexts such as operating systems and applications. The study found that both novice and advanced home users appeared vulnerable to security risks. The authors concluded with a call for the development of new models of engagement and awareness raising.

Rao and Pati surveyed home users in India to understand their levels of awareness of security threats and usage of security measures (password protection, antivirus, firewall, patching, data backup, and parental controls) [36]. The study revealed poor understanding of security threats, and low levels of adoption of recommended security controls. The authors concluded that the security in the home can be improved through awareness and user-friendly security controls. Similarly, Ng and Rahim studied factors that influence a home computer user's intention to practice computer security [29]. They surveyed 233 home computer users on the use of antivirus software, data backup, and personal firewall.

Ion et al. studied security practices that different experts and non-experts consider to be the most important in protecting their security online [23]. They conducted 40 semi-structured interviews with security experts, and used the results to design a survey. 231 security experts and 294 non-security experts were surveyed, and the practices of the two groups compared. The studied practices included installing software updates, using antivirus software, account security (using password managers, writing down passwords, changing passwords frequently, and using two-factor authentication), and mindfulness (visit only known websites, check

if HTTPS, clear browser cookies, and email habits). The results showed discrepancies between the most important security practices of the two groups. The authors concluded that more work is needed to improve the practices of non-experts, and identified three key recommendations: install software updates, use password managers, and use two-factor authentication for online accounts.

Dourish et al. [11] investigated how users respond to security issues in their daily lives and found that people ask for assistance or delegate security activities to knowledgeable family members (similar to [15]), friends, or roommates. They also found a reliance on technology (e.g. SSL for data connections, ssh tunneling for email, or trust wired Ethernet to be more secure than a traditional wireless medium) for protection; others reported delegating security to institutions such as financial companies. Likewise, Nthala and Flechais [31] found that some home users turn to trusted others (colleagues, IT professionals, relations, and peers) for help with security issues.

2.2 What Influences Security Behaviours?

Research has been conducted to investigate and understand the factors that motivate different security behaviours. Several studies [38, 29, 31] have shown that social influence has an impact on the security behaviours of home users. Das et al. [8, 9] studied in more detail how this social influence plays a role in the security behaviours of home users. They found that social influence affected the security behaviours of those involved through social processes (observing and learning from friends, social sense-making, pranks and demonstrations, negative experience of others, and device sharing), and conversations about security (a finding similar to Rader et al. [35]).

Wash [43] carried out a qualitative study of iterative interviews to investigate the existence of folk models of security for home computer users, aiming to increase our understanding of mental models of security for home computer users. The study focussed on finding out how home computer users understand and think about potential threats. Wash identified eight folk models categorised into models of viruses and other malware, and models of hackers and break-ins.

Herley [21] argued that users perform an implicit cost-benefit analysis when making a security decision. The cost is the effort required to follow security advice, while the benefit is the avoidance of potential harm that a successful attack might cause. The harm includes monetary loss (if any) that victims endure, but also the time and effort they must spend resolving the situation. Similarly, [31] found that the cost of protection also influences the outcome of security decisions in the home.

In a study investigating why users accept or reject different advice about secure behaviours, Redmiles et al. [38] found that users reject advice due to too much marketing information, inconvenience and threatening users' privacy. In addition, the study reported that trust was a clear factor that influenced the choice of a source of security advice.

Other related work has focussed on understanding practices around home network security, highlighting the differences in responsibility between Internet Service Providers and home users [32].

3. METHODOLOGY

We started our study with a scoping study of 15 semi-structured interviews. The aim of the scoping study was to make an initial exploration of security practices (which we consider to consist of (i) security behaviours and (ii) the decisions that lead to such behaviours) in the home, from which we would identify a research gap for further exploration. Our research questions would then be refined based on the initial results. Respondents for this study were chosen from a snowball sample [7] of home users in the UK. Two research questions guided our interviews during the scoping study:

1. What influences security decision-making in the home?
2. What kinds of security behaviours exist in the home?

We analysed the data using Grounded Theory (see section 3.2.2) to identify all the key themes emerging from the data. Our analysis identified a number of factors that influence the outcome of security decisions in the home, all of which were consistent with previous studies discussed in section 2.2. These included inconvenience, trust, cost, and availability of too much marketing material. Analysis of the data on security behaviours revealed two separate categories of the behaviours which we categorised as: *security work* and *security support*.

Security work is highly contextual and specific to technology platforms, comprising behaviours such as installing and using firewalls, antivirus software, patching, data backup, and parental controls. As reviewed in section 2.1, our findings were consistent but much less comprehensive than previous surveys in this area.

Security support, on the other hand, comprises two subcategories; support seeking and support giving. The work of Dourish et al. on delegation [11], Nthala and Flechais [31] on security support, and Redmiles at al. [38] on advice seeking and giving, all fall under security support. We noted that little work has been done to explore security support that is required or available in the home in great detail.

This led us to focus our research on understanding security support in the home, and the reasoning behind it. We thus refined our main research questions to:

1. What influences security decision-making in the home?
2. What are the characteristics of security support in the home?
3. Where do home users get support?

To answer these questions, we adapted the research methodology proposed by the Productive Security research team of Beautement et. al. [6]. We conducted a two-part study aiming to increase our understanding of security support in the home, and the reasoning that surrounds it. In the first part, a detailed understanding of the problem domain arises out of studying a few individuals and exploring their perspectives in great depth. In the second, a more generalisable understanding of the issues identified in the first part can be gained from examining a large sample and assessing responses to a few variables.

Part 1 of our research consisted of 50 targeted semi-structured interviews with a broad range of individuals and families within the home context. As the interview data was being collected, it was qualitatively analysed using Grounded

Theory (see section 3.2.2) in order to identify the significant themes to answer our research questions. The themes were used to generate scenarios and questions from which a survey was developed and run in the second part of our study. By tailoring our survey to the home context, we ensured that the questions were relevant and recognisable to the participants.

Part 2 made use of *Unipark* to run an online survey and *Prolific Academic* to identify a representative sample (in terms of age, gender, and educational level) of 1128 participants. The survey results were analysed and aimed to validate the findings of the qualitative data analysis, and support the generalisability of these results to a wider home user population. This was meant to provide clear evidence on which future work can draw to improve education, technology, and practices for home data security.

The study was ethically reviewed and approved by the Social Sciences and Humanities Inter-divisional Research Ethics Committee at our institution.

3.1 Recruitment

We recruited for the interviews by advertising through community centres, newspapers (in print and online), and other social groupings, and by putting up posters at the National Museum of Computing. The recruitment was conducted in different locations in the UK. Before starting an interview, we collected demographic information including age, gender, highest educational level, ethnicity, marital status, and occupation from the respondents to ensure we cover a broad range of home users. Each participant was compensated with a £10 Amazon voucher for an approximately one-hour interview session.

Participants for the survey were recruited through Prolific Academic, and each participant was compensated with £1.70 for an approximately twenty-minute session.

3.2 Procedure

3.2.1 Semi-structured Interviews

We followed a semi-structured interview protocol utilising an interview guide to maintain direction while keeping the interview open for both depth and breadth topic exploration. Prior to the interview, participants were asked to complete a demographic form, which included questions regarding the devices and services they use. Our interview guide is appended in D.

3.2.2 Grounded Theory

The interview data was analysed using Grounded Theory [19]. Grounded theory allows researchers to examine topics and related behaviours from many different angles, leading to comprehensive explanations. It is used to uncover beliefs and meaning that underlie action, and to examine both rational and non-rational aspects of a behaviour [41]. This makes it the ideal choice for studying security support and any issues that surround it. Our approach was consistent with that described by Strauss and Corbin [41].

Three researchers were involved in the analysis. The primary researcher, who conducted the interviews, did the initial open coding of the interview transcripts. To ensure credibility of the codes, a second researcher cross-checked all the codes against the interview transcripts. At the same time, the third researcher reviewed the initial codes and all quotes supporting each code. Any differences and/or issues arising from the initial coding were discussed and resolved among the three researchers. A codebook consisting of 130 codes emerged from the initial coding. These codes were then applied across other interviews through constant comparison, while new codes were added as they emerged and were deemed necessary. In further analysis, the three researchers discussed and grouped the codes into themes (axial coding) and categories (selective coding), based on the properties and dimensions of each theme. Regular coding meetings were held to discuss any emerging codes and to group the codes into families.

3.2.3 Survey Development

The survey tool was developed from the Grounded Theory analysis of the interview data to test a number of significant themes. Scenarios used in the survey were developed from analysis of anecdotes from the interviews, and themes that emerged from the analysis. The aim was to ensure that the participants were presented with scenarios they are familiar with, hence reducing the effect of unknown personal preferences. We made sure that our options to the scenarios were testing the construct under study. Hence, options with factor loadings less than .30 were dropped.

Prior to running the full survey, the tool was piloted and tested with seven participants. To ensure we tested for both clarity and usability of the tool (face validity), we developed and tested it on the platform it would run on (Unipark). The questionnaire went through three iterations of testing, and modification with our participants (four non-experts and three experts – two in usable security research and one in human-centred computing studies).

Two non-experts tested the instrument online, followed by *cognitive interviews* [46]. The participants were asked how they understood and interpreted each question; how easy they found it to understand each question and respond; how easy it was to navigate through the whole questionnaire; and how they viewed the general outline of the questionnaire. This was followed by *expert interviews* as applied in [37], where each expert was asked to first test the survey online, and then review each item on the survey tool in terms of biases, question ordering, clarity, sensitivity of questions, and other issues; all in line with the aim of the study. After this phase, the last two non-experts tested the tool, followed by cognitive interviews.

During each of these phases, the tool was updated based on feedback from the interviews. Once a consensus was reached on all issues affecting different aspects of the tool, we published the study on Prolific Academic targeting 1128 UK only respondents. We asked participants about demographic information including age, gender, and educational level. Survey questions revolved around factors that influence security decision-making (survival/outcome bias, confidence in a security measure, and availability and quality of support), characteristics support (duty of care and continuity of care), and preference and sources of support.

To check the quality of responses, we applied three kinds of checks. First, we used Prolific's start and finish times to check for *speeders*. During testing of the questionnaire, the

average completion time was fifteen minutes. After publishing the survey on Prolific, we applied demographic filters of the survey platform on the first set of fifty responses to get a representative sample of the demographics shown in Figure 2. The average completion time remained fifteen minutes, with a minimum of twelve minutes. We set our minimum acceptable response time at ten minutes. Responses below the limit were rejected. Second, we checked for and rejected *straight-liners* - responses that all have the same answers, and *pattern responses* - answers in a pattern. Third, we included a *binary red herring question* which read, "I am randomly answering the questions" with a "Yes" or "No" answer. We placed one towards the middle of the questionnaire, and another towards the end. Responses bearing a "Yes" to any of these questions were rejected.

Due to the ordinal nature of our data, we tested for reliability of different constructs - each measured by a scale of items - on the final questionnaire by computing their ordinal alpha coefficients (Ordinal α) [18]. The constructs had the following coefficients: survival/outcome bias, .75; confidence in a security measure, .74; duty of care - motivate others, .91; duty of care - be motivated by others, .83; and duty of care - social responsibility, .81. Since our test for continuity of care involved repeated measures, we tested for the reliability of the eight pairs of items using Spearman rank correlation coefficient (r_s). There were positive correlations between each of the eight pairs of items, all significant at $p < 0.05$. The Spearman coefficients for the pairs were, $r_s(1085) = .594, .672, .601, .583, 638, .564, .499, .530$ for pairs A through H discussed in section 4.3.2 respectively.

3.2.4 Survey Analysis

For the survey data, we present descriptive statistics for the different variables. We also run inferential tests on the data including Friedman [16] and Wilcoxon Signed-rank [45] tests for analysis of matched-pair data and rank-ordered data. These non-parametric tests were selected on the basis of the ordinal nature of our data, where the chances of getting valid results from parametric tests were minimal or unclear.

3.3 Limitations

Our study has some limitations. First, all are participants are residents of the UK. This might raise questions regarding generalisability of our results. However, we have documented the procedure we followed in this study, which makes it possible for other researchers to replicate it elsewhere.

Second, common to all qualitative studies, researcher bias is a concern. A single researcher, trained to conduct research interviews, conducted all the 65 interviews. The researcher avoided leading questions, and ensured participants felt comfortable to respond to questions. The researcher avoided interrupting participants, and probed for more information when required. To further mitigate bias, two other researchers reviewed and were part of the data analysis to enhance consistency in data coding. Our research design explicitly aims to mitigate potential bias by also running an extensive survey to test how generalisable the qualitative findings are.

Third, given that security is a sensitive topic, social desirability could bias some of the responses to the survey, specifically for the two scenarios developed to study survival/outcome bias and confidence in a security measure. To mitigate this, we took three measures: 1) we did not reveal at the onset that the main purpose of the survey was to study security practices of the participants. Instead, we stated that the aim was to understand decision-making in the daily use of technology. 2) We employed a self-administered questionnaire [28], hence no interviewer presence and a high degree of anonymity. 3) We used indirect (structured, projective) questioning [13] in those two scenarios, where respondents answered from the perspective of another person.

Demographic	Category	# Participants
Age	12-17	2
	18-34	22
	35-64	24
	65+	2
Gender	Male	26
	Female	24
Highest educational level	No schooling completed	1
	High School	11
	Trade/technical/vocational training	2
	Undergraduate	8
	Graduate	12
	Postgraduate	16
Ethnicity	White	39
	Hispanic/Latino	1
	Black/African/Caribbean	5
	Asian/Pacific Islander	5
Marital Status	Single	28
	Married	18
	Divorced	3
	Separated	1
Employment status	Employed	28
	Retired	3
	Self-employed	8
	Not working	2
	Student	12

Figure 1: Interview participant demographics

Lastly, our data consists of only what people say. This makes it hard to understand how our results translate into actual behaviour in the home. Future work would aim to employ relevant approaches to study these behaviours in context.

4. RESULTS

In this section, we detail the findings of our study. We start by presenting the demographics of our participants, and then discuss the key findings from our study organised according to the research questions. First, we discuss the factors that influence the outcome of security decisions in the home. Second, we explain the different factors that our participants reported using to evaluate the quality and source of security support. Finally, we detail the characteristics and sources of security support in the home.

4.1 Participants

Our scoping study comprised 9 male and 6 female participants, with ages ranging from 18 to 34, and an ethnicity of 4 Asians, 5 Whites, 4 Africans, and 2 Black Americans. For the targeted semi-structured interviews, we selected 60 people to interview, 50 of which attended. We kept a balance between male and female participants, as well as a diversity of age, ethnicity, education, and employment status.

Demographics for our 50 participants are shown in Figure 1. Two participants indicated being both students and em-

		Gender	
		Male	Female
Age	18 - 34	253	254
	35 - 64	275	269
	65+	12	24

Education	Age			Gender	
	18 - 34	35 - 64	65+	Male	Female
No schooling completed					
High school	131	134	13	137	141
Trade/technical/vocational training	54	88	9	88	63
Undergraduate	148	87	3	126	112
Graduate	102	149	4	114	141
Postgraduate	69	83	6	72	86

Figure 2: Survey participant demographics

ployed, while one indicated being both employed and self-employed. 52% of our participants were male, 48% were female. 44% belonged to the 18-34 age group, 48% belonged to the 35-64 age bracket. During the interviews, these two age groups were noted to be the ones responsible for making most of the security decisions in the home environment. The other two age groups, 12-17 and 65+, made up 4% of the participants each. 32% of the participants hold postgraduate degrees, 24% have graduate degrees, 16% completed undergraduate studies, 4% completed trade/technical/vocational training, 22% completed high school, and 2% did not complete any school level.

1128 respondents took part in the survey. After running quality checks on the data, 41 responses were excluded, leaving 1087 responses. Fifty percent of our participants were male, and fifty percent female. Forty seven percent were between the age range of 18 - 34, fifty percent between 35 and 64, while three percent were above 65 years old. Of all the participants, less than one percent had not completed any education, twenty six percent had completed high school, fourteen percent had done trade/ technical/ vocational training, twenty two percent had undergraduate degrees, twenty four percent had graduate degrees, and fifteen percent had postgraduate degrees. The demographics of our participants are summarised in Figure 2.

4.2 Security Decision-Making

We asked our interview participants questions regarding their security decision-making process in order to identify factors that influence the outcome of such decisions. In addition to other factors (knowledge and skill, inconvenience, cost, trust, and influence) that have been reported by other studies before (ref. Section 2), we identified three other areas that have not been explored yet. These include survival/outcome bias, other factors that induce or undermine one's confidence in a security measure, and the availability and quality of support. We discuss these in detail below.

4.2.1 Survival/Outcome Bias

Our analysis of the interviews reveals a tendency for participants to concentrate on practices that have survived security breaches, and to overlook those that have not. This was a reason some participants gave for not implementing recommended security measures. They believe that as long as something bad has not happened yet, they are safe: *"For me, until something happens, I will be safe"* - P4.

Even in the face of a security concern, some participants report not engaging in security action because *"I think it's probably the fact that as far as I'm aware of, I haven't had serious breaches of personal data, or data security breaches. Not that I'm aware of, no. I think if I was exposed to something which was quite serious, then I would probably change my look quite a lot"* - P6 or *"I don't think I have because I have not had any reason to. That's why personally I just feel like as long as it has not done anything that would cause direct harm to like my information or anything like that, [it is secure]. I haven't felt the need to do any other security check to keep up with any security information because I haven't experienced anything that would cause me to do that. So I feel like until I have that experience with maybe an application, then I might either delete the application, or look for some security measures that I might take"* - P1.

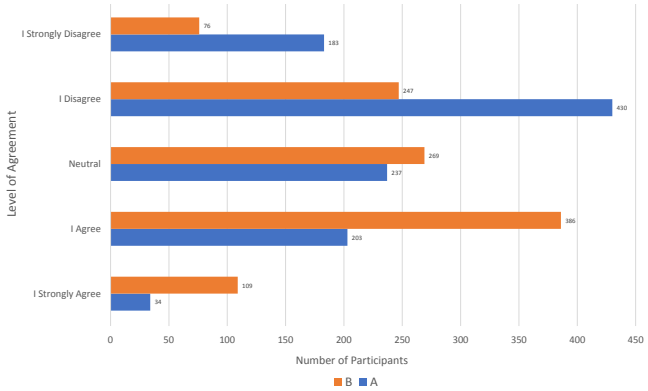


Figure 3: Survival/Outcome Bias

While realising that statistical validation of this factor requires some complex and detailed study design as shown in [4], we crafted a scenario to make a preliminary exploration of the availability of this factor. We presented the respondents with two options, both indicating survival/outcome bias. Shown below is the scenario:

> For the past 5 years, your friend John has been downloading free music, videos, and software from different websites including torrent sites without any problem. One day, he reads an article about the dangers of free downloads such as viruses, adware, Trojan horses, worms and spyware. For each of the following options, how much do you agree that it is a good choice for John?
> A - *Continue downloading free files from any website as usual. He has been doing it for 5 years without a problem, chances of being affected are very small.*
> B - *Restrict the downloads to those websites John has already used before. He has used them for 5 years without a problem, he trusts them to be secure.*

The options were evaluated on a 5-point Likert scale ranging from Strongly Disagree to Strongly Agree. The results showed that about 22% agreed with option A, while about 46% of the participants agreed with option B (cf. figure 3). While there was a statistically significant difference between options A and B ($Z = -18.058, p = 0.000$), our aim was to make an *initial exploration of the availability of survival/outcome bias*, and not to study types or levels of survival/outcome bias, or factors that affect the construct.

4.2.2 Other Factors That Induce or Undermine Confidence in a Security Measure

In our analysis of the interviews, we found that where a security measure was in place and the participants were confident in it's effectiveness, they would trust the service or action to be secure; *"With financial, there was one time when my credit card was charged to two transactions that I did not recognise. I immediately contacted the bank, and I was able to describe why I couldn't recognise them, and the bank believed me and refunded my money... That made me confident in using online shopping, and financial services"* - P7 ... and similarly *"I am less concerned about banking because I find that the banking services I use to be secure, and I am often reassured by the fact that if something were to go wrong, the bank is likely to compensate me for any fraud or any security breaches that would result in the loss of my money"* - P21. This confidence is not always to do with security measures implemented by a service provider however; *"If they have got work stuff on their laptop, or they are one of those people that have a word document with all their passwords on it, people do that, then I would probably advise them to think about high level security, or at least password-protecting files because I think it's very interesting that there has been an increase in people holding data hostage, and say pay us this, and you can have your files back. That for me would be like, ok you can keep it. I am not that bothered. Any photos I have got are uploaded to the cloud, there is nothing on my desktop that I need that can't be replaced. But for a lot of people, that obviously is not the case."* - P5.

To explore this factor, we crafted the following scenario:

> Your friend Felicity is a college student. She owns a laptop. She stores assignments and study materials on it. Felicity visits her friend, Laurel, whom she finds watching a very interesting movie. Felicity asks Laurel if she can share the movie with her, as well as some of the music Laurel downloaded. Laurel copies all the files to a USB stick, and hands it over to Felicity. On their way out, Laurel tells Felicity that she thinks her laptop might have a virus because she could not open one of her word documents to study, and this has happened to her a number of times. For each of the following options, how much do you agree that it is a good choice for Felicity?
> A - *Felicity could copy the movie and music to her laptop. Laurel probably got a corrupted file, there is nothing to fear.*
> B - *Felicity could copy the files to her laptop. She has an antivirus which will keep her data secure.*
> C - *Felicity could take and maintain a backup of her files in a USB stick, phone storage, cloud storage, external drive, another computer, etc. She could hence copy the movie and music to her laptop. She can always get the files from the backup when needed.*

We introduced option *A* to indicate taking no action, here serving the purpose of a control variable. The other two options, *B* and *C*, were used to test the participants' confidence in the implemented security measures and the subsequent behaviour following from their confidence. These options were also evaluated on a 5-point Likert scale ranging from Strongly Disagree to Strongly Agree. The results (cf. figure 4) showed that about 14% agreed with option A, about 26% agreed with option B, and about 46% agreed with option C.

A Wilcoxon signed-rank test showed that the introduction of an antivirus in *B* resulted in a significant statistical difference between option A and B ($Z = -16.473, p = 0.000$). Similarly, the Wilcoxon signed-rank test showed a significant difference between options A and C ($Z = -21.855, p = 0.000$), where a backup was introduced as a security measure. While there was also a significant statistical difference ($Z = -14.497, p = 0.000$) between options B and C, it was not our aim to compare different security solutions, and we hypothesize that this might have occurred due to the participants' perceptions, preferences, needs and experiences.

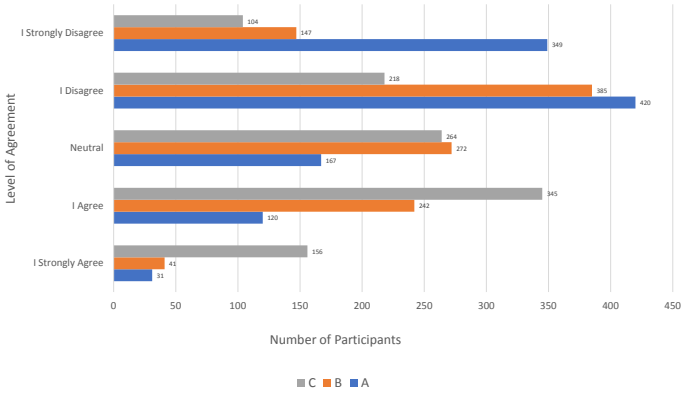


Figure 4: Confidence in a Security Measure

4.2.3 Availability and Quality of Security Support

Our analysis of the interviews surrounding security decision-making in the home revealed that our participants constantly need support in their endeavour to be secure. Previous studies have explored support in terms of security advice or information [38, 37, 21, 34]. While this is a common trend, there is evidence [3, 33, 17, 27] of a low success rate of such form of support. We thus set out to first identify the kind of support that is needed or exists in the home regarding security. Our analysis revealed a number of different kinds of support currently present and/or needed in the home: *information, advice, and technical help.*

While there might be some differences between information and advice, we noted that participants treated the two as the same. This challenge is also seen in other studies [38, 34] that have been done on this topic, where they interchangeably refer to the two without any difference. To avoid introducing discrepancies in the results, we therefore treated these two as one, and referred to it broadly as security advice. Our analysis pointed out the following kinds of advice that our participants talked about:

- Advice on available security tools or controls
- Reviews about a particular security tool or control
- Information on the cost of protection
- Opinion or recommendation for a particular security-related action, e.g. permissions requested by applications
- Advice on privacy settings
- The risks for a specific environment, service, or tool
- Where they can get support with a particular problem

Technical support was reported to be common mostly among the social circles of the participants. This included some aspect of responsibility where someone, who is perceived to be more competent or feels responsible, assumed the responsi-

bility of making security decisions on behalf of others (that is, decide and act on their behalf). Parents for example reported making decisions for or offer advice to their children; *"I give that [advice] as a concerned parent just as I would encourage them to look both ways when they cross the road. They don't ask me for that advice."* - P4, *"I don't think anyone is really responsible for the household. Myself and my wife will have some say in what the children can or can't do on their devices. But no one person is responsible for that."* - P30; friends on behalf of their friends, *"One of my friends is good with computers. He does all the security stuff for me when he comes."* - P48.

We were particularly interested in how participants choose where to seek this support and/or whether or not to accept any unsolicited support that is offered to them. In this regard, we identified five factors that are used to assess a source and/or the quality of support: perceived competence, trust, availability, cost, and closeness to a source.

i. Perceived Competence: The notion of *better than me* was common among the participants when talking about a source of security support. We understood this to mean the perceived competence of the source of support; and 91% of the survey participants agreed to consider competence in seeking or offering support. The participants reported making a comparison between their self efficacy and the perceived competence of a potential source.

We sought to identify the metrics that are used in this comparison, or in other words, how the different participants understand competence in security. Our interview results showed that for some it means someone who *works in data security*; 86% of the survey participants nodded to this. For others, it means someone who *works for a technical company*, regardless of whether their job is technical or not; 24% of the survey participants agreed to consider this metric. More than that, it also means someone *whose job is technical*; 24% of the survey participants agreed with this.

Another metric used in assessing someone's competence involves identifying someone with *more experience in using technical devices than the one seeking help*; 51% of our survey participants agreed with this. 27% consider someone who *has studied/studies* a technical course. 7% go for someone who is *more educated than the one seeking help*. 78% said they choose someone who *has studied/studies data security*. 39% seek help from those who have *experienced a data security incident before*; and only 4% said they do not consider any of these factors when choosing a source of support. The survey participants were asked to select more than one metric they consider, hence the percentages total more than 100.

In addition to selecting the metrics the participants consider in assessing the competence of a potential source of support, we also asked the participants to rank these metrics in order of preference. A Friedman Test on the metric rankings showed that there was a statistically significant difference ($X^2(7) = 3218.784, p<.05$). Post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level of $p = 0.002$. There were no significant differences between options A and D ($Z = -.339, p = 0.735$), or between A and H ($Z = -1.320, p = 0.187$), or between B and H ($Z = -1.744, p = 0.081$), or between D and H ($Z = -1.646, p = 0.100$); however, B was ranked higher than A ($Z = -4.662, p = 0.000$), and higher than D ($Z = -3.909, p = 0.000$). The overall ranking is:

1. F: He/she works in data security.
2. G: He/she studied or studies data security.
3. C: He/she has more experience than you in using or working with technical devices and services.
4. B: His/her job is technical.
5. A,D,H (A: He/she works for a technical company; D: He/she studied or studies a technical course; H: He/she has experienced a data security incident before.)
6. E: He/she is more educated than you.

ii. Trust: Previous studies [38, 37, 31] reported that trust plays a role when users choose a source of security advice. Similarly, our study found that trust influences the choice of a source of support among our participants. Characterising this in our study was the availability of a social relationship between those involved. This is also reflected in the preferences of a source of support, discussed in 4.3.1. When seeking advice for instance, *"because they are my closest friends and I kind of trust what they have to say. I know that they give me an honest opinion"* - P29; and *"they are my parents. So I am their closest relation. I think they trust me a lot"* - P2. 89% of the survey participants indicated considering trust when they seek or accept security advice or help.

iii. Cost: Our study confirmed what other researchers [21, 31] have reported about the importance of cost in security. We went further to identify two dimensions of cost among our participants that are considered in deciding when, and where to seek support. First, *cost to the one seeking help*, which includes money, favours, and gifts. Second, there is *cost to the source of support*, which is characterised by effort, and inconvenience. These dimensions were evident in reported (from interviews) security support sought and offered among the social relationships of the participants. In the survey, we asked the participants to choose which of the two they took into consideration when choosing a source of support. 49% indicated that they consider the cost to the one seeking support as an important factor, and 36% consider the cost to the source of support to be a significant factor.

iv. Closeness: When we tried to find out about the sources of security support in the home in our interviews, one thing that was not clear was whether the preference of the sources was determined by (constant) availability of the source, or how close one is to the source. Phrases such as "my friends", "my dad", and "my work colleague" could not explicitly clarify which of the two was in play. When asked why they chose such sources, the common responses were "because they are better than me", "they know me", or "I trust them". We hence separated the two, *closeness* and *availability*, and surveyed them as separate factors. 31% of the survey participants indicated that they consider closeness as a significant factor in selecting a source of and accepting support for their security.

v. Availability: Our analysis of the interviews indicates a common pattern in the sources of security support, be it advice or technical help. Such consistencies included friend-to-friend, parent-to-child, between couples or within a fam-

ily, among work colleagues, and client-to-commercial IT Services Professional. In the survey, we asked the participants if constant availability of a potential source of support is an important factor. 31% of the participants indicated that they consider availability as a significant factor.

Only 1% of the survey participants indicated that they do not consider any of these factors when selecting a source of security support. We also asked the participants to rank these factors in order of preference. A Friedman Test on the ranked factors showed that there was a statistically significant difference ($X^2(5) = 2444.265, p<.05$). Post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level of $p = 0.003$. There was no significant difference between *availability* and *cost to you (money, favour, gifts, etc)* ($Z = -.835, p = 0.404$). The overall ranking therefore is as shown below:

1. Competence
2. Trust
3. Availability and Cost to you (money, favours, gifts)
4. Closeness
5. Cost to the source of advice/help (effort, inconvenience)

In the next section, we discuss what characterises security support in the home. We detail the how the evaluation of the five factors discussed in this section impact the sources of support, and the reasoning behind the choices and practices.

4.3 Characteristics of Security Support

Our analysis of the interviews reveals that participants mostly had the same sources for advice and technical help. These included family, friends, work colleagues, service providers, and IT repair shop professionals; with family and friends being the most common source. This corroborates other studies [38, 37, 11, 17]. Other sources include search engines (*"I searched online for people with the same problem and got many results. People gave many solutions and I tried several of them until I got one that seemed to work."* - P23), and specific websites (*"Sometimes you go to sites that you think are credible like stackoverflow... some credible sites or sites that look credible to me. I just read about what people have experienced and how they went about it."* - P11).

None of the sixty five interviewees cited any security awareness websites as a source of security advice. We did not expect our participants to recall details of websites they visit for security information, but this is consistent with the findings of Furnell et al. [17], who found that the majority of their respondents had not heard of public awareness websites (including Get Safe Online: https://www.getsafeonline.org/, and Webwise: http://www.bbc.co.uk/webwise).

Our analysis shows that the preference and choice of a source or recipient of security support in the home is characterised by two main attributes: duty of care and continuity of care.

4.3.1 Duty of Care

Participants consider security support in the home a moral obligation to ensure the safety or well-being of others. This duty of care is expressed through the following modalities.

i. Delegation: As explained in section 4.2.3, support for security in the home involves seeking or accepting advice, but also encompasses users taking security responsibility for others to ensure their well-being. We found that some people delegate the responsibility for security to competent, and trusted others; a result shared by Dourish et al. [11], who found that people *"delegate to another individual, such as a knowledgeable colleague, family member, or roommate"*. Some of our participants said; *"Me! Mum always. I guess because my husband thinks I'm more knowledgeable about computers and about settings for the internet"* - P7; and *"Oh! My husband, because he has always been keen on computers and adopting technology, and that is a big part of his work. So he is the one who does that [all security tasks]"* - P45. A similar finding is also presented in [31], *"There is a friend who usually comes here. Mostly he is the one. If the laptop has a virus, I give it to him."*

ii. Motivation: A second way in which duty of care is expressed is by motivating others to behave securely. This generally includes offering unsolicited support. Our interview data shows two aspects of unsolicited support: 1) when somebody notices a practice they believe to be insecure and they intervene (e.g. *"they just feel like they can send a young person like 'go and check my email', and they give you all the details to check the emails and I'm like, it's supposed to be private."* - P1); and 2) when there is nothing specifically wrong but support is offered (e.g. *"My parents, I do advise a lot about different security issues. They are just aware of it"* - P43). Unsolicited support without noticing a particular need was common in cases where there was delegation and participants felt responsible for the security of another.

We asked survey participants how likely they are to offer unsolicited advice and technical help to someone they believe to be less competent in security than them. Since the interviews show that this practice is common among relatives, friends, and colleagues, we sought to explore in our survey how widely held such behaviour is. Our survey shows that about 56% of the respondents are likely to offer unsolicited support to a relative; about 47% to a friend; about 27% to a work colleague; and about 12% to other sources.

We also asked the participants to rank who they would likely offer unsolicited support to, in order of preference. A Friedman Test on the ranked order of preference showed that there was a statistically significant difference ($X^2(3) = 2127.517, p<.05$). Post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level of $p = 0.008$. The overall ranking in order of preference is as shown below:

1. Relative
2. Friends
3. Work colleague
4. Others

But offering unsolicited support is only one side of the coin – to fully explore this, we also asked participants how likely they are to accept unsolicited advice or help with data security from different sources of support. About 63% of respondents reported being likely to accept it from a relative; 63% from a friend; 48% from a work colleague; 44% from a service provider/manufacturer help desk; 40% from an IT repair shop professional; and about 12% from other sources.

We asked the participants to rank these sources in order of preference. A Friedman Test on the ranked sources of

support showed that there was a statistically significant difference ($X^2(5) = 1987.664$, $p<.05$). Post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level of $p = 0.003$. There were no significant differences between Relatives and Friends ($Z = -2.153, p = 0.31$), or between Work colleague and Service Provider/manufacturer help desk ($Z = -1.990, p = 0.047$). The overall ranking in order of preference is as shown below:

1. Relatives and Friends
2. Work colleagues and Service Provider/Manufacturer help desk
3. IT repair shop professional
4. Others

We sought to understand the extent of care and intervention in cases where the participants notice a practice they believe to be insecure, and crafted the following scenario:

> Assume you have a sister named Vanessa, and you believe her to be less competent than you in data security. One day you visit her, and while you use her laptop, you notice that her antivirus is not set to automatically scan removable media, such as USB sticks, when they are plugged in. For each of the following options, how much do you agree that it is a good choice?
> A - *Change the settings of the antivirus to enable auto-scan of removable media, and say nothing.*
> B - *Change the settings of the antivirus to enable auto-scan of removable media, and tell Vanessa what you have done.*
> C - *Leave the settings as they are. It is Vanessa's choice to disable auto-scan.*
> D - *Leave the settings as they are. It is not your responsibility.*
> E - *Ask Vanessa why auto-scan is disabled.*

The results showed that 27% of the participants agreed with option A; 68% with option B; 23% with C; 19% with D; and 90% with option E. A Friedman Test on the ranked order of preference showed that there was a statistically significant difference ($X^2(4) = 1634.910, p<.05$) in the choice of the options. Post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level of $p = 0.005$. The overall ranking in order of preference is:

1. E: Ask Vanessa why auto-scan is disabled.
2. B: Change the settings of the antivirus to enable auto-scan of removable media, and tell Vanessa what you have done.
3. C: Leave the settings as they are. It is Vanessa's choice to disable auto-scan.
4. A: Change the settings of the antivirus to enable auto-scan of removable media, and say nothing.
5. D: Leave the settings as they are. It is not your responsibility.

iii. Social Responsibility: As evidenced in the last scenario regarding responsibility towards the security of others, option D received the least agreement (19%), and was the lowest ranked. Our interviews reveal that participants consider security support in the home as an obligation to act for the benefit of *society*. What is more interesting is the scope of this society; who do the participants consider part of their *security/secure society*? "I give it [security advice] to a certain level... I am not an expert in security, but people ask me and I tell them my thoughts... *whoever* asks me... *anyone*.. I mean *colleagues at work, my friends, my relations*" - P40. "[I give advice] to help her... [and to] *everyone if I know them* and I am sympathetic to them" - P36.

We asked our survey participants how likely they are to seek advice or help from a source of support that they believe to be more competent than them. The sources included relative, friend, work colleague, service provider /manufacturer help desk, IT repair shop professional, and others. We found that about 80% are likely to seek advice or help from a relative; about 85% from a friend; about 71% from a work colleague; about 58% from a service provider/manufacturer help desk; about 51% from an IT repair shop professional; and about 16% would seek support from other sources.

We also asked the participants to rank these sources in order of preference. A Friedman Test on the ranked order of preference showed that there was a statistically significant difference ($X^2(5) = 2066.482, p<.05$). Post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level of $p = 0.003$. There was no significant difference between Relatives and Friends ($Z = -0.684, p = 0.494$). The overall ranking in order of preference is as shown below:

1. Relative and Friend
2. Service provider/Manufacturer help desk
3. Work colleague
4. IT repair shop professional
5. Others

There is a significant difference ($Z = -5.618, p = 0.000$) in the likelihood of seeking support from a work colleague (71%) and a Service Provider/manufacturer help desk (58%). However, the rankings indicate a significant difference in reverse; the Service provider/Manufacturer help desk was preferred over a work colleague. We hypothesize this might be because 1) some service providers or device manufacturers do not provide support with security, and 2) the range of services and devices available in homes is too broad, and expecting participants to go to many service providers and manufacturers for assistance is contrary to the finding in [11] where users expect a unitary solution to security problems.

Given the common trend during the interviews where most of the participants indicated that they seek support from friends, relatives, and work colleagues, we wanted to know how likely our participants are to offer support to those that approach them for help. Asked how likely they are to offer advice or technical help when asked by someone they believe to be less competent than them in data security, the results showed that about 80% would likely offer support to a relative; 78% are likely to help a friend; 67% are likely to assist a work colleague; and 41% are likely to offer support to any other people who seek it from them.

4.3.2 Continuity of care

The second characteristic of support in the home that we identified from the interviews is continuity of care. Our participants look for a continuous caring relationship with an identified competent and trusted individual. This is evidenced by the preference for availability (ranked third from

competence and trust), as shown in section 4.2.3. From our analysis, two reasons explain this need: 1) In the case of delegation, one needs someone who will be constantly available, and as [11] also reports that people used to delegate to a "person who had helped them in a previous context, such as in discussing what to get, helping them set up the computer, etc.", and similarly *"I was involved in helping them set up in the first place... I helped a lady buy a computer, I helped her to get it online. So she comes to me all the time for information and she keeps asking me questions. I consult and then go back to her"* - P36; and 2) If something goes wrong as a result of the support someone offered, the victim can easily go back and seek further assistance.

Our study showed that participants are likely to take responsibility for consequences resulting from support they offered; *"I may help to solve the problem"* - P28, *"I would consider that as my responsibility, if it was compromised"* - P47. To verify how widely shared this belief and practice is, we crafted two scenarios: one without indicating that a compromise was due to advice that the participant might have given; the second indicating that the compromise was due to advice that they had offered beforehand. We presented the participants with the same answers to both scenario so that we could test the significance of the difference in taking or accepting responsibility. The first scenario read:

> Assume you have a friend, Catherine, who you believe to be less competent than you in data security. She comes to you for help because she had corrupted files on her computer and thinks she has a virus. What would you do?
> A - *Do nothing.*
> B - *Fix it, if you feel you can.*
> C - *Tell Catherine what to do to fix the problem herself, if you know the solution.*
> D - *Tell Catherine to look for help elsewhere if you feel/find that you cannot fix it.*
> E - *Arrange for a trusted contact to fix it, if you feel/find that you cannot.*
> F - *Arrange for a third party to fix it. You offer to pay.*
> G - *Arrange for a third party to fix it. You offer to help pay (share the cost).*
> H - *Arrange for a third party to fix it. You expect Catherine to pay.*

The results showed that 3% of the participants agreed with option A; 87% agreed with B; 70% agreed with C; 81% agreed with D; 73% agreed with E; 7% agreed with F; 7% agreed with G; and 56% agreed with option H.

While maintaining options A - H, we then presented respondents with an updated scenario as follows:

> Assume you have a friend, Catherine, who you believe to be less competent than you in data security. She comes to you for help because she had corrupted files on her computer and thinks she has a virus. You recall that three months ago, Catherine was trying to install a piece of software, but was failing. She asked for your help. You were busy and told her the antivirus was the problem, and to try turning it off. You now notice the antivirus is off. What would you do?

The results showed that 4% agreed with option A; 90% agreed with option B; 74% agreed with option C; 79% agreed with option D; 77% agreed with option E; 21% agreed with option F; 28% agreed with option G; and 40% agreed with option H.

Test Statistics[a]

	A2 - A1	B2 - B1	C2 - C1	D2 - D1	E2 - E1	F2 - F1	G2 - G1	H2 - H1
Z	-.994[b]	-3.035[b]	-4.136[b]	-4.099[c]	-1.262[b]	-15.572[b]	-16.103[b]	-11.571[c]
Asymp. Sig. (2-tailed)	.320	.002	.000	.000	.207	.000	.000	.000

a. Wilcoxon Signed Ranks Test. b.Based on positive ranks. c. Based on negative ranks.

Figure 5: Test for continuity of care

We ran a Wilcoxon signed-rank test against respective pairs of options to check if the changes in the responses were significant. The test showed significant changes in options B, C, D, F, G, and H. These results are summarised in figure 5, where the options are presented as $x1$ for options from the first scenario, and $x2$ for options from the second scenario; where x represents the respective letter for a given option.

5. DISCUSSION

5.1 Evaluating Security Decisions and Support

Our study has uncovered that participants look for evidence, specifically impact, of security problems for them to feel motivated to practice security. The perceived absence of harm (to themselves or their social circles) is seen as evidence of good security decisions. However, harm arises only when an attack is attempted and then successful: a perceived lack of harm is not sufficient evidence to validate a good security decision for the following reasons.

First is the case where harm occurred but was not perceived by the home user: for instance a user might download malware that steals information in the background without their knowledge. Another instance where the perception of harm can fail is in the situation where a successful attack harms a third party outside the notice of the home user: publicised examples of this are the DDoS attack on DyN DNS servers [5] through compromised IoT devices and the 2014 Lizard Squad attack on XBox live and the Playstation Network [26] through compromised home routers.

Second is the case where harm genuinely did not happen, however this is not always evidence of a good security decision either. In the case where no attack was attempted, a lack of harm is no evidence of effectiveness: vulnerabilities might still be exploitable or countermeasures ineffective. Another situation is where an attack was attempted, but was stopped by a third party before material harm occurred. For instance, a home users' credit card details might have been stolen while shopping on an illegitimate website, but the bank stopped the attacker from using the details.

Only in the third case, where attempted attacks are genuinely mitigated down to no harm, does the perceived absence of harm actually demonstrate evidence of a good security decision. We believe that this is strong evidence that survival/outcome bias is a key element in poor security decisions, and that the wider challenge of evaluating a good security decision is a difficult problem for home computer users (and arguably the wider security community).

Related to the difficulties of evaluating good security deci-

sions is the challenge that home users face when evaluating the competence of those they seek support from. For example, participants reported that the ability to use technical devices better than them was used to support the assessment of competence, however this is not clear evidence of security competence. This problem is somewhat mitigated when home users seek support from people within their social circles, where trust and remedial help may be available in the case where problems arise. However, outside of established relationships and remediation, the challenge remains difficult for home users in telling the difference between a genuinely competent individual, an incompetent individual (who may or may not be aware of the fact), and in the worse case a malicious attacker seeking to take advantage by masquerading as a helpful individual.

Home users need to be able to evaluate the quality of a security decision or source of support. In the absence of clear indicators of quality, a variety of different practices have emerged, yet their effectiveness is questionable. A key challenge remains to uncover the means of making quality more evident to non-experts both for security products/practices, and for the skills, knowledge, and characteristics of those who offer support. This is a hard challenge, particularly where such indicators might then be spoofed by malicious actors, however we believe it is still important to work at making good security evident to non-experts considering the wide variety of non-malicious situations where they may need to make a decision or seek support.

5.2 The Role of Social Networks in Home Security

We have explored the role that social relationships play in security practice in the home. While the need for continuity of care may seem odd, it also reflects common security practices in organisational settings. Even though employees are offered security training and awareness, there are usually support people to whom they can turn to when they have issues. In addition to resolving problems, security support is also responsible for carrying out proactive security activities such as firewall configuration, system patching, network monitoring, and many more. In contrast to this, every home is considered to be responsible for its own security, whether it is competent to do so or not. As a result, a wide variety of different practices exist around seeking and giving support for security in the home context. As Dourish et al. [11] observe, the knowledge and skill of a trusted and competent person is one element of a person's defense against potential threats. In this paper, we have discussed social relationships in the context of informal support networks that exist in the home environment. We postulate that these existing networks can be leveraged to provide appropriate and relevant support to home users.

Prior work has investigated how the security behaviour of home users can be changed. Different improvements to security awareness techniques have been proposed and tested, yet evidence [17] shows that despite claims of being aware, home users still do not practice security. One reason for this is that while awareness might impart knowledge, it does not cover skills; a very essential aspect of security practice. Based on our findings, we argue that the *security posture* of the home is more likely to improve by targeting the support network rather than the user directly for two reasons:

First by targeting the support network, change is introduced at the point where security work is more likely to occur. We believe that by providing tools, training, education, and incentives to those who provide help to others, there is a better chance of achieving a measurable beneficial change to the security of homes.

Second, given the importance of social relationships and the trust placed in the support networks of home computer users, we believe that leveraging these is also a promising approach for transferring both security knowledge and skills to home computer users. Owing to the cost of building a support infrastructure that meets all the requirements discussed in this paper, we believe a fruitful approach is to investigate how social relationships could be leveraged through collaborative technology, social media, and training that focuses on building independent competent communities.

6. CONCLUSIONS

Our research has focussed on the key role of social relationships in home data security, and the reasons behind these informal support networks. We have also uncovered two important factors that explain why some home users do not behave securely: *outcome bias* and *confidence in security measures*. Based on our findings, we put forward the following recommendations:

Leverage existing social relationships: While awareness is important, current practice has focussed on improving the security awareness of individuals or end-users. We suggest focussing on finding ways of targeting existing informal networks of support: building competence, targeting tools, and fostering a sense of trust and recognition. This leverages two characteristics of support currently sought in the home – duty of care and continuity of care.

Simple and useful tools: We need more tools targeted at home users. First, tools that non-experts (especially the existing informal support workers) can use to manage security configurations for different devices and services in the home. Currently, the proliferation of networked devices and services in the home makes the task of managing security complex, and security configurations need to be done on each and every device and service separately. As Dourish et al [11] state, people expect a unitary solution to a number of security problems. Developing tools to manage security configurations of a number of devices and/or services centrally would motivate home users and simplify this task.

Second, tools need to be developed to help the informal support workers that currently assist home users. This might include remote assistance, network monitoring, or incident management tools. It is important to note that this also raises a wide variety of different challenges pertaining to consent, privacy, and standards of care, in addition to fundamental security considerations.

Evidence-based security: Finally, our work has shown that home users look for evidence of harm to evaluate the quality of their security decisions, and to be motivated to make changes. We hypothesise that this might be due to current mechanisms failing to effectively convey knowledge of an attempted or successful incident. This suggests that there is a need to find ways of detecting and communicating (in a simple, concise, and understandable manner) any attempted, successful, and failed attacks.

7. ACKNOWLEDGMENTS

This work was supported by the Research Institute in Science of Cyber Security (RISCS) under grant No. BLR01330, and grant No. BLR01790. Recruitment of participants was done in collaboration with the National Museum of Computing, and Community Centres in Oxford. Norbert is funded by the Rhodes Scholarship (Malawi & Linacre, 2015) and the Oxford-Linacre African Scholarship (2015/16).

8. REFERENCES

[1] C. L. Anderson and R. Agarwal. Practicing safe computing: a multimedia empirical examination of home computer user security behavioral intentions. *Mis Quarterly*, 34(3):613–643, 2010.

[2] AOL/NCSA. Online safety study. https://library.educause.edu/resources/2004/1/aolncsa-online-safety-study, 2017. Online; accessed on 25-August-2017.

[3] K. Aytes and T. Connolly. Computer security and risky computing practices: A rational choice perspective. *Journal of Organizational and End User Computing (JOEUC)*, 16(3):22–40, 2004.

[4] J. Baron and J. C. Hershey. Outcome bias in decision evaluation. *Journal of personality and social psychology*, 54(4):569, 1988.

[5] BBC. Smart home devices used as weapons in website attack. http://www.bbc.co.uk/news/technology-37738823, 2016. Online; accessed on 03-April-2017.

[6] A. Beautement, I. Becker, S. Parkin, K. Krol, and A. Sasse. Productive security: A scalable methodology for analysing employee security behaviours. In *12th Symposium on Usable Privacy and Security (SOUPS)*, 2016.

[7] P. Biernacki and D. Waldorf. Snowball sampling: Problems and techniques of chain referral sampling. *Sociological methods & research*, 10(2):141–163, 1981.

[8] S. Das, T. H.-J. Kim, L. A. Dabbish, and J. I. Hong. The effect of social influence on security sensitivity. In *Proc. SOUPS*, volume 14, 2014.

[9] S. Das, A. D. Kramer, L. A. Dabbish, and J. I. Hong. Increasing security sensitivity with social proof: A large-scale experimental confirmation. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 739–749. ACM, 2014.

[10] Z. Dong, V. Garg, L. J. Camp, and A. Kapadia. Pools, clubs and security: designing for a party not a person. In *Proceedings of the 2012 workshop on New security paradigms*, pages 77–86. ACM, 2012.

[11] P. Dourish, R. E. Grinter, J. D. De La Flor, and M. Joseph. Security in the wild: user strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing*, 8(6):391–401, 2004.

[12] ENISA. Mirai malware, attacks home routers. https://www.enisa.europa.eu/publications/info-notes/mirai-malware-attacks-home-routers, 2016. Online; accessed on 03-April-2017.

[13] R. J. Fisher. Social desirability bias and the validity of indirect questioning. *Journal of consumer research*, 20(2):303–315, 1993.

[14] O. for National Statistics. Internet access - households and individuals 2015. http://www.ons.gov.uk/ons/dcp171778412758.pdf, 2017. Online; accessed on 01-April-2017.

[15] A. Forget, S. Pearman, J. Thomas, A. Acquisti, N. Christin, L. F. Cranor, S. Egelman, M. Harbach, and R. 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.

[16] M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

[17] S. Furnell, P. Bryant, and A. D. Phippen. Assessing the security perceptions of personal internet users. *Computers & Security*, 26(5):410–417, 2007.

[18] A. M. Gadermann, M. Guhn, and B. D. Zumbo. Estimating ordinal reliability for likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, 17(3), 2012.

[19] B. G. Glaser and A. L. Strauss. *The discovery of grounded theory: Strategies for qualitative research*. Transaction publishers, 2009.

[20] P. Gutmann. Applying problem-structuring methods to problems in computer security. In *Proceedings of the 2011 workshop on New security paradigms workshop*, pages 37–44. ACM, 2011.

[21] C. Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Proceedings of the 2009 workshop on New security paradigms workshop*, pages 133–144. ACM, 2009.

[22] A. E. Howe, I. Ray, M. Roberts, M. Urbanska, and Z. Byrne. The psychology of security for the home computer user. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 209–223. IEEE, 2012.

[23] I. Ion, R. Reeder, and S. Consolvo. "... no one can hack my mind": Comparing expert and non-expert security practices. In *SOUPS*, pages 327–346, 2015.

[24] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong. Teaching johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)*, 10(2):7, 2010.

[25] Y. Li and M. T. Siponen. A call for research on home users' information security behaviour. In *PACIS*, page 112, 2011.

[26] P. Lunsford and M. C. Boahn. How the lizard squad took down two of the biggest networks in the world. 2015.

[27] M. S. Mendes, E. Furtado, G. Militao, and M. F. de Castro. Hey, i have a problem in the system: Who can help me? an investigation of facebook users interaction when facing privacy problems. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 391–403. Springer, 2015.

[28] A. J. Nederhof. Methods of coping with social desirability bias: A review. *European journal of social psychology*, 15(3):263–280, 1985.

[29] B.-Y. Ng and M. Rahim. A socio-behavioral study of home computer users' intention to practice security. *PACIS 2005 Proceedings*, page 20, 2005.

[30] M. Nouh, A. Almaatouq, A. Alabdulkareem, V. K.

Singh, E. Shmueli, M. Alsaleh, A. Alarifi, A. Alfaris, et al. Social information leakage: Effects of awareness and peer pressure on user behavior. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 352–360. Springer, 2014.

[31] N. Nthala and I. Flechais. "if it's urgent or it is stopping me from doing something, then i might just go straight at it": A study into home data security decisions. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 123–142. Springer, 2017.

[32] N. Nthala and I. Flechais. Rethinking home network security. In *European Workshop on Usable Security (EuroUSEC)*, 2018.

[33] B. P., F. S.M., and P. A.D. Improving protection and security awareness among home users. *Advances in Networks, Computing and Communications 4*, 2008.

[34] E. Rader and R. Wash. Identifying patterns in informal sources of security information. *Journal of Cybersecurity*, 1(1):121–144, 2015.

[35] E. Rader, R. Wash, and B. Brooks. Stories as informal lessons about security. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, page 6. ACM, 2012.

[36] U. H. Rao and B. P. Pati. Study of internet security threats among home users. In *Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on*, pages 217–221. IEEE, 2012.

[37] E. M. Redmiles, S. Kross, and M. 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.

[38] E. M. Redmiles, A. R. Malone, and M. 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.

[39] J. Rowe, K. Levitt, and M. Hogarth. Towards the realization of a public health system for shared secure cyber-space. In *Proceedings of the 2013 workshop on New security paradigms workshop*, pages 11–18. ACM, 2013.

[40] I. L. Stats. Internet users. http://www.internetlivestats.com/internet-users/, 2017. Online; accessed on 25-August-2017.

[41] A. Strauss and J. Corbin. Basics of qualitative research: Procedures and techniques for developing grounded theory, 1998.

[42] E. U.S. Department of Commerce and S. Administration. Computer and internet use in the united states: 2013. www.census.gov/, 2017. Online; accessed on 01-April-2017.

[43] R. Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, page 11. ACM, 2010.

[44] R. Wash and E. 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.

[45] F. Wilcoxon and R. A. Wilcox. *Some rapid approximate statistical procedures*. Lederle Laboratories, 1964.

[46] G. B. Willis. *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications, 2004.

APPENDIX

A. INTERVIEW DEMOGRAPHIC FORM

1. *Age*: a) 12 - 17, b) 18 - 34, c) 35 - 64, d) 65+
2. *Gender:* a) Male, b) Female
3. *Location:* a) Rural, b) Suburban, c) Urban
4. *What is the highest level of school you have completed?*
a) No schooling completed, b) Nursery, c) High School, d) Trade/technical/vocational training, e) Undergraduate, f) Graduate, g) Postgraduate
5. *Choose one option that best describes your ethnic group or background:*
a) White, b) Hispanic/Latino, c) Black/African/Caribbean, d) Asian/Pacific Islander, e)Other:
6. *Choose the technology devices you own/use in your home:*
a) Mobile Phone, b) Telephone, c) Tablet/iPad, d) Laptop, e) PC, f) Game Console, g)TV, h) Camera, i) Wearable device, j) Other:
7. *Choose the services you use:*
a) Online/Mobile banking, b) Online shopping, c) Social networking, d) Communication, e) Education, f) Entertainment, g) Work, h) Home security, i) TV streaming, j) Health services, k) Other:
8. *How would you rate your general skills in using technology devices, services, and applications?*
a) Novice, b) Competent, c) Expert
9. *How would you rate your general skills in computer security and privacy (e.g. understanding threats, vulnerabilities, and countermeasures)?*
a) Novice, b) Competent, c) Expert
10. *Would you briefly describe the composition of your household?*
A. Marital status: a) Single, b) Married, c) Widowed, d) Divorced, e) Separated
B. Number of people in your household:
C. Relationship with other residents:
D. Age ranges of other residents:
E. Employment status: a) Student, b) Employed, c) Retired, d) Self-employed, e) Not working

B. INTERVIEW GUIDE

B.1 Introductory questions

1. Can you rank these services in order of importance, from the most important to the least important?

B.2 Data Security Concerns and Breaches

2. Do you have any data security concerns with these devices/services/applications?
3. Have you or people you know experienced any data security breaches in the past?

B.3 Security Controls/Tasks

4. What was done to address the data security concerns, and breaches? Who did this?
5. Do you think this was enough to keep your data secure? If not, why?

Open problems in security decision making	Data security concerns	Factors influencing security decisions	Home responsibility
Evaluating the effectiveness or quality of security solution	Loss	Convenience	Source of Support
Unable to have a relevant solution	Loss of control	Cost	Relative
Good Security Practices	Loss of money	Ease of use	Friend
Guidelines and rules for security decision making	Loss of Privacy	Experience	Service provider
Ask the more knowledgeable	Nuisance	Experience in using a security measure	IT shop
Disconnect from the internet when not needed	Uncertainty	Experienced a security breach	Work colleague
Follow advice from a service provider	Security practice	Professional experience	Online forum
Use a tier system of passwords	Insecure practices	Knowledge and skill	Search engine
Don't give out personal details to someone you don't know	Secure Practices	Professional - education	Technical help
Responsibility	Non-security-technology practices	Professional job-related experience	Awareness
Attitude - Giving advice and post breach reaction	Pre-emptive practices	Obligation	Identifying risks
Attitude - Problems arising from well-intended individuals	Pro-active Damage Limitation	Survival/Outcome bias	News
Attitude - Responsible stakeholders	Reactive practices	Perceived Competence	Devices
Boundaries of responsibility	Security-technology practices	experience in using or working with technical devices and services	Services
Understanding responsibility	Reactive - Incident Management	Level of education	Anecdotes
Abrogate responsibility	Noticing a breach	Personal negative experience	Incident reporting behaviour
Noticing responsibility	Risk attitude	Studied or studies a technical course	Security evaluation
Taking responsibility	It's not a risk	Studied or studies data security	Cost of protection
Stakeholders	Not understanding the risk	Technicality of a job	Where to get support
Support	Risk evaluation	Works for a technical company	Reviews
Characteristics of Support	Perceived value of impact	Works in data security	Available security tools or measures to a problem
Continuity of Care	Perceived gain for attacker	Significance	Unsolicited support
Duty of Care	Security incidents experienced	Time pressure (Urgency)	Solicited support
Delegation	Identifying incidents	Trust	Trust evaluating practices
Motivation	Harm	Sharing devices, services and passwords	Relationship with others
Social Responsibility	Security alert	Extent of sharing	Knowledge and skill level
Types of support	Security warning	Purpose of sharing	Closeness to source
Advice	Intuition	Trust cues	Visual cues
Types of advice	Support giving	Availability heuristic	Kinds of information
Opinion	Support seeking	Brand recognition	Confidence in security measure
Recommendation	Availability of support	Interaction	Reviews about a security tool
Information	Quality of support		

Table 1: Grounded Theory Codebook

6. Did you face any problems with the solution?
7. Have you ever adopted or avoided a device/service/application for data security reasons? What prompted you to do this?
8. Have you ever changed settings or abandoned/uninstalled a device/service/application for data security reasons? What prompted you to do this?
9. Is there a particular time when you had data security concerns with a device/service/application but you chose to continue using the device/service/application? Why did you do so?
10. Who is generally responsible for making data security decisions in your home? Why?
11. In the particular scenarios you have mentioned, who made these data security decisions? Why? Were there any difficulties in deciding what to do?
12. If you were to make these decisions for your friend, what would you do? Why?

B.4 Capability and Support

13. Are there any guidelines or rules you follow when making data security decisions? Where do these come from? In the scenarios you mentioned, did you follow these? If not, why?
14. What kind of information/resources do you need when you want to make a data security decision?
15. Where or from who do you seek such information/resources?
16. If you needed advice or technical assistance with data security, where would you seek it?

B.5 Delegation

17. Have you ever given advice/recommendation about data security to other people? Who were they? What kind of advice/recommendation did they want? How much effort did you put in (what did you do)?
18. Why do you think they chose to seek advice/recommendation from you? Why did you give advice/recommendation?
19. Have you made data security decisions and acted on them on behalf of someone? For who was this done? What kind of decisions were these? Why did you do it?
20. If you have given bad advice/recommendation or wrongly decided and acted on behalf of someone and something happened, what would you do? Has this ever happened to you?

B.6 Attitude towards data security

21. Can you give me examples of what you consider good and bad data security (measures/practices)?
22. Who do you think is responsible for implementing this kind of data security in the different devices/services/applications you use?
23. Do you personally follow these measures? If not, why?
24. Do you think any of your actions in using the devices/services/applications could expose other people to data security risks? What are some of these actions and how do you think they might affect others? What do you do about it?

C. SURVEY TOOL

C.1 Demographics

1. Please select your age range: a) 18 - 34, b) 35 - 64, c) 65+
2. Please select your gender: a) Male, b) Female
3. What is the highest educational level you have completed?
a) No schooling completed, b) High school,
c) Trade/technical/vocational training, d) Undergraduate,
e) Graduate, f) Postgraduate

C.2 Survial/Outcome Bias

For the past 5 years, your friend John has been downloading free music, videos, and software from different websites including torrent sites without any problem. One day, he reads an article about the dangers of free downloads such as viruses, adware, Trojan horses, worms and spyware. For each of the following options, how much do you agree that it is a good choice for John?
(Responses: I strongly agree, I agree, Neutral, I disagree, I strongly disagree)
A. Continue downloading free files from any website as usual. He has been doing it for 5 years without a problem, chances of being affected are very small.
B. Restrict the downloads to those websites John has already used before. He has used them for 5 years without a problem, he trusts them to be secure.

How would you rank the options from the scenario above in order of preference?

C.3 Assessing Other's Security Competence

How do you assess if someone is more competent than you in data security? (Please select all that apply.)
A. He/she works for a technical company.
B. His/her job is technical.
C. He/she has more experience than you in using or working with technical devices and services.
D. He/she studied or studies a technical course.
E. He/she is more educated than you.
F. He/she works in data security.
G. He/she studied or studies data security.
H. He/she has experienced a data security incident before.
I. None of the above.

How would you rank the options selected in the question above in order of preference?

C.4 Seeking Support

Assuming you believe each of the following to be more competent than you in data security, how likely are you to seek advice or help with data security from him/her?
(Responses: Very Likely, Likely, Neutral, Unlikely, Very Unlikely)
A. Relative
B. Friend
C. Work colleague
D. Service provider/Manufacturer help desk
E. IT repair shop professional
F. Others

How would you rank the options in the question above in order of preference?

C.5 Accepting Unsolicited Support

Assuming you believe each of the following to be more competent than you in data security, how likely are you to accept unsolicited (not asked for) advice or help with data security from him/her?
(Responses: Very Likely, Likely, Neutral, Unlikely, Very Unlikely)
A. Relative
B. Friend
C. Work colleague
D. Service provider/Manufacturer help desk
E. IT repair shop professional
F. Others

How would you rank the options in the question above in order of preference?

C.6 Giving Solicited Support

Assuming you believe each of the following to be less competent than you in data security, if they ask you for advice or help with data security, how likely are you to offer it?
(Responses: Very Likely, Likely, Neutral, Unlikely, Very Unlikely)
A. Relative
B. Friend
C. Work colleague
D. Others

C.7 Quality Check

I am randomly answering the questions.
A. Yes
B. No

C.8 Giving Unsolicited Support

Assuming you believe each of the following to be less competent than you in data security, how likely are you to offer unsolicited (not asked for) advice or help with data security to him/her?
(Responses: Very Likely, Likely, Neutral, Unlikely, Very Unlikely)
A. Relative
B. Friend
C. Work colleague
D. Others

How would you rank the options in the question above in order of preference?

C.9 Assessing the Quality and Source of Support

Which of the following do you take into consideration when seeking data security advice or help from someone? (Please select all that apply)
A. Competence
B. Availability
C. Trust
D. Closeness to you
E. Cost to you (money, favours, gifts, etc)
F. Cost to the source of advice/help (effort, inconvenience, etc)
G. None of the above

How would you rank the options in the question above in order of preference?

C.10 Confidence in a Security Measure

Your friend Felicity is a college student. She owns a laptop. She stores assignments and study materials on it. Felicity visits her friend, Laurel, whom she finds watching a very interesting movie. Felicity asks Laurel if she can share the movie with her, as well as some of the music Laurel downloaded. Laurel copies all the files to a USB stick, and hands it over to Felicity. On their way out, Laurel tells Felicity that she thinks her laptop might have a virus because she could not open one of her word documents to study, and this has happened to her a number of times. For each of the following options, how much do you agree that it is a good

choice for Felicity?
(Responses: I strongly agree, I agree, Neutral, I disagree, I strongly disagree)
A. Felicity could copy the movie and music to her laptop. Laurel probably got a corrupted file, there is nothing to fear.
B. Felicity could copy the files to her laptop. She has an antivirus which will keep her data secure.
C. Felicity could take and maintain a backup of her files in a USB stick, phone storage, cloud storage, external hard drive, another computer, etc. She could hence copy the movie and music to her laptop. She can always get the files from the backup when needed.

How would you rank the options in the question above in order of preference?

C.11 Duty of Care

Assume you have a sister named Vanessa, and you believe her to be less competent than you in data security. One day you visit her, and while you use her laptop, you notice that her antivirus is not set to automatically scan removable media, such as USB sticks, when they are plugged in. For each of the following options, how much do you agree that it is a good choice?
(Responses: I strongly agree, I agree, Neutral, I disagree, I strongly disagree)
A. Change the settings of the antivirus to enable auto-scan of removable media, and say nothing.
B. Change the settings of the antivirus to enable auto-scan of removable media, and tell Vanessa what you have done.
C. Leave the settings as they are. It is Vanessa's choice to disable auto-scan.
D. Leave the settings as they are. It is not your responsibility.
E. Ask Vanessa why auto-scan is disabled.

How would you rank the options in the question above in order of preference?

C.12 Quality Check

I am randomly answering the questions.
A. Yes B. No

C.13 Continuity of Care - Scenario 1

Assume you have a friend, Catherine, who you believe to be less competent than you in data security. She comes to you for help because she had corrupted files on her computer and thinks she has a virus. What would you do?
(Responses: I strongly agree, I agree, Neutral, I disagree, I strongly disagree)
A. Do nothing.
B. Fix it, if you feel you can.
C. Tell Catherine what to do to fix the problem herself, if you know the solution.
D. Tell Catherine to look for help elsewhere if you feel/find that you cannot fix it.
E. Arrange for a trusted contact to fix it, if you feel/find that you cannot.
F. Arrange for a third party to fix it. You offer to pay.
G. Arrange for a third party to fix it. You offer to help pay (share the cost).
H. Arrange for a third party to fix it. You expect Catherine to pay.

C.14 Continuity of Care - Scenario 2

Assume you have a friend, Catherine, who you believe to be less competent than you in data security. She comes to you for help because she had corrupted files on her computer and thinks she has a virus. You recall that three months ago, Catherine was trying to install a piece of software, but was failing. She asked for your help. You were busy and told her the antivirus was the problem, and to try turning it off. You now notice the antivirus is off. What would you do?
(Responses: I strongly agree, I agree, Neutral, I disagree, I strongly disagree)
A. Do nothing.
B. Fix it, if you feel you can.
C. Tell Catherine what to do to fix the problem herself, if you know the solution.
D. Tell Catherine to look for help elsewhere if you feel/find that you cannot fix it.
E. Arrange for a trusted contact to fix it, if you feel/find that you cannot.
F. Arrange for a third party to fix it. You offer to pay.
G. Arrange for a third party to fix it. You offer to help pay (share the cost).
H. Arrange for a third party to fix it. You expect Catherine to pay.

D. SUMMARY STATISTICS

Survival/Outcome Bias:

	Strongly Agree	*Agree*	*Neutral*	*Disagree*	*Strongly Disagree*
A.	34 (3.1%)	203 (18.7%)	237 (21.8%)	430 (39.6%)	183 (16.8%)
B.	109 (10%)	386 (35.5%)	269 (24.7%)	247 (22.7%)	76 (7%)

Confidence in a Security Measure:

	Strongly Agree	*Agree*	*Neutral*	*Disagree*	*Strongly Disagree*
A.	31 (2.9%)	120 (11%)	167 (15.4%)	420 (38.6%)	349 (32.1%)
B.	41 (3.8%)	242 (22.3%)	272 (25%)	385 (35.4%)	147 (13.5%)
C.	156 (14.4%)	345 (31.7%)	264 (24.3%)	218 (20.1%)	104 (9.6%)

Assessing the Quality and Source of Support:

Which of the following do you take into consideration when seeking data security advice or help from someone? (Please select all that apply)

A. Competence	993 (91.4%)	E. Cost to you (money, favours, gifts, etc)	530 (48.8%)
B. Availability	334 (30.7%)	F. Cost to the source of advice/help (effort, inconvenience, etc)	394 (36.2%)
C. Trust	971 (89.3%)	G. None of the above	11 (1%)
D. Closeness to you	339 (31.2%)		

How would you rank the options in the question above in order of preference?

	0 (No rank)	*1*	*2*	*3*	*4*	*5*	*6*
A.	6 (.6%)	660 (60.7%)	273 (25.1%)	79 (7.3%)	39 (3.6%)	17 (1.6%)	13 (1.2%)
B.	22 (2%)	21 (1.9%)	77 (7.1%)	293 (27%)	283 (26%)	232 (21.3%)	159 (14.6%)
C.	4 (.4%)	321 (29.5%)	541 (49.8%)	118 (10.9%)	62 (5.7%)	32 (2.9%)	9 (.8%)
D.	28 (2.6%)	20 (1.8%)	71 (6.5%)	212 (19.5%)	195 (17.9%)	224 (20.6%)	337 (31%)
E.	21 (1.9%)	50 (4.6%)	82 (7.5%)	234 (21.5%)	252 (23.2%)	248 (22.8%)	200 (18.4%)
F.	24 (2.2%)	14 (1.3%)	38 (3.5%)	140 (12.9%)	232 (21.3%)	304 (28%)	335 (30.8%)

Assessing Other People's Security Competence:

How do you assess if someone is more competent than you in data security? (Please select all that apply.)

A. He/she works for a technical company	255 (23.5%)	F. He/she works in data security	938 (86.3%)
B. His/her job is technical	255 (23.5%)	G. He/she studied or studies data security	846 (77.8%)
C. He/she has more experience than you in using or working with technical devices and services	559 (51.4%)	H. He/she has experienced a data security incident before	428 (39.4%)
D. He/she studied or studies a technical course	289 (26.6%)	I. None of the above	47 (4.3%)
E. He/she is more educated than you	75 (6.9%)		

How would you rank the options in the question above in order of preference?

	0 (No rank)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
A.	36 (3.3%)	38 (3.5%)	47 (4.3%)	106 (9.8%)	176 (16.2%)	180 (16.6%)	233 (21.4%)	188 (17.3%)	83 (7.6%)
B.	32 (2.9%))	38 (3.5%)	41 (3.8%)	116 (10.7%)	201 (18.5%)	269 (24.7%)	229 (21.1%)	131 (12.1%)	30 (2.8%)
C.	21 (1.9%)	96 (8.8%)	73 (6.7%)	231 (21.3%)	252 (23.2%)	164 (15.1%)	140 (12.9%)	96 (8.8%)	14 (1.3%)
D.	32 (2.9%)	16 (1.5%)	38 (3.5%)	103 (9.5%)	173 (15.9%)	254 (23.4%)	249 (22.9%)	190 (17.5%)	32 (2.9%)
E.	39 (3.6%)	11 (1%)	15 (1.4%)	24 (2.2%)	41 (3.8%)	36 (3.3%)	65 (6%)	188 (17.3%)	668 (61.5%)
F.	8 (.7%)	730 (67.2%)	163 (15%)	71 (6.5%)	38 (3.5%)	33 (3%)	20 (1.8%)	18 (1.7%)	6 (.6%)
G.	18 (1.7%)	115 (10.6%)	634 (58.3%)	138 (12.7%)	55 (5.1%)	45 (4.1%)	32 (2.9%)	25 (2.3%)	25 (2.3%)
H.	22 (2%)	42 (3.9%)	68 (6.3%)	284 (26.1%)	124 (11.4%)	74 (6.8%)	84 (7.7%)	208 (19.1%)	181 (16.7%)

Duty of Care: Motivation - Offer Unsolicited Support:

	Very Likely	*Likely*	*Neutral*	*Unlikely*	*Very Unlikely*
A. Relative	209 (19.2%)	396 (36.4%)	180 (16.6%)	209 (19.2%)	93 (8.6%)
B. Friend	137 (12.6%)	376 (34.6%)	223 (20.5%)	250 (23%)	101 (9.3%)
C. Work colleague	55 (5.1%)	236 (21.7%)	268 (24.7%)	351 (32.3%)	177 (16.3%)
D. Others	24 (2.2%)	107 (9.8%)	275 (25.3%)	350 (32.2%)	331 (30.5%)

How would you rank the options in the question above in order of preference?

	0 (No rank)	*1*	*2*	*3*	*4*
A.	4 (.4%)	756 (69.5%)	196 (18%)	97 (8.9%)	34 (3.1%)
B.	5 (.5%)	216 (19.9%)	746 (68.6%)	113 (10.4%)	7 (.6%)
C.	5 (.5%)	96 (8.8%)	123 (11.3%)	812 (74.7%)	51 (4.7%)
D.	12 (1.1%)	18 (1.7%)	17 (1.6%)	58 (5.3%)	982 (90.3%)

Duty of Care: Motivation - Accept Unsolicited Support:

	Very Likely	*Likely*	*Neutral*	*Unlikely*	*Very Unlikely*
A. Relative	195 (17.9%)	485 (44.6%)	232 (21.3%)	134 (12.3%)	41 (3.8%)
B. Friend	174 (16%)	514 (47.3%)	253 (23.3%)	117 (10.8%)	29 (2.7%)
C. Work colleague	103 (9.5%)	424 (39%)	331 (30.5%)	172 (15.8%)	57 (5.2%)
D. Service provider/ Manufacturer help desk	135 (12.4%)	347 (31.9%)	266 (24.5%)	222 (20.4%)	117 (10.8%)
E. IT repair shop professional	118 (10.9%)	322 (29.6%)	253 (23.3%)	242 (22.3%)	152 (14%)
F. Others	26 (2.4%)	109 (10%)	424 (39%)	283 (26%)	245 (22.5%)

How would you rank the options in the question above in order of preference?

	0 (No rank)	*1*	*2*	*3*	*4*	*5*	*6*
A.	7 (.6%)	405 (37.3%)	192 (17.7%)	211 (19.4%)	119 (10.9%)	119 (10.9%)	34 (3.1%)
B.	8 (.7%)	220 (20.2%)	403 (37.1%)	167 (15.4%)	203 (18.7%)	76 (7%)	10 (.9%)
C.	9 (.8%)	95 (8.7%)	134 (12.3%)	456 (42%)	127 (11.7%)	235 (21.6%)	31 (2.9%)
D.	10 (.9%)	232 (21.3%)	149 (13.7%)	129 (11.9%)	332 (30.5%)	186 (17.1%)	49 (4.5%)
E.	8 (.7%)	123 (11.3%)	191 (17.6%)	93 (8.6%)	227 (20.9%)	378 (34.8%)	67 (6.2%)
F.	12 (1.1%)	11 (1%)	13 (1.2%)	22 (2%)	67 (6.2%)	81 (7.5%)	881 (81%)

Duty of Care: Motivation:

	Strongly Agree	*Agree*	*Neutral*	*Disagree*	*Strongly Disagree*
A.	61 (5.6%)	233 (21.4%)	239 (22%)	397 (36.5%)	157 (14.4%)
B.	350 (32.2%)	393 (36.2%)	171 (15.7%)	134 (12.3%)	39 (3.6%)
C.	56 (5.2%)	189 (17.4%)	319 (29.3%)	402 (37%)	121 (11.1%)
D.	48 (4.4%)	160 (14.7%)	269 (24.7%)	414 (38.1%)	196 (18%)
E.	539 (49.6%)	436 (40.1%)	63 (5.8%)	36 (3.3%)	13 (1.2%)

How would you rank the options in the question above in order of preference?

	0 (No rank)	*1*	*2*	*3*	*4*	*5*
A.	9 (.8%)	56 (5.2%)	131 (12.1%)	413 (38%)	118 (10.9%)	360 (33.1%)
B.	5 (.5%)	255 (23.5%)	509 (46.8%)	96 (8.8%)	191 (17.6%)	31 (2.9%)
C.	9 (.8%)	54 (5%)	160 (14.7%)	268 (24.7%)	503 (46.3%)	93 (8.6%)
D.	10 (.9%)	26 (2.4%)	108 (9.9%)	163 (15%)	237 (21.8%)	543 (50%)
E.	2 (.2%)	695 (63.9%)	174 (16%)	140 (12.9%)	28 (2.6%)	48 (4.4%)

Duty of Care: Social Responsibility - Seek Support:

	Very Likely	*Likely*	*Neutral*	*Unlikely*	*Very Unlikely*
A. Relative	383 (35.2%)	485 (44.6%)	128 (11.8%)	71 (6.5%)	20 (1.8%)
B. Friend	362 (33.3%)	561 (51.6%)	124 (11.4%)	29 (2.7%)	11 (1%)
C. Work colleague	207 (19%)	562 (51.7%)	214 (19.7%)	79 (7.3%)	25 (2.3%)
D. Service provider/ Manufacturer help desk	224 (20.6%)	402 (37%)	279 (25.7%)	150 (13.8%)	32 (2.9%)
E. IT repair shop professional	186 (17.1%)	367 (33.8%)	262 (24.1%)	206 (19%)	66 (6.1%)
F. Others	46 (4.2%)	123 (11.3%)	561 (51.6%)	239 (22%)	118 (10.9%)

How would you rank the options in the question above in order of preference?

	0 (No rank)	*1*	*2*	*3*	*4*	*5*	*6*
A.	10 (.9%)	361 (33.2%)	205 (18.9%)	229 (21.1%)	119 (10.9%)	132 (12.1%)	31 (2.9%)
B.	8 (.7%)	243 (22.4%)	393 (36.2%)	173 (15.9%)	198 (18.2%)	62 (5.7%)	10 (.9%)
C.	10 (.9%)	98 (9%)	138 (12.7%)	410 (37.7%)	166 (15.3%)	236 (21.7%)	29 (2.7%)
D.	11 (1%)	232 (21.3%)	159 (14.6%)	151 (13.9%)	305 (28.1%)	196 (18%)	33 (3%)
E.	9 (.8%)	145 (13.3%)	176 (16.2%)	97 (8.9%)	235 (21.6%)	357 (32.8%)	68 (6.3%)
F.	14 (1.3%)	6 (.6%)	9 (.8%)	17 (1.6%)	52 (4.8%)	90 (8.3%)	899 (82.7%)

Duty of Care: Social Responsibility - Seek Support:

	Very Likely	*Likely*	*Neutral*	*Unlikely*	*Very Unlikely*
A. Relative	479 (44.1%)	388 (35.7%)	80 (7.4%)	95 (8.7%)	45 (4.1%)
B. Friend	454 (41.8%)	398 (36.6%)	86 (7.9%)	102 (9.4%)	47 (4.3%)
C. Work colleague	286 (26.3%)	439 (40.4%)	161 (14.8%)	136 (12.5%)	65 (6%)
D. Others	147 (13.5%)	302 (27.8%)	307 (28.2%)	215 (19.8%)	116 (10.7%)

Continuity of Care - Scenario 1:

	Strongly Agree	*Agree*	*Neutral*	*Disagree*	*Strongly Disagree*
A.	12 (1.1%)	24 (2.2%)	116 (10.7%)	452 (41.6%)	483 (44.4%)
B.	389 (35.8%)	561 (51.6%)	80 (7.4%)	42 (3.9%)	15 (1.4%)
C.	146 (13.4%)	613 (56.4%)	188 (17.3%)	114 (10.5%)	26 (2.4%)
D.	327 (30.1%)	556 (51.1%)	100 (9.2%)	77 (7.1%)	27 (2.5%)
E.	263 (24.2%)	535 (49.2%)	192 (17.7%)	81 (7.5%)	16 (1.5%)
F.	11 (1%)	66 (6.1%)	131 (12.1%)	464 (42.7%)	415 (38.2%)
G.	20 (1.8%)	72 (6.6%)	148 (13.6%)	442 (40.7%)	405 (37.3%)
H.	153 (14.1%)	459 (42.2%)	281 (25.9%)	130 (12%)	64 (5.9%)

Continuity of Care - Scenario 2:

	Strongly Agree	*Agree*	*Neutral*	*Disagree*	*Strongly Disagree*
A.	14 (1.3%)	29 (2.7%)	98 (9%)	491 (45.2%)	455 (41.9%)
B.	424 (39%)	549 (50.5%)	62 (5.7%)	37 (3.4%)	15 (1.4%)
C.	200 (18.4%)	604 (55.6%)	160 (14.7%)	98 (9%)	25 (2.3%)
D.	240 (22.1%)	615 (56.6%)	131 (12.1%)	69 (6.3%)	32 (2.9%)
E.	252 (32.3%)	584 (53.7%)	161 (14.8%)	61 (5.6%)	29 (2.7%)
F.	63 (5.8%)	168 (15.5%)	219 (20.1%)	367 (33.8%)	270 (24.8%)
G.	65 (6%)	235 (21.6%)	211 (19.4%)	331 (30.5%)	245 (22.5%)
H.	90 (8.3%)	347 (31.9%)	315 (29%)	241 (22.2%)	94 (8.6%)

Share and Share Alike? An Exploration of Secure Behaviors in Romantic Relationships

Cheul Young Park, Cori Faklaris, Siyan Zhao, Alex Sciuto, Laura Dabbish, Jason Hong

Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA, USA
cheulyop@andrew.cmu.edu, {cfaklari, siyanz, dabbish, jasonh}@cs.cmu.edu,
sciutoalex@gmail.com

ABSTRACT

Security design choices often fail to take into account users' social context. Our work is among the first to examine security behavior in romantic relationships. We surveyed 195 people on Amazon Mechanical Turk about their relationship status and account sharing behavior for a cross-section of popular websites and apps (e.g., Netflix, Amazon Prime). We examine differences in account sharing behavior at different stages in a relationship and for people in different age groups and income levels. We also present a taxonomy of sharing motivations and behaviors based on the iterative coding of open-ended responses. Based on this taxonomy, we present design recommendations to support end users in three relationship stages: when they start sharing access with romantic partners; when they are maintaining that sharing; and when they decide to stop. Our findings contribute to the field of usable privacy and security by enhancing our understanding of security and privacy behaviors and needs in intimate social relationships.

1. INTRODUCTION

Sharing digital accounts is a common practice for various social groups and individuals. Recent Twitter discussion among members of the UK's Parliament sharing their account credentials shows that password sharing is widespread even among groups that require maximum levels of information security [23]. Studies report employees share account credentials with their colleagues, as sharing can facilitate trust and productivity [7, 24, 30]. Sharing is more common among intimate social groups such as families and friends. Researchers found people share accounts to overcome resource limitations [41], while convenience, combined with proximity, also motivates sharing [14, 34]. In a broader context, sharing has been recognized as a token of "trust," which enables a society to perform its functions [8, 30, 34, 41].

Sharing is gaining traction in security research community as the emphasis on the "human side" of computer security is growing [1, 35, 40]. Researchers are beginning to focus on designing secure systems that accommodate sharing. Still, many designs of online systems assume a single user – an assumption that would be considered ridiculous if those systems were situated in an offline environment. More than a decade ago, Grinter et al. showed that a home entertainment system designed for a single user can be unsuitable for a multi-user scenario and even create conflict among household members [18]. Recent work by Matthews et al. shows that while households may share devices and accounts in daily use, there is scarce support for sharing among current technologies [34].

In this regard, research on the sharing practices of couples in romantic relationships can inform future designs of security technologies that afford sharing behaviors. Further, dyadic romantic relationships are the most pervasive social constructs, but they have been left mostly unexplored concerning cybersecurity.

To address this gap in the literature, we conducted an online survey in 2017. The survey was distributed on Amazon Mechanical Turk, targeting people who have experienced romantic relationships. We collected quantitative data on what accounts people share with their partners, demographics, relationship duration, cohabitation duration, and qualitative responses on how and why they share. We were interested in 1) how sharing behaviors differ individually and 2) how tendencies of sharing for various types of accounts differ with the progress of a relationship.

We found that account sharing among couples emerges both from needs to fulfill functional goals such as sharing finances, as well as from desires to satisfy each other's emotional needs. Our findings suggest that account sharing plays a critical role in the progression of romantic relationships, supporting the notion of creating affordances for shared usage in online accounts. We also report hiding behaviors and examine underlying rationales. Finally, we present design recommendations to support sharing in different stages of a relationship.

The contributions of our work are as follows:

- We provide a snapshot of account sharing behaviors of people in romantic relationships.
- We extend the literature on account and password sharing to the context of romantic relationships.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

- We provide guidelines for designers and developers of security systems to better support account sharing behaviors of romantic couples in different relationship stages.

2. RELATED WORK

2.1 The Social Context of Security Behaviors

We are applying a social psychology lens to problematic security behaviors. This framing specifically builds on the work of Das et al. [9, 10, 11, 12] in gathering and analyzing empirical data about end users' triggers for security behavior change. These triggers include observations of friends' and loved ones' security behaviors, social sensemaking of security practices and beliefs, pulling pranks and otherwise demonstrating to peers and family various security behaviors, and sharing account access and passwords with close ties. Das' findings have been echoed by others such as Redmiles et al. [39], who found in a 2016 census-representative survey of N=526 U.S. residents that family and friends, along with media, were the most prevalent sources of security advice.

Other authors have also examined security behaviors in a social context. Singh et al. [41] reported results of a 2005-2006 qualitative study of how people in Australia use banking services and manage money in the context of their personal relationships and in their broader socio-economic contexts. The data collected through open-ended interviews with a total of N=108 Australians of largely European heritage, indigenous "yarning circles" and focus groups of people with disabilities found that couples in relationships share PINs as an expression of trust and that sharing of confidential or private security information is inevitable under certain life circumstances, such as when accessing a service, is difficult due to factors such as remoteness or disability.

More recently, Matthews et al. [34] found in a 2016 mixed-methods study that households' sharing of devices and accounts is common. Participants in a survey of N=99 households, followed by a 25-day diary study of N=25 individuals and interviews with N=24, reported a fluid boundary of what is perceived as "personal," such as mobile devices lying around the house. Trust and convenience were found to be major influences on sharing. These findings are congruent with those of boyd [8] and Singh et al. [41] among others on how family environments socialize family members to share passwords, and other researchers such as Herley on how end users judge costs and benefits in applying security behaviors [21]. Additionally, Matthews et al. developed a taxonomy of sharing with six categories (borrowing, mutual use, setup, helping, broadcasting, and accidental) that suggests a guide for our interpretations of participant sharing data.

Our work extends this literature on the social context of security behaviors to the specific context of romantic relationships. While couples have comprised a subset of the participant groups in prior work, ours is among the first studies, and is the first that we are aware of, to focus exclusively on romantic partners as a user population.

2.2 Password Sharing

Singh et al. and Kaye were among the first HCI researchers to specifically examine reasons for and methods of password sharing. Singh et al., [41] in the study noted above, found that the distance and difficulty of travel to a physical bank branch were major factors that led to password sharing among those with physical disabilities and inhabitants of remote and poor villages in Australia. Participants who shared accounts with partners or family also needed to share passwords to facilitate their access to the accounts.

Kaye's sample, by contrast, was drawn from a U.S.-based convenience sample of friends, family and their own ties reached through online communication and social media. In his primarily qualitative study with N=122 participants published in 2011 [28], he reported that gender and age were positively correlated with password sharing, with password sharing the highest among men ages 46-49. Participants who were in a relationship or married had on average 2.8 (SD=3.5) instances of password sharing, whereas people who were single and not in a relationship had on average 1.4 (SD=1.5) instances. This data suggests that password sharing is becoming a behavioral norm in the U.S. for those in romantic relationships and/or heads of households, for which older men traditionally have managed finances and account logistics.

In a 2013 YouGov Norway survey of N=1003 employees age 18 to 64, Helkala and Bakås [20] found that 31% of participants said they share passwords with a partner. The authors noted that many were confused or misguided about how to create and manage strong passwords, reusing passwords across accounts and showing a lack of understanding as to which accounts contained confidential or private information.

Separately, Whitty et al. [43] found in a 2013 online survey of N=497 U.K. professionals age 18-72 that younger people were more likely to share passwords than older people. High scores on scales measuring certain personality traits (lack of perseverance, suggesting boredom or unenthusiasm for tasks; and the tendency of self-monitoring, which implies sensitivity to social and situational cues) were positively correlated with password sharing. However, knowledge of cybersecurity was not correlated with password sharing. This suggests that social and individual psychological factors may be as important, if not more so, than training or access to information about best practices for understanding some individuals' security behaviors.

Our work builds on this prior research by contributing data from a sample population of romantic couples about their password and account sharing behaviors.

2.3 Partner as "Insider Threat"

At least one participant in Kaye's 2011 study reported having a negative experience with password sharing, as her now-ex-boyfriend made use of his knowledge to send threatening emails and delete accounts [28]. Such experiences with intimate-partner harassment and even abuse or violence using shared security information and device access are sadly not uncommon [15, 16]. Freed et al. advocate incorporating safety reviews for such types of attacks into UI evaluations and penetration testing protocols [16], though they acknowledge the difficulty of designing systems to hamper usability for intimate-partner attackers while preserving usability for targeted or third-party users, all of whom may reside in the same households.

In a 2013 study, Muslukhov et al. [36] reported 12% of those surveyed or interviewed reported a negative experience

with unauthorized access of their smartphone, for instance a housemate looking at personal photos and making costly calls while the phone's owner slept. The authors argued for expanding the adversarial threat model used by smartphone security designers and engineers to include threats posed by "insiders (e.g., friends)" who have proximity to users' smartphones and/or knowledge of their everyday behavior. Follow-up studies [31, 32, 42] from the same research group reinforce the notion that perpetrators of security intrusions can be among our most intimate ties, as Marques et al. estimate that as many as 1 in 3 people have snooped on someone else's smartphone, and Usmani et al., that more than 1 in 5 have snooped on someone else's Facebook account. The latter authors identified fun, curiosity, jealousy, animosity and utility as motivations for these intrusions [42].

End users may become more aware of threats, and more likely to hide some data even from intimate partners, due to their increased use of computing devices for social media [32] and for employment activities. Kang et al. [27] found that social media users who are younger and more educated put more personal information online, but also seek more anonymity and hide more components of their identity than those older or less educated or both. In their comparison of a survey sample drawn from Amazon Mechanical Turk and one more representative of the broader U.S. population, the U.S. MTurk users were found to be more likely to seek anonymity and hide identity and to be more worried about their online information than the U.S. public, regardless of their age, gender, education, and social media use. They also found that MTurk workers hide more information from family members, a romantic partner, friends and coworkers than other groups.

Our work attempts to extend this prior research by adding to the knowledge of "hiding" as a distinct user behavior for partners in romantic relationships. While our survey does not specifically address snooping or intimate partner abuse or violence, our findings on hiding could contribute to the overall understanding of the spectrum of possible antisocial security behaviors by users that designers and developers should take into account.

3. METHOD

We used Amazon Mechanical Turk to reach a broader sample of participants in a variety of relationship and cohabitation situations. Although our results may not be generalizable to the entire population, we did not want to limit our scope geographically. Past study has also shown that MTurk subjects are more representative than student and local convenience samples [6], hence supporting our choice of crowdworkers as a primary survey target.

3.1 Survey Design

The survey[1] consisted of three parts; first, we asked our participants what accounts they own; second, we looked into security and account sharing behaviors for each account; and third, we asked participants about their demographics. Before these questions, for screening purpose, we asked our participants for their relationship status, relationship duration, and cohabitation duration.

We initially drafted a list of popular websites in the U.S. from Alexa.com[2]. However, as it did not provide distinct groupings, we reorganized accounts based on their usages and created 17 original categories. For each category, we selected 15 websites ranked most popular by Alexa.com. The list of categories and accounts is in Appendix 1.

3.2 Survey Items

Once participants completed screening, they were asked to select accounts they own from our list. For each chosen account, we asked for its ownership, the usage of an account by both participants and their partners, and the access to an account by partners. Participants were also prompted to enter up to 3 additional accounts if they did not find any account they own from the given list, but those additional entries were excluded from the analysis.

For ownership, we asked participants whether an account is owned by them, by their partner, jointly by both them and their partners, or separately as individual accounts. For the usage of an account, we asked how frequently participants and their partners use an account respectively. We then assessed how easily a partner can access an account. In each of 17 categories, we asked participants to write a short response describing their reasons and methods for sharing any accounts, and the same for hiding any accounts. Lastly, we asked for participants' demographics, which included: age, gender, ethnicity, sexual orientation, household income, and education level. Detailed questions are in Appendix 2.

3.3 Recruitment and Participants

Between August 30 - September 6, 2017, 244 participants were recruited in three batches on Amazon Mechanical Turk. Participation was limited to the U.S. residents aged above 18 with an approval rating over 95% and had more than 1,000 tasks approved. The survey was titled "Romantic Couples and Cybersecurity," and had a description as the following: "What online accounts and devices you and your partner own and share with (or hide from) each other? You must 1) have ever been in (or are in) a romantic relationship; 2) been in a relationship for $>$ 1 month or broke up $<$ 1 yr ago; and 3) aged 18 or more." Once Turkers accepted the HIT, they were redirected to the Qualtrics survey.

Participants were notified that their participation is voluntary and they can terminate their sessions at any time. Before publishing the survey, we pilot-tested the survey with 25 people and asked them to provide feedback on survey taking experience. Based on the received feedback, we made minor modifications to the interface and the flow of the survey. We estimated the survey to take about 25 - 30 minutes to complete and paid $4 to each participant. On average, participants took 36.9 minutes (SD=37.4) to finish the survey, and the median session duration was 26.7 minutes.

3.4 Data Cleaning

From the total of 306 responses, we removed 25.1% of responses (N=77) which were incomplete or entered by Turkers outside the U.S. We also excluded 34 logically faulty responses which included accounts being used by neither participants nor their partners from the rest of 229 responses. We analyzed the remaining 195 responses each from a unique participant. Only 4 among 195 responses did not report any

[1] http://cmu.ca1.qualtrics.com/jfe/form/SV_beZL6a2GYEOjgwt

[2] https://www.alexa.com/topsites/countries/US

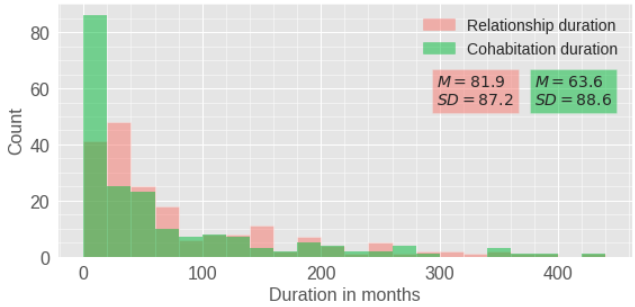


Figure 1: Distributions of relationship duration and cohabitation duration with respective means and standard deviations (N=195).

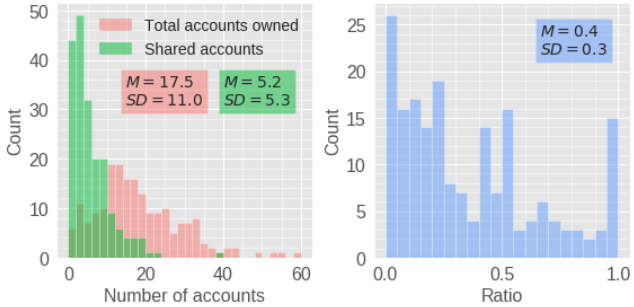


Figure 2: Distributions of a total number of accounts, number of shared accounts, and the ratio of sharing (N=195).

account. We removed 265 duplicate entries of accounts from the total of 3,686 accounts to prevent double-counting. We also refined our account categories as our initial categorization of accounts was ambiguous and not suitable for the analysis. The new categories are in Appendix 1.

Whether an account is shared or not was determined with the following criteria. While a partner must have ready access or be able to access whenever needed, 1) a partner must use an account more than never if a participant owns an account, or 2) a participant must use an account more than never if a partner owns an account, or 3) an account is jointly owned by both a participant and his/her partner.

For hiding, an account was considered actively hidden if a participant selected "Partner doesn't know and I'm actively hiding the account" for the question asking partner's access to an account. However, we noticed many participants mentioned hiding in their open-ended responses although they did not explicitly indicate active hiding of accounts in prior questions.

4. RESULTS

We examined what factors affect sharing of accounts with quantitative data and identified themes that categorize people's motivations and methods for sharing from qualitative responses. 3 authors participated in iteratively developing the taxonomy of sharing reasons from the textual data.

4.1 Sample Characteristics

In our sample, 4% of participants (N=8) were not currently in a relationship, 62% (N=122) were dating someone, and 34% (N=65) were married. 140 participants responded that they are currently living together with their partners and 55 responded they are not. The relationship duration of the participants varied from the minimum of two months to the maximum of 434 months (M=81.9, SD=87.2). Cohabitation duration also varied widely from zero for those who are not living together to the maximum of 434 months (M=63.6, SD=88.6). Figure 1 shows distributions of relationship duration and cohabitation duration.

Previous studies have shown that U.S. Turkers are distinct from the general U.S. population. Researchers found Turkers tend to be younger, more educated, less wealthy, more white, and predominantly females [22, 26, 33, 37]. The characteristics of our survey sample are mostly consistent with that of MTurk populations studied in the past. Ages of our participants ranged from 19 to 63 years old, with 33 as the median (M=34.2, SD=8.91). 81 participants reported education above bachelor's degree, and the median level of education was an associate or technical degree with 120 participants above the median. Our sample consisted of 111 males and 83 females (male/female ratio=1.34), and one who reported being non-binary. The median income range was $40,000 to $59,999 with 55 participants, and the largest number of participants belonged to the range of $20,000 to $39,999 with 65 of them in the group. 153 participants in our sample identified themselves as white, followed by 18 black or African American, 13 Asian or Pacific Islander, 6 Hispanic or Latino, 1 Native American or American Indian, and 6 other ethnicities. Overall, our sample was younger, less wealthy, and more educated compared to the general U.S. public. Appendix 3 shows differences in demographics of our sample and the U.S. population in greater detail.

4.2 Factors Affecting Sharing

To eliminate the chance of difference in sharing across groups stemming from one group having more accounts than the other group, we used the ratio of sharing (the number of shared accounts divided by the number of owned accounts) as our response variable instead of the absolute number of shared accounts. In doing so, we hypothesized based on findings from the security literature that who are younger and more educated will share less [25, 27, 28], while who have less income and were in a relationship/cohabiting for a longer time will share more [34, 41]. As we tested multiple hypotheses simultaneously, we applied Bonferroni correction and used the critical value of 0.05/22=0.0023.

The number of accounts owned and accounts shared were distributed as shown in Figure 2. Overall, 84.6% (N=165) participants out of 195 were sharing at least one account, and one participant sharing 39 accounts was the maximum. The median for number of shared accounts was 4 and sharing ratios were distributed as shown in the right subgraph of Figure 2.

4.2.1 Individual differences based on demographics and relationship characteristics

In our analyses, we used the subset of 174 participants excluding 8 participants who were not in a relationship, 8 with outlying ages, 4 who did not report any account, and 1 participant of non-binary gender. With binary variables including gender, marriage, and cohabitation, we compared the ratio of sharing across two groups (male vs. female, married vs. unmarried, and cohabiting vs. not cohabiting). For categorical or continuous variables such as income, education, age, relationship duration, and cohabitation duration, we

Table 1: Differences in sharing due to demographics and relationship characteristics (N=174).

	Summary statistics								
Explanatory variables	U	p	d	N_1	SD_1	Mdn_1	N_2	SD_2	Mdn_2
Gender (1=female, 2=male)	3719	0.98	0.00	75	0.27	0.27	99	0.32	0.25
Age (1=above median, 2=below median)	3883	0.77	0.03	86	0.26	0.27	88	0.33	0.25
	724	0.82	0.03	37	0.27	0.25	38	0.28	0.31
	1306	0.54	0.07	46	0.30	0.36	53	0.34	0.23
Marriage (1=married, 2=unmarried)	4426	0.001*	0.30	59	0.27	0.43	115	0.31	0.21
	843	0.13	0.21	35	0.24	0.41	40	0.29	0.21
	1292	0.001*	0.44	24	0.27	0.47	75	0.32	0.22
Cohabitation (1=cohabiting, 2=not cohabiting)	4350	<0.001**	0.52	130	0.28	0.39	44	0.32	0.07
	590	<0.001**	0.68	64	0.26	0.33	11	0.28	0.05
	1617	<0.001**	0.49	66	0.29	0.41	33	0.33	0.07
Relationship duration (1=above median, 2=below median)	4902	<0.001**	0.30	87	0.27	0.42	87	0.32	0.21
	692	0.91	-0.02	37	0.24	0.27	38	0.30	0.29
	1730	<0.001**	0.41	49	0.29	0.42	50	0.33	0.14
Cohabitation duration (1=above median, 2=below median)	4977	<0.001**	0.32	86	0.27	0.42	88	0.32	0.20
	870	0.07	0.24	36	0.24	0.41	39	0.29	0.22
	1843	<0.001**	0.50	49	0.27	0.43	50	0.33	0.10
Education (1=above median, 2=below median)	3442	0.31	-0.09	84	0.26	0.24	90	0.33	0.31
	620	0.47	-0.10	32	0.22	0.24	43	0.30	0.30
	265	0.84	-0.05	6	0.12	0.28	93	0.33	0.25
Income (1=above median, 2=below median)	3060	0.80	0.03	47	0.27	0.33	127	0.31	0.24
	593	0.91	0.02	22	0.26	0.35	53	0.28	0.24
	957	0.80	0.04	25	0.27	0.33	74	0.34	0.24

† ** p<0.001, * p<0.0023. For each major row except gender, the top subrow shows the result of a test including both males and females, while the middle and the bottom subrows show results of tests with only females or males respectively.

‡ Column 1 through 3 under summary statistics each show a U-statistic for Mann-Whitney U test, a p-value, and Cliff's delta (effect size). Column 4 through 6 are sample size, standard deviation, and median sharing ratio for group 1, and column 7 through 9 are the same but for group 2.

split the data at corresponding medians to get two groups: one above the median (group 1) and one below the median (group 2). Although splitting data at the median age of 32 or the median relationship duration of 50.5 months is arbitrary, it was necessary for testing differences across variables which were distributed non-normally. For the same reason, we used the nonparametric Mann-Whitney U test instead of the t-test. The summary of results is in Table 1.

The results show that there are no significant differences in sharing due to gender, age, education, and income. Only marriage, cohabitation, relationship duration, and cohabitation duration were significant with positive effect sizes.

One explanation is that marriage and cohabitation, per se, work as a "leap of faith" that triggers a considerable proportion of sharing. Researchers have noted the linear progression of self-disclosure in the developmental trajectory of personal relationships [4, 17, 38], which explains positive associations of relationship duration and cohabitation duration with sharing. Another interesting observation is variables that positively affect sharing show greater significance in males than females. While many factors may be in play, it is possible that our results reflect the tendency of males being registered owners of jointly owned properties in relationships traditionally.

4.2.2 *Combined effects of variables*

Hierarchical logistic regression was conducted with the same subset of 174 participants to study combined effects of variables on sharing. We used the variable indicating if the ratio of sharing is above or below the median of 0.258 (25.8%) as our dependent variable. Transforming sharing ratio to a binary variable rather than treating it as a numeric variable led to a loss of information. However, a linear model with numeric sharing ratio as its dependent variable did not meet assumptions required for a general linear model, e.g., the normal distribution of residuals and the zero mean of residuals, and failed to provide satisfactory explanations for our data. We also tried log-transforming sharing ratio after adding 1 to all values, but the distribution of ratios

Table 2: Hierarchical logistic regression to test effects of multiple variables on sharing (N=174).

	Independent variables								
Model	*Marriage*	*Cohabitation*	*Age*	*Rel. duration*	*Cohab. duration*	*Gender*	*Income*	*Education*	R^2
1	2.11 (1.22, 3.63)*	-	-	-	-	-	-	-	0.032
2	-	1.50 (1.06, 2.13)*	-	-	-	-	-	-	0.022
3	1.82 (0.90, 3.70)	1.16 (0.73, 1.84)	-	-	-	-	-	-	0.033
4	2.47 (1.18, 5.15)*	4.17 (1.83, 9.49)*	0.96 (0.94, 0.98)*	-	-	-	-	-	0.101
5	2.48 (1.05, 5.89)*	4.18 (1.82, 9.61)*	0.96 (0.94, 0.98)*	1.00 (0.98, 1.02)	1.00 (0.98, 1.02)	-	-	-	0.101
6	2.50 (1.05, 5.96)*	4.43 (1.91, 10.30)*	0.96 (0.94, 0.99)*	1.00 (0.99, 1.02)	1.00 (0.98, 1.01)	0.67 (0.33, 1.35)	-	-	0.106
7	2.46 (1.02, 5.92)*	4.38 (1.88, 10.22)*	0.97 (0.94, 0.99)*	1.00 (0.99, 1.02)	1.00 (0.98, 1.01)	0.66 (0.32, 1.34)	1.33 (0.61, 2.93)	0.65 (0.33, 1.27)	0.113

† * $p<0.05$. The table shows odds ratios with 95% CI in brackets. An odds ratio is significant at 0.05 level if the confidence interval does not contain 1.0.

‡ Marriage, cohabitation, gender, income, and education are binary variables, while age, relationship duration, and cohabitation duration are numeric variables.

was still non-normal. Hence we performed logistic regression with sharing ratio as a binary variable and observed the positive/negative directions of odds ratios. For independent variables, we used marriage, cohabitation, age, gender, relationship duration, cohabitation duration, income, and education. The results are summarized in Table 2.

When marriage or cohabitation is the only predictor, it predicts the ratio of sharing above the median positively and is highly significant. This outcome reaffirms results we obtained from hypothesis tests and is intuitive as married or cohabiting couples are likely to share more accounts than unmarried couples, with more of their life and activities overlapping.

However, neither marriage nor cohabitation is significant when they are both included as predictors. The reason is likely that cohabitation is a confounding factor associated positively with both marriage and sharing. When participants are grouped by marriage and cohabitation, the largest group is who are cohabiting but not married with 72 participants. On the contrary, only one participant is married but not cohabiting. Remaining 101 participants are either married and cohabiting (N=43) or just dating (N=58). This incongruence in cohabitation and marriage is likely due to people's propensity to cohabit before marrying, to experiment the viability of a more committed relationship. Thus, including cohabitation along with marriage in the model decreases the overestimated effect of marriage on sharing.

Marriage and cohabitation are significant with positive odds ratios when age is added as a third predictor in the model, which is also significant but with a negative effect. This is in contrast to our observation that the ratio of sharing is not significantly different across groups above and below the median of age. However, Whitty et al. studied password sharing practices in the UK and also found that younger people have higher chances of sharing passwords. They suggested that a younger population may have more family and friends active online compared to an older population, hence have more opportunities to share accounts [43].

None among relationship duration, cohabitation duration, gender, income, and education is neither significant nor affects the power of marriage, cohabitation, and age in the model. Hypothesis tests have shown that gender, income, and education do not contribute to differences in sharing, but the insignificance of relationship duration and cohabitation duration opposes our previous observations. This result may indicate that sharing of accounts does not undergo drastic changes during a relationship, but occurs at a specific point, e.g., after a couple decides to cohabit or marry. Research on self-disclosure has also shown successful couples often engage in a higher level of interaction earlier in their relationships then exhibit a decline in disclosure after establishing a sufficient level of confidence [5, 19].

4.3 Account Types and Sharing

In our data with 3,421 accounts, 29.8% accounts were shared (N=1,019), and among them, 39.5% were joint accounts (N=402). Figure 3 shows accounts shared by at least ten participants and Figure 4 shows proportions of shared accounts for each category of accounts.

We defined joint account as an account that is set up solely for sharing, owned by both participants and their partners. As we collected data on different types of accounts, we were interested in knowing whether sharing behaviors differ with types of accounts. For example, are some accounts more likely to be shared than other accounts? Also, are people more likely to share a particular type of account when they are earlier/later in their relationships? To answer these

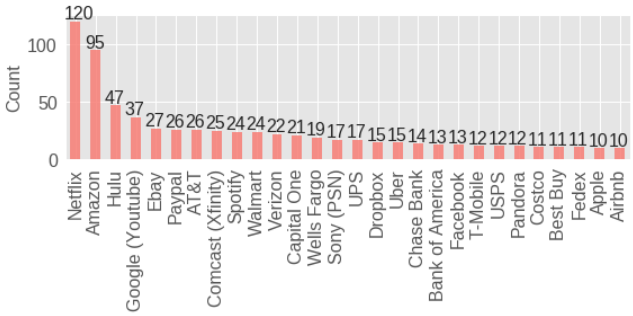


Figure 3: Number of users for accounts shared by more than ten participants.

Figure 4: Proportions of accounts shared by categories of accounts.

questions, we analyzed the effect of relationship duration on different types of accounts. The contingency table summarizing the results of the analysis is in Appendix 4.

We defined new relationships as relationships less than 7 months in their duration, based on an observation that infatuation and fusion, the first stage of a romantic relationship, lasts about 6 months [13]. With this definition, we calculated Chi-square tests of independence and found that people new in relationships share significantly more entertainment accounts ($\chi^2[1,1019]=15.7$, $p<0.0001$), but significantly fewer finance accounts ($\chi^2[1,1019]=7.29$, $p<0.01$). For other types of accounts besides entertainment and finance, we did not find a statistically significant relationship between the stage of relationship and sharing. We also found people who are not new in relationships share more joint accounts with their partners ($\chi^2[1,1019]=15.8$, $p<0.0001$). These results suggest people first share information of less importance such as entertainment accounts before they disclose more private information that carries a higher personal value.

4.4 Taxonomy of Reasons for Sharing

To understand why romantic couples share accounts, we conducted an iterative coding of participants' open-ended responses with 3 of the authors. Initially, 25 reasons for sharing emerged from all the responses, and 6 codes grouping together a set of reasons were identified. Then coders independently coded 50 randomly sampled responses and discussed their rationale. This process was repeated with a new sample until the acceptable level of inter-rater reliability was reached. Once consensus seemed sufficient, we proceeded to code all responses on sharing. Table 3 shows the breakdown of themes and needs with Krippendorff's alphas for each code. The list of reasons for sharing and associated codes are in Appendix 5.

We identified two overarching goals for account sharing from this analysis: *functional* and *emotional*. Specifically, four themes emerged: *convenience* and *household maintenance*, to fulfill a couple's functional needs, and *trust* and *relationship maintenance*, to satisfy their emotional needs. Among the four themes, relationship maintenance and household maintenance contain subcategories: relationship well-being and support within relationship maintenance, and economics and logistics within household maintenance. While convenience, economics, logistics, and trust were observed in previous studies, maintaining relationship well-being and providing support as reasons for account sharing are our novel findings, which we are the first to report according to our knowledge. Note that these categories are not mutually exclusive. Therefore, the total proportion of categories does not add up to 100%.

4.4.1 *Sharing theme 1: Convenience*

In our data, 63.8% of responses mentioned sharing accounts with romantic partners for simplicity and ease of access or usage. They mentioned not wanting the hassle of creating and maintaining a separate account as a reason for sharing. It may also occur by default if two people share a device, and hence, the account on the device. Example comments about sharing because of convenience:

> *"These are common streaming accounts that we share. There is no need for us to have our own accounts when it comes to streaming."*
>
> *"We both use the prime account part of Amazon, and it is easier for both of us to have the email and passwords."*

Unlike [34], we did not see a clear distinction between borrowing and mutual use in our responses. This may be due to cognitive interdependence, a unique characteristic of romantic relationships where individuals in a relationship have greater perceived unity of self and partner [3]. Therefore, sharing of devices and accounts occurs naturally. This is reflected in phrases often used in the responses, such as *"[sharing] just makes sense"* and *"there's no need [to create separate accounts]."*

4.4.2 *Sharing theme 2: Household maintenance*

Household maintenance (85% of responses) refers to sharing accounts in order to complete house-related or financial tasks. House-related chores include running the household and making daily arrangements, e.g., food, clothing, shelter, and travel. We labeled these activities logistics (67.3% of responses). An example of this: *"We choose to share this account because we both use [it] for ebay.com purchases and returns."*

Financial tasks are those that involve currency, such as paying utility bills, managing bank accounts, collecting/using reward points, and managing properties and/or investments. These are grouped as a sub-category named economics (60.1% of stories). Here is an example of economics: *"We both use it [the Amazon account]/share the Prime account to keep costs down ..."*

In our analysis, we found that logistics and economics often overlap (147 comments – 58.6% of logistics comments; 65.6% of economics comments). For example:

Table 3: Taxonomy of 4 themes for why couple share accounts – identified from open-ended questions.

Needs	Themes (Description)	Codes (IRR)	% sharing stories (N=373)
Functional	Convenience (for simplicity and ease of access or usage)	Convenience (0.72)	63.8%
	Household Maintenance (to complete house-related or financial tasks)	Economics (0.79)	60.1%
		Logistics (0.49)	67.3%
Emotional	Trust (to establish trust – intimacy and belief)	Trust (0.75)	45.3%
	Relationship Maintenance (to improve relationship well-being or to provide and receive support)	Relationship Well-Being (0.53)	20.9%
		Support (0.67)	5.6%

> *"We have been married for 7 years so far and have 2 kids. We both need to know what we have in the accounts in order to make purchase and pay bills. It[']s important we have a working knowledge of the money we share."*

Convenience and household maintenance are part of the functional needs to share accounts.

4.4.3 Sharing theme 3: Trust

We characterize sharing out of trust as a statement about intimacy and belief in the partner and the relationship. 45.3% of responses mentioned trust as the reason to share. For instance:

> *"I choose to share for utilities because I trust my partner, and believe both people should have access to them."*

Other variations in expressions of trust include *"... we are in this together," "... because we are married,"* and *"It [sharing accounts] is ... transparent and makes us feel comfortable to know what the other is doing."* This theme is also found in similar works in the past [34, 41].

4.4.4 Sharing theme 4: Relationship maintenance

Relationship maintenance refers to sharing accounts as a measure to improve relationship well-being or to provide and receive support. It accounts for 24.4% of total responses.

Relationship well-being (20.9% of total responses) happens when people actively put in the effort to maintain and improve the quality of a relationship. This often takes the form of sharing activities together. For example: *"[We share accounts] to discuss sports and see highlights of the night before. [We] use [it] for different content also."* Relationship well-being differs from trust in that relationship well-being suggests active effort, while trust is a reflection of the state of a relationship. Another way to differentiate between them is that relationship well-being can be framed as "we-do" statements, e.g., *"we travel together,"* and trust is "we-are" statement, e.g., *"we are in this together."*

The other component of relationship maintenance is support (5.63% of responses), which we defined as the act of receiving and providing help to a partner. An example comment of support:

Provide support: *"I already had a netflix account before we started dating. ... I gave her my password so she could watch when we weren't together."*

Receive support: *"He does not use them but we share them because he knows he can use them and that they exist. I share them because I want him to know about them and have access to them if anything happens to me."*

While relationship well-being is bi-directional (e.g., sharing activities together), support is unidirectional and may be non-reciprocal (e.g., I help my partner without my partner helping me).

4.5 Reasons for Hiding

In contrast to responses on sharing, only 13 responses mentioning active hiding of accounts were collected. We used the same iterative coding procedure from reasons for sharing to code reasons for hiding and found three main reasons for hiding an account: hiding relationships with other people, hiding what could bring up an argument or damage their relationship, and hiding what is irrelevant to the relationship. These three reasons were distributed as 69.2%, 76.9%, and 23.1% in responses. Examples are as follows:

Hiding relationships: *"I just do not want them to see what I post or to see my conversations with other people. "*

Avoid conflict: *"I choose to hide my Peebles[credit card] account because my partner is unaware that I have opened it. She would be angry if she found out I took on another bill when we can barely afford the bills we have."*

Irrelevant to a relationship: *"I don't see a reason for her to know about my Tinder account, I'm sure she has one to[o] but I don't see the point in bringing it up."*

All three reasons for explicit hiding involve a motivation to conceal what a partner may consider wrongdoing [2]. This observation is not surprising considering conventional circumstances where hiding most frequently occurs, such as in illicit liaisons. However, other responses reveal hiding can occur due to reasons that are not necessarily undesirable. Although some of these responses were not marked for active hiding, we find them worthy of mentioning as they reveal neutral or even positive aspects of hiding, as opposed to our intuitions. For example, the following responses demonstrate how hiding occurs to maintain one's personal space:

> *"I choose to hide these accounts by not telling her about it. I choose to do this because [I] want my social media accounts separate and for my own view only."*

"I have a separate gmail account... sometimes, it's okay to have an account that's just yours and yours only..."

As observed in past studies, individual privacy is an essential matter for couples in romantic relationships [29, 38]. Concerning studies on intimate partner abuse and a partner as an "insider threat," above responses put further emphasis on designing technologies that provide better defined personal boundaries [15, 16, 32].

While above responses display conventionally expected motivations for hiding, other responses reveal rare instances where hiding comes from a good-natured motivation:

"I am not hiding anything besides when I am trying to get her a surprise gift. I just try to make sure the browser is closed."

"I don't usually hide my Amazon account but my partner doesn't have the password to it. I do make sure there isn't any e-mails from Amazon if I'm buying a gift for my partner and want it to be a surprise."

As shown, hiding can be employed as a device to strengthen one's relationship by facilitating gift giving. Another response shows hiding can also serve a protective function:

"My spouse spends money badly so I do not want him to spend everything."

Similar to parent-children relationships, adult relationships can involve restrictions intended to promote healthier attitudes that can mutually benefit who are involved in a relationship.

4.6 Sharing Methods

Among the open-ended responses, 49.7% of responses reported methods of sharing. These methods can be categorized under eight general sharing methods, with the most common methods being: 1) keeping the account logged in so it is automatically signed in when needed, 2) storing passwords in a password manager, and 3) sharing/storing the passwords digitally in files or via digital communication, e.g., email. Table 4 shows the eight categories and their frequencies in our responses.

Of concern to us were the 11.8% of the responses that mentioned sharing methods that do not follow general best practices for account security. These included using a familiar or easy password (4.28% of responses), using passwords based on personal information (3.21% of responses), reusing common password-ID combinations (10.7%), and sharing through email. This supports a need to encourage more secure password sharing.

5. DISCUSSION

Our study paints a rich picture of how romantic relationships influence security behaviors and extends the existing knowledge of how individuals approach cybersecurity in social contexts [9, 10, 12, 28, 34, 41]. With the majority of our participants either dating, living with a partner, or married, our data show the array of accounts and behaviors that result from combining lives with another person. We have found it difficult, in coding many of the open-ended responses, to disentangle pragmatic from emotional reasons for sharing behaviors, or even for methods – for instance, is a couple's practice of creating passwords together from personal information more for the ease of memorization, or for the pleasure of memorializing their emotional bond in everyday activities? Often our answer was, "It could be both practical and emotional," which we argue is a complete perspective to bring to security research.

Table 4: Account sharing methods observed in open-ended responses.

Sharing methods	# (%) sharing stories (N=376)
Auto sign-in	58 (31.0%)
Password manager	35 (18.7%)
Electronically stored/shared	31 (16.6%)
Reusing common password/id	20 (10.7%)
Memorizing	17 (9.09%)
Creating credentials together	12 (6.42%)
Writing down on paper	11 (5.88%)
Verbally telling password to partner	6 (3.21%)

At the very least, our data show the need for security designers and engineers to consider socio-cognitive factors when generating ideas for system features, evaluating the usability of security systems, and conducting user evaluations with romantic couples and family households, not just with individuals. Our research has identified four factors motivating online account sharing among couples – relationship maintenance, household maintenance, trust, and convenience – that echo prior works among platonic roommates and other social groups [10, 12, 34, 41]. Security user interface and architecture designers can use these as criteria for evaluating whether the proposed or developed systems or features support usability for those in romantic relationships both as individuals and as a couple. They are also likely to help those in other sharing situations, such as people with disabilities who rely on household helpers for errands or extended families who share resources and logistical burdens such as shopping or banking.

Moreover, decisions about whether and to what extent to share access to accounts and devices with a partner (either by intent or default) are not products of a single moment. They occur in stages and follow the life cycle of the romantic relationship itself. We offer the following observations and suggestions for security design for this relationship lifecycle, broken into three stages: the start of relationship sharing, the maintenance of relationship sharing, and the end of relationship sharing.

5.1 Design Recommendations for Couples

The start of relationship sharing is characterized by individuals starting to grant partner access to some, though not all, of their individually owned accounts and devices. Our data showed that people in the early stage of relationship share significantly more entertainment accounts and fewer finance accounts. Sharing can happen either proactively, e.g., actively sharing passwords, or by default, e.g., watching the same TV.

Sharing at the first stage may be uncertain. In our data, one

participant commented that *"[w]e don't share any [accounts] yet. We're trying to figure that out as our relationship moves on."* We recommend building security features that ease the feeling of uncertainty at the beginning of relationship sharing. For example, allowing multiple PINs or passwords for a single device can segment device accessibility, preserving the access of new romantic partners to some apps while fencing off access to others. Another way to facilitate relationship sharing at this stage is to prompt the account's original owner, on a regular basis, to review his or her current security settings and account sharing status. This can remind users that their accounts are currently being shared and offer options to revoke sharing access if necessary.

Unsurprisingly, our data suggest that couples who have been in a longer-term relationship or who are cohabiting or married tend to share more accounts than those who are in the early stages of a relationship and that they begin to create accounts for joint use. Couples in our study who had been dating longer or who were cohabiting or married indicated sharing more financial accounts, such as individual or joint banking accounts and investment accounts. However, certain accounts remain personal, with participants reporting keeping individual banking accounts and email accounts. Hiding behaviors are likely to occur to preserve privacy and maintain personal spaces.

A design recommendation for this relationship-maintenance stage is to establish a model where multiple users can share one account while user profiles remain independent of each other. Existing services such as Netflix and Hulu allow users to create individual profiles, but this feature is not implemented pervasively. In our data, participants' comments about their practices of account sharing imply benefits they may enjoy if existing services adopted such one-account-multiple-user-profile structure more widely:

> *"The amazon account is automatically signed in. We both use it/share the Prime account to keep costs down and use our own credit cards attached to it."*

This shows there exists a demand for account sharing among the users of services that currently employ one account-one user model. Anecdotally, another example where the current one account-one user model breaks down is two-factor authentication for joint accounts. Authentication information is typically sent to one phone number that is not shared between two people.

Given the popularity of shared account usage and shortcomings in the current implementation of many accounts for couples' needs, it is worth considering a wider range of user configuration options in a one account-multiple-users model, where individuals in a relationship have the freedom to customize their account information and security settings while being able to maintain only one login information. Such account might appear as a single account on the surface, but it would allow each user to maintain his or her personal security settings under the hood, e.g., viewing access to personal information, possibly with an additional layer of identification (e.g., 2FA). It can further help alleviate the "insider threat" of a vengeful or negligent partner being able to sabotage or failing to safeguard account information by limiting access of the partner while still sharing the same account login. Another benefit of this account sharing model is that it can assist its users to monitor for malicious attacks on partners' account, even if it is not actively requested.

Another issue with the current account sharing is that people grant access to their existing individual accounts to their partners. This sharing behavior carries security concerns because login information to individual accounts may contain personal information unique to their original owners. To address this issue, future security systems could make use of machine learning algorithms to identify when users have been sharing access with a romantic partner for an extended period and timely prompt them to review account settings, such as password, viewing permissions, emergency contact, or beneficiary.

A separate aspect of the maintenance of account sharing is safe and secure password sharing. From our data, we noticed many insecure password sharing practices, e.g., reusing passwords for convenience and sharing through email. This poses an opportunity for security researchers to innovate different methods to enable secure sharing of passwords between romantic couples. Equally important may be the need to educate users on secure password sharing protocols.

5.2 Supporting Users in Breakups

Of course, many relationships will not endure forever. At this third stage, individuals are likely to attempt to remove or disable a partner's access to accounts and will need to split up jointly owned property. In our qualitative data, one participant mentioned resetting passwords to all their accounts after breaking up with their ex-partner. Currently, this is a tedious and challenging process and poses security concerns if the user forgets which accounts are shared and which are not. We suggest that the design of account sharing should support users to effortlessly separate their accounts from their partners' and help owners monitor their accounts for ex-partners' login attempts. One design recommendation is to develop login notifications to notify account owners if individuals without sharing access are getting into accounts.

Furthermore, devices in a home network or personal mobile devices should be set, by default, to send notifications to private emails or text accounts about any installation of keyloggers, GPS trackers or other spyware. Accounts should also periodically prompt users to review their security settings. This will trigger owners' memory and help them retrieve access permissions from ex-partners.

Many times, the end of account sharing also triggers account ownership issues, i.e., who should own accounts that are used to be joint accounts? Account sharing features for romantic couples can keep track of the frequency of individual activities and show this information to couples to help them make an ownership decision. Alternatively, an account splitting feature can also help mediate this issue.

In general, sharing between romantic couples is a complicated behavior involving many nuances. While the majority of relationships are fulfilling and desirable, there are many examples of poor relationships, such as "insider threat" and domestic abuse. It is essential to consider these various contexts when designing account sharing and hiding features for romantically involved individuals and how different people will use the features. Supporting couples' practical and emotional needs while maintaining security for each user should

be the cornerstone of designing account sharing features for romantic couples.

6. LIMITATIONS AND FUTURE WORK

Only U.S. residents participated in our study and our findings are not representative of all sharing behaviors. We lacked data on individuals of non-binary gender and non-heterosexual couples and excluded their responses from quantitative analyses. Responses to hiding are tame given their small quantity. A future study may put greater emphasis on hiding and extend its scope to sharing among marginalized groups to amend these issues.

Self-reported responses may have resulted in an inaccurate recall and social desirability bias. The vague wording of the question asking for "active" hiding possibly misled some participants to overlook reporting past behaviors. As our work is exploratory, our design recommendations are non-technical and speculative. In general, our work could benefit from a more thorough exploration of behaviors and their diverse contexts. For example, we did not ask our participants in an open-ended question what types of online accounts they have, and likely have missed some online accounts (e.g., an online account for a municipal library) and associated account sharing behaviors.

Nevertheless, our work opens up an ample room for future works, which may look into: sharing behaviors violating terms and conditions if any, differences between sharing of remote accounts and machine (device) accounts, sharing of phone unlock patterns, and comparison of sharing behaviors between romantic relationships and other close relationships such as family and friends.

7. CONCLUSION

Security design choices often fail to take into account users' social context. Our work is among the first to examine security behavior in romantic relationships. We surveyed 195 people on Amazon Mechanical Turk about their relationship status and account sharing behavior for a cross-section of popular websites and apps. We examined differences in account sharing behavior at different stages in a relationship and for people in different age groups and income levels. We also constructed a taxonomy of sharing motivations and behaviors based on the iterative coding of open-ended responses, many of which are excerpted in this paper.

Based on this taxonomy, we presented design recommendations to support end users in three relationship stages: when they start sharing access with romantic partners; when they are maintaining that sharing; and when they decide to stop. Our findings contribute to the field of usable privacy and security by enhancing our understanding of security and privacy behaviors and needs in intimate social relationships and providing empirical evidence of the need to move beyond a simple one-user-one-account model of security design and system development.

8. ACKNOWLEDGMENTS

We thank Yang Wang who helped us to improve this work as our shepherd. We also thank paper reviewers who contributed valuable feedback and our participants who shared their experiences generously. This material is based upon work supported by the U.S. National Science Foundation under Award No. SaTC-1704087.

9. REFERENCES

[1] A. Adams and M. A. Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999.

[2] W. A. Afifi and J. K. Burgoon. "we never talk about that": A comparison of cross-sex friendships and dating relationships on uncertainty and topic avoidance. *Personal Relationships*, 5(3):255–272, 1998.

[3] C. R. Agnew, P. A. Van Lange, C. E. Rusbult, and C. A. Langston. Cognitive interdependence: Commitment and the mental representation of close relationships. *Journal of personality and social psychology*, 74(4):939, 1998.

[4] I. Altman, A. Vinsel, and B. B. Brown. Dialectic conceptions in social psychology: An application to social penetration and privacy regulation. In *Advances in experimental social psychology*, volume 14, pages 107–160. Elsevier, 1981.

[5] J. H. Berg and M. S. Clark. Differences in social exchange between intimate and other relationships: Gradually evolving or quickly apparent? In *Friendship and social interaction*, pages 101–128. Springer, 1986.

[6] A. J. Berinsky, G. A. Huber, and G. S. Lenz. Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, 20(3):351–368, 2012.

[7] J. Blythe, R. Koppel, and S. W. Smith. Circumvention of security: Good users do bad things. *IEEE Security & Privacy*, 11(5):80–83, 2013.

[8] d. boyd. How parents normalized teen password sharing. `http://www.zephoria.org/thoughts/archives/2012/01/23/how-parents-normalized-teen-password-sharing.html`, 2012. Accessed: 2018-02-14.

[9] S. Das. Social cybersecurity: Reshaping security through an empirical understanding of human social behavior. 2017.

[10] S. Das, T. H.-J. Kim, L. A. Dabbish, and J. I. Hong. The effect of social influence on security sensitivity. In *Proc. SOUPS*, volume 14, 2014.

[11] S. Das, A. D. Kramer, L. A. Dabbish, and J. I. Hong. Increasing security sensitivity with social proof: A large-scale experimental confirmation. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 739–749. ACM, 2014.

[12] S. Das, A. D. Kramer, L. A. Dabbish, and J. I. Hong. The role of social influence in security feature adoption. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1416–1426. ACM, 2015.

[13] P. David. Stages of development in intimate relationships.

[14] S. Egelman, A. Brush, and K. M. Inkpen. Family accounts: A new paradigm for user accounts within the home environment. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 669–678. ACM, 2008.

[15] D. Freed, J. Palmer, D. Minchala, K. Levy, T. Ristenpart, and N. Dell. Digital technologies and intimate partner violence: A qualitative analysis with multiple stakeholders. *PACM: Human-Computer Interaction: Computer-Supported Cooperative Work and Social Computing (CSCW) Vol*, 1, 2017.

[16] D. Freed, J. Palmer, D. Minchala, K. Levy, T. Ristenpart, and N. Dell. "a stalker's paradise": How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 667. ACM, 2018.

[17] K. Greene, V. J. Derlega, and A. Mathews. Self-disclosure in personal relationships. *The Cambridge handbook of personal relationships*, pages 409–427, 2006.

[18] R. E. Grinter, W. K. Edwards, M. W. Newman, and N. Ducheneaut. The work to make a home network work. In *ECSCW 2005*, pages 469–488. Springer, 2005.

[19] R. B. Hays. A longitudinal study of friendship development. *Journal of personality and social psychology*, 48(4):909, 1985.

[20] K. Helkala and T. H. Bakås. National password security survey: Results. In *EISMC*, pages 23–33, 2013.

[21] C. Herley. So long, and no thanks for the externalities: The rational rejection of security advice by users. In *Proceedings of the 2009 workshop on New security paradigms workshop*, pages 133–144. ACM, 2009.

[22] C. Huff and D. Tingley. "who are these people?" evaluating the demographic characteristics and political preferences of mturk survey respondents. *Research & Politics*, 2(3):2053168015604648, 2015.

[23] T. Hunt. The trouble with politicians sharing passwords. `https://www.troyhunt.com/the-trouble-with-politicians-sharing-passwords/`, 2017. Accessed: 2018-02-14.

[24] P. G. Inglesant and M. A. Sasse. The true cost of unusable password policies: Password use in the wild. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 383–392. ACM, 2010.

[25] I. Ion, R. Reeder, and S. Consolvo. "... no one can hack my mind": Comparing expert and non-expert security practices. In *SOUPS*, volume 15, pages 1–20, 2015.

[26] P. G. Ipeirotis. Demographics of mechanical turk. 2010.

[27] R. Kang, S. Brown, L. Dabbish, and S. Kiesler. Privacy attitudes of mechanical turk workers and the us public. In *Symposium on Usable Privacy and Security (SOUPS)*, volume 4, pages 37–49, 2014.

[28] J. Kaye. Self-reported password sharing strategies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2619–2622. ACM, 2011.

[29] A. E. Kelly. *The Psychology of Secrets*. Springer Science & Business Media, 2002.

[30] I. Kirlappos and M. A. Sasse. Fixing security together: Leveraging trust relationships to improve security in organizations. In *Proceedings of the NDSS Symposium 2015*. Internet Society, 2015.

[31] D. Marques, L. Duarte, and L. Carriço. Privacy and secrecy in ubiquitous text messaging. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services companion*, pages 95–100. ACM, 2012.

[32] D. Marques, I. Muslukhov, T. Guerreiro, L. Carriço, and K. Beznosov. Snooping on mobile phones: Prevalence and trends. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, 2016.

[33] W. Mason and S. Suri. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

[34] T. Matthews, K. Liao, A. Turner, M. Berkovich, R. Reeder, and S. Consolvo. She'll just grab any device that's closer: A study of everyday device & account sharing in households. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5921–5932. ACM, 2016.

[35] B. D. Medlin, J. A. Cazier, and D. P. Foulk. Analyzing the vulnerability of us hospitals to social engineering attacks: How many of your employees would share their password? *International Journal of Information Security and Privacy (IJISP)*, 2(3):71–83, 2008.

[36] I. Muslukhov, Y. Boshmaf, C. Kuo, J. Lester, and K. Beznosov. Know your enemy: The risk of unauthorized access in smartphones by insiders. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*, pages 271–280. ACM, 2013.

[37] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. 2010.

[38] S. Petronio. Boundaries of privacy. *State University of New York Press, Albany, NY*, 2002.

[39] E. M. Redmiles, S. Kross, and M. L. Mazurek. How i learned to be secure: Advice sources and personality factors in cybersecurity. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 666–677. ACM, 2016.

[40] M. A. Sasse, S. Brostoff, and D. Weirich. Transforming the 'weakest link'–a human/computer interaction approach to usable and effective security. *BT technology journal*, 19(3):122–131, 2001.

[41] S. Singh, A. Cabraal, C. Demosthenous, G. Astbrink, and M. Furlong. Password sharing: Implications for security design based on social practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 895–904. ACM, 2007.

[42] W. A. Usmani, D. Marques, I. Beschastnikh, K. Beznosov, T. Guerreiro, and L. Carriço. Characterizing social insider attacks on facebook. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3810–3820. ACM, 2017.

[43] M. Whitty, J. Doodson, S. Creese, and D. Hodges. Individual differences in cyber security behaviors: An examination of who is sharing passwords. *Cyberpsychology, Behavior, and Social Networking*, 18(1):3–7, 2015.

APPENDIX

1. Categories and Accounts

Table A1. The list of 17 categories and accounts as presented in the survey, with categories revised for the analysis.

Revised categories	Initial categories	Accounts
Finance	Banking and Real Estate	Chase Bank, Bank of America, Wells Fargo, Capital One, American Express, Discover, U.S. Bank, TD Bank, SunTrust Banks, PNC, Zillow, Realtor, LoopNet, Trulia, Redfin
	Financial Services	Fidelity, Vanguard, American Century Investments, T. Rowe Price, Geico, Charles Schwab Corp., TD Ameritrade, TIAA, Progressive, Allstate, State Farm, Esurance, Metlife, Paypal, Venmo
	Utilities	Comcast (Xfinity), AT&T, Verizon, T-Mobile, Sprint, CenturyLink, MetroPCS, Con-Edison, People's Natural Gas, California Edison, Ameren UE, Georgia Power, National Grid, Eversource Energy, North American Power
Social	SNS, Blogging, and Forum	Facebook, Twitter, Instagram, Linkedin, MySpace, Wordpress, Imgur, Pinterest, Reddit, Tumblr, Snapchat, Blogger (Blogspot), Flickr, Squarespace, 4chan
	Social, Lifestyle, and Art	Meetup.com, Change.org, Patreon, HappyCow, Cohousing.org, Petfinder.com, Jw.org, Lds.org, Flexjobs.com, Skype[2], WhatsApp, Viber, Discord, Telegram, imo.im
	Web Portal (1)	Google - Gmail, Google Drive, etc.[1], Microsoft (MSN) - Outlook Mail, Bing, MS Onedrive, Office.com, etc.[2], Yahoo - Yahoo Mail, Yahoo Answers, etc., AOL - Aol Mail, etc., Apple - iCloud Mail, etc.[3], Easy.com, Lycos, Excite, Craigslist
Entertainment	Video/Music Streaming	Youtube[1], Vimeo, Hulu, Netflix, Soundcloud, Amazon Prime Streaming (Amazon)[4], Spotify, Pandora, Bandcamp, Tidal, Apple Music (iTunes)[3], Directv, Pandora, Google Play[1], iHeartRadio
	Sports, Gaming, and Entertainment	ESPN, MLB.com, NBA.com, NFL.com, Goal.com, Bleacher Report, CBS Sports, Steam, Roblox.com, Battle.net, Xbox.com, Ign.com, League of Legends, Sony Entertainment (Playstation Network), Twitch.tv
Lifestyle	E-Commerce	Amazon[4], Target, Best Buy, Ikea, Macy's, Kohl's, Walmart, The Home Depot, Costco, Staples, Lowe's, Ebay, Etsy, Groupon, Salesforce
	Logistics and Delivery	UPS, Fedex, USPS, DHL, Postmates, Grubhub, Seamless, DoorDash, OnTrac, Blue Apron, GoPuff, Foodler, EatStreet, Instacart, XPO Logistics
	Transportation and Rentals	Uber (UberEATS), Lyft, Uhaul, Penske, Budget, Hertz, Zipcar, Megabus, Greyhound, BoltBus, United Airlines, American Airlines, Delta Airlines, Southwest Airlines, JetBlue
	Fitness and Health	WebMD, Myfitnesspal.com (Under Armour), Mayo Clinic, Drugs.com, Medscape.com, Strava, Prevention.com, Self.com, 24 Hour Fitness, Gold's Gym, American Council on Exercise (Acefitness.org), Freeletics, Freetrainers.com, Peak Pilates, Men's Health
	Leisure and Travel	Booking.com, TripAdvisor, Expedia, Hotels.com, Kayak.com, Marriott.com, Priceline, Hilton.com, easyJet, VRBO, Orbitz, Lonely Planet, Couchsurfing.com, Airbnb, Yelp (Yelp Eat24)
Information & Learning	Creativity and Productivity	Github, Adobe Create Cloud, DeviantArt, unity3d.com, Autodesk.com, Shutterstock, Fanfiction.com, Instructables, MindTools, Framer, VSCO, Epicurious, Allrecipes, Wix.com, Sketch
	Learning and References	Coursera, Duolingo, Codecademy, edX.org, Lynda.com, Khan Academy, Udacity, Stack Overflow, Quora, Wikia, IMDb, MIT Opencourseware, Alison.com, Masterclass.com, Wikipedia
	News and Magazine	CNN, NYTimes, The Guardian, The Washington Post, Forbes, Fox News, Bloomberg, USA Today, The Wall Street Journal, CNBC, Time, The Atlantic, Buzzfeed, Wired, Queerty

	Web Portal (2)	Amazon - Amazon Drive, Amazon Web Services, etc.[4], Oracle - Oracle Cloud Storage Service, etc., Dropbox, Box.com, Mega.nz, SpiderOak
	Dating	OkCupid, Happn, Coffee Meets Bagel, Bumble, Tinder, Down, Lulu, Match.com, Zoosk, Grindr, Hinge, eHarmony, Badoo, PlentyofFish, Ashley Madison

* Web portal was later grouped under two revised categories. Accounts with email features were grouped under social and the rest were grouped under information & learning. Dating was left as a separate category.
[1] Google - Gmail, Google Drive, etc., Youtube, Google Play were coded as Google (Youtube).
[2] Microsoft (MSN) - Outlook Mail, Bing, MS Onedrive, Office.com, etc. and Skype were coded as Microsoft.
[3] Apple - iCloud Mail, etc. and Apple Music (iTunes) were coded as Apple.
[4] Amazon, Amazon Prime Streaming, and Amazon - Amazon Drive, Amazon Web Services, etc. were coded as Amazon.

2. Survey Questions

Note: We only present here questions relevant to the analysis. Questions here are renumbered for presentation, and visual details are removed for concision.

Screening

1. Have you ever been in (or are currently in) a romantic relationship?
☐ Yes, I have been in (or am currently in) a romantic relationship.
☐ No, I have never been in a romantic relationship.

2. Are you currently in an exclusive romantic (dating/marital) relationship?
☐ Yes, I am currently dating someone.
☐ Yes, I am currently married.
☐ No, I am not currently in an exclusive romantic relationship.

3. Have you been in your current relationship for more than a month?
☐ Yes, I have been in my current relationship for more than a month.
☐ No, I have not been in my current relationship for more than a month.

4. If you are not currently in a relationship, did you end your last relationship more than a year ago?
☐ Yes, I broke up from my last relationship more than a year ago.
☐ No, I did not break up from my last relationship more than a year ago.

5. Did your previous relationship last longer than one month?
☐ Yes, my previous relationship lasted longer than one month.
☐ No, my previous relationship did not last longer than one month.

Relationship Details
1. How long have you been in your current relationship? Years ____ Months ____
2. Are you currently living together with your partner? ☐ Yes ☐ No
3. For how long have you been living with your partner? Years ____ Months ____
4. How long did your previous romantic relationship last? Years ____ Months ____
5. Did you live together with your last romantic partner? ☐ Yes ☐ No
6. For how long did you live with your last romantic partner? Years ____ Months ____

Account Usage and Access

Note: There were 17 sections and each section corresponded to a category. A section had two pages, and an introductory paragraph was shown at the beginning of the first page to remind participants about definitions of terms we used throughout the survey. Following questions recurred for each account selected by participants. The part on devices was structured similarly, but we do not explain in detail as it was excluded from the analysis.

1. Do you have any [category] accounts that you commonly use? Choose all accounts that you OR your partner own from the following list. As a reminder…
- By accounts, we mean any website which you use an ID and password to access services or content.
- By sharing, we mean any situation in which you and your partner use a single account/device, either at the same time or taking turns.
- By own, we mean either you own and your partner accesses or that your partner owns and you access. While the questions assume you own the account, you should treat the questions similarly if your partner is the primary owner.
- By joint accounts, we mean any accounts which you and your partner have set up solely for sharing, owned by both you and your partner.
Also, if you have any accounts which you share with or hide from your partner, you will be asked to write few lines to describe why and how you share or hide those accounts.

☐ Account 1	☐ Account 2	☐ Account 3
☐ Account 4	☐ Account 5	☐ Account 6
☐ Account 7	☐ Account 8	☐ Account 9
☐ Account 10	☐ Account 11	☐ Account 12
☐ Account 13	☐ Account 14	☐ Account 15

2. For each account which you selected or entered on the previous page, pick statements those best describe how you and your current/last partner use(d) an account. From each column: 1) choose a statement indicating ownership of an account, 2) choose a statement describing how your partner use(d) an account, and 3) choose a statement about how you use(d) an account.

	Partner regularly uses this account (once a week or more)	Partner sometimes uses this account (once a month)	Partner rarely uses this account (once every few months)	Partner never uses this account
Account	☐	☐	☐	☐

	I am the primary owner of this account	My partner is the primary owner of this account	This account is a joint account	We have separate accounts
Account	☐	☐	☐	☐

	I regularly use this account (once a week or more)	I sometimes use this account (once a month)	I rarely use this account (once every few months)	I never use this account
Account	☐	☐	☐	☐

3. For each account which you selected or entered on the previous page, pick a statement that best describes how your current/last partner access(ed) an account.

	Partner has ready access to this account (e.g. knows password)	Partner can access this account if needed (e.g. can guess password or knows where you store passwords)	Partner doesn't have easy access to this account (i.e., has to ask you, or you login manually)	Partner doesn't know about this account but I'm not actively hiding it	Partner doesn't know and I'm actively hiding the account
Account	☐	☐	☐	☐	☐

4. If you share any [category] accounts with your partner, then could you describe why you choose to share those accounts with your partner, and how you share passwords? (e.g., By using a password manager, by keeping accounts signed in, etc.) You can skip this question if you don't share any accounts.

5. If you are actively hiding any [category] accounts from your partner, then could you describe why you choose to hide those accounts from your partner, and how you hide them? (e.g., By using incognito mode, by deleting the browsing history, by physically hiding the usage, etc.) You can skip this question if you don't hide any accounts.

3. Comparison of Survey Sample and the U.S. Population

Table A2. The comparison of demographic characteristics of the survey sample and the U.S. population.

	U.S. population	Survey sample
N	**249M**	**191**
Age		
18-24	12.4%	7.3%
25-34	17.8%	53.9%
35-44	16.3%	25.7%
45-54	17.1%	9.4%
55-65	16.6%	3.7%
65+	19.7%	0%
Education		
High school or less	39.7%	17.3%
Some college	29%	36.1%
College and more	31.3%	46.6%
Income		
Less than $19,999	45.1%	11.5%
$20,000 to $39,999	24%	34%
$40,000 to $59,999	11.6%	27.7%
More than $60,000	19.3%	26.7%
Gender		
Female	51.3%	43.2%
Male	48.7%	56.8%

* Percentages for the U.S. population were calculated from 2016 American Community Survey (ACS) 1-year Estimates that was released September 14, 2017. 2016 ACS 1-year estimates are based on data collected from January 1, 2016 to December 31, 2016.

4. Differences in Sharing of Entertainment Accounts and Finance Accounts

Table A3. Contingency table for the number of finance accounts and entertainment accounts shared in different stages of a relationship.

Count Total % Col %	Is not new in a relationship	Is new in a relationship	Total
Is not an entertainment account	655 64.3% 66.7%	13 1.28% 35.1%	668 65.6%
Is an entertainment account	327 32.1% 33.3%	24 2.36% 64.9%	351 34.5%
Is not a finance account	704 69.1% 71.7%	34 3.34% 91.9%	738 72.4%
Is a finance account	278 27.3% 28.3%	3 0.29% 8.11%	281 27.6%

Note: The first two rows of the table are comparing the number of entertainment accounts and non-entertainment accounts shared by those who are new in relationships and those who are not. For example, the first column of the first row shows who are not new in relationships share 655 non-entertainment accounts, which constitute 66.7% of accounts they share. Comparing that with 327 in the row below, which is the number of entertainment accounts shared by who are not new in relationships, shows who are not new in relationships share more non-entertainment accounts than entertainment accounts. On the contrary, the second column shows the reversed pattern of entertainment account sharing for who are new in relationships, with 64.9% (24) of accounts shared by them being entertainment accounts, while only 35.1% (13) of accounts shared are not entertainment accounts. Numbers in bottom two rows show that who are new in relationships share more non-finance accounts than who are not new in relationships (91.9% vs. 71.7%), while who are not new in relationships share more finance accounts that who are new in relationships (28.3% vs. 8.11%).

5. Reasons for Sharing

Table A4. List of 25 reasons for sharing with descriptions and associated codes (C=Convenience, E=Economic, L=Logistics, T=Trust, R=Relationship well-being, S=Support).

Reasons (Associated codes)	Description	Example
1 - Joint finance (T, E, C)	Sharing an account because of merged finance	*... We share these accounts to help keep track of our spending. This allows us to budget for the whole month. We both need to know how much money is being spent.*
2 - To keep track of activities (C, L)	Sharing an account to keep track of activities such as spending and order shipments	*We both order and ship stuff and log in to these from time to time to deal with the tracking and history aspects. I keep them logged in but she knows the passwords.*
3 - Similar interest (R, L)	Sharing an account because of shared interests, but not necessarily doing same activities or for same contents	*My [partner]loves those kinds of websites for reading, so everything is in his name. I have access and go on too read it when I want to.*
4 - Simultaneous activities (C, L, R)	Sharing an account to engage in some activity simultaneously	*We both use the Directv information since we like to watch TV together and we can also check on specials this way. He will sometimes access my Google Play if we are going to watch a movie together.*
5 - Shared devices (C, T, E, L)	Couples share an account as they share a device that uses the account	*We keep the accounts signed in on the devices that they are used on. We also know each others passwords to the account should they get signed out.*
6 - Shared friends/family (T, C)	Couples share an account as it lets them connect to shared friends/family	*It's easier when dealing with family and mutual friends to use our joint gmail account. We both keep track of our own passwords.*
7 - Easier management/usage (C, E, L)	Sharing an account to make its management or usage convenient	*Both of our names are on this joint account and it's our main credit card we use. We decided to share and create the password together in order t make it easier to manage account and payments. We use passwords that both of us can remember based on personal information and it is saved in a file.*
8 - For transparency (T)	Sharing an account for transparency/openness	*I share my SNS account passwords with my partner because I don't have anything to hide from him. We are completely open with each other so there isn't any reason why I would not allow him to access my accounts. Since we share the same laptop and I sometimes use his smartphone, I am usually already signed into my accounts so he can access them as well.*
9 - To know what the other is doing (T)	Sharing an account to know what the other is doing	*All passwords for these are often saved on sign in and are accessible by both of us. We choose to share accounts because we share devices and play the same games most of the time. We have the same friends and play buddies so it is much easier for us to manage one account rather than two. It is also transparent and makes us feel comfortable to know what the other is doing.*
10 - Because of relationship/marriage (C, T, E, L)	Sharing an account as it makes their marriage or relationship stronger	*Even though we don't live together, we spend the majority of our time together. It just makes life easier to share these accounts now, since we do plan on marrying in the next year.*
11 - No reason to hide/no sensitive information (T)	Sharing an account as there is no reason to hide, the account contains no sensitive information	*I share the Pacific Gas & Electric and AT&T accounts with my partner because there is NO reason to hide anything ...*

12 - Mutual usage (C, L)	Couples share an account as they mutually use the account (or its contents), but not necessarily at the same time, or for the same purpose	*We share an Uber account so that we can both get around the city. It just makes it easier to have the same account so that it charges to the same card.*
13 - Shared objectives (R, L)	Couples share an account to achieve a mutual goal or purpose	*I like all of our pictures in one place, so I have given my partner the password to Dropbox*
14 - Trust (T, R)	Sharing an account because of trust, or for trust	*I shared my google account password with my partner because trust my partner. My partner know this password and saved in web-browser for easy access.*
15 - Shared business/investments (E, L)	Sharing an account for a shared business, or for shared investments	*WE share it so we can both sell things on it and have a better rating we both know the password for the account*
16 - To help/get help (T, S)	Sharing an account to get help or give help	*Everything is in my name in our marriage, so we share everything. He helps pay bills and helps deposit money so it makes sense for him to have access to all of the accounts. \n\nAlso he likes to make sure I'm not spending too much on my credit cards.*
17 - For emergency (T, S, L)	Sharing an account in preparation for an emergency	*We share these accounts so that either one of us can call if we have problems or questions. We both know the passwords to each account.*
18 - To care for the other (S, L, R)	One shares an account to care for the other	*i share the account so that my fiance could keep up on current events with me and so that he can read funny articles. i share the password by just telling him what it is so that he always has access to it.*
19 - To reduce costs/share benefits (E)	Couples share an account as sharing reduces costs or increases benefits earned from using the account	*I choose to share this account because it would save us a lot of money if we used this individually which makes sense. If we have to put a password in, it is in our little notebook we have to check.*
20 - Living together (C, L)	Couples share an account as they live together	*There is no need for both of us to have a Netflix account since we live in the same house. We stay signed in to this account.*
21 - Because there is a feature that support sharing (C, L)	Sharing an account because it has a feature that supports sharing	*We share the account because we pay for the account together and can have multiple users. There would be no reason to pay for two accounts. We usually just stay signed in on the account on the tv.*
22 - To delegate responsibilities (merged with 16)	Couples share an account to delegate responsibilities besides paying bills when needed	*I share this account with my husband cause sometimes I work late and he needs to order groceries from the app. With my job, I cannot stop and get on my phone to order. I let him know the password when I initially signed up.*
23 - No reason to make a new account (removed)	One sees no reason to make a new account or is reluctant to create a new account	*These are common streaming accounts that we share. There is no need for us to have our own accounts when it comes to streaming. We both know the password and both use these accounts regularly.*
24 - Because sharing was necessary/was asked to do so (removed)	Sharing an account as it was necessary or were asked when creating the account	*In the case of Groupon, I use it far less frequently than my wife and she often forwards me deals that may be of interest to me. Therefore, there is little point in my creating my own account when I can simply use hers. For Costco, we were asked to create a single account when we became Costco members, and it was easy for my wife to remember the username and password.*
25 - Laziness (C, L)	Sharing an account because of laziness/don't want to create a new account	*We share the accounts out of laziness mostly. She uses Ebay and Etsy though, while I don't have any interest in them.*

Characterizing the Use of Browser-Based Blocking Extensions To Prevent Online Tracking

Arunesh Mathur
Princeton University
Princeton, NJ
amathur@cs.princeton.edu

Jessica Vitak
University of Maryland
College Park, MD
jvitak@umd.edu

Arvind Narayanan
Princeton University
Princeton, NJ
arvindn@cs.princeton.edu

Marshini Chetty
Princeton University
Princeton, NJ
marshini@princeton.edu

ABSTRACT

Browser-based blocking extensions such as Ad blockers and Tracker blockers have provisions that allow users to counter online tracking. While prior research has shown that these extensions suffer from several usability issues, we know little about real world blocking extension use, why users choose to adopt these extensions, and how effectively these extensions protect users against online tracking. To study these questions, we conducted two online surveys examining both users and non-users of blocking extensions. We have three main findings. First, we show both users and non-users of these extensions only possess a basic understanding of online tracking, and that participants' mental models only weakly relate with their behavior to adopt these extensions. Second, we find that that each type of blocking extension has a specific primary use associated with it. Finally, we find that users report that extensions only rarely break websites. However when websites break, users only disable their extensions if they trust and are familiar with the website. Based on our findings, we make recommendations for designing better protections against online tracking and outline directions for future work.

1. INTRODUCTION

Online tracking presents numerous privacy risks to users. Third-party trackers present on multiple websites [13] collect sensitive information such as users' personal information, activities, and interests [26] without necessarily alerting users to this type of tracking. Many such third-parties also transmit the information they collect over insecure channels, impeding HTTPS adoption [13, 29]. Given the fact that tracking is on the rise and is often undesirable, users have been advised by numerous agencies, including the Federal Trade Commission (FTC) [14, 9], to take adequate steps to shield their information from such online tracking.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

Users can protect themselves from online tracking by deploying browser-based blocking extensions, which studies [15, 29, 16] have found to be effective to various degrees in blocking third-party trackers. However, while industry surveys [32, 18, 6, 3] have shown that users primarily adopt Ad blocker extensions for user experience (UX) benefits, we lack a comprehensive understanding of how and why users adopt various browser-based blocking extensions in the real world. To improve the privacy protections offered by blocking extensions, we need to better understand users' motivations behind adopting these extensions in the first place, their understanding of the online tracking ecosystem, and whether these extensions work effectively in shielding them against online tracking.

To answer these questions, we conducted two large scale online surveys with current users and non-users of three types of blocking extensions (Ad blockers, Tracker blockers and Content blockers) on Amazon Mechanical Turk (MTurk). We asked three research questions. First, how much do users understand online tracking, and does heightened knowledge about online tracking relate with users adopting such blocking extensions? We investigated this question through the lens of mental models, which prior research has shown influence attitudes and behaviors [20]. Second, do users consciously adopt various blocking extensions to protect themselves from online tracking? Knowing users' intentions can help us understand whether the extensions function to according to users' expectations and if privacy protections are a motivating factor in adoption. Third, when and how do users disable their extensions and accept being tracked? We asked this question because extensions can fail to distinguish between content and trackers, and consequently break websites, potentially forcing users to choose between online tracking protection and accessing content [29].

We have three main findings which both confirm and extend previous work:

1. First, our results show that blocking extension usage only weakly relates with an advanced understanding of online tracking in the real world. Indeed, current blocking extension users were able to better articulate certain aspects of online tracking but these differences were small—despite them having used these extensions for long periods of time. This supports findings from

previous research [37] studying first-time users of these extensions in a lab setting.

2. Second, we report evidence to confirm the expected: most Ad blocker users adopt these extensions primarily to improve their UX on the web and not to block online tracking. On the other hand, tracker blocker users adopt these extensions primarily to block online tracking. However, in an unexpected and new result, we found that most Content blocker users also adopt these extensions primarily to improve their UX on the web as opposed to block online tracking.
3. Third, our results show that current users report that they rarely experience website breakages because of their blocking extensions. However, when blocking extensions break websites, about half of all users disable their extensions so that they can access the content they desire. Their decision to give up tracking protection is based on the perceived value and importance of the content they are obstructed from accessing.

Based on our findings, we make the following recommendations. First, given users' lack of understanding of online tracking, we suggest that system designers should focus their efforts on building systems that automatically enforce tracking protection as opposed to having users take action to protect themselves (such as by installing an extension). We argue that browser vendors can play an important role in facilitating this type of default privacy protection. Second, we suggest that blocking extensions can be further improved by better understanding how website developers embed third-party trackers and deliver content through their websites so that non-use (disabling) is not forced upon users.

2. RELATED WORK

In this section, we touch upon relevant research on online tracking, use of different types of browser-based blocking extensions to prevent online tracking, and studies examining the usability and effectiveness of these extensions.

2.1 Online Tracking

When people visit a website, they interact with a *first* party and often, several *third* parties. The first party is the website or service people visit and intend to use, while third parties are embedded services and trackers that people indirectly and inadvertently interact with. First parties typically include third-party trackers to collect analytics about their customer base, show targeted advertisements, or to include functionality such as social media sharing links [36]. As an example, when someone visits The New York Times (NYT) website[1], the first party is The New York Times—the website that people directly interact with—and one of the third parties—at the time of writing this article—is Google Tag Manager[2], which provides the NYT with analytics about their visitors and marketing support. Another such third-party on the NYT website is Google Publisher Tags[3], which serves the NYT with targeted advertisements—often called Online Behavioral Advertising (OBA)—that are based on peoples' interests, demographics and browsing histories.

[1]https://www.nytimes.com
[2]https://www.google.com/analytics/tag-manager/
[3]https://developers.google.com/doubleclick-gpt/

Extension Studied	Blocking Method
Ad blockers	
AdBlock AdBlock Plus	EasyList, EasyPrivacy (Not default)
Tracker blockers	
Ghostery	Ghostery Blocklist
PrivacyBadger	Heuristics
Disconnect	Disconnect Blocklist
Content blockers	
uBlock uBlock Origin	EasyList, EasyPrivacy, Misc. lists

Table 1: Summary of the browser-based blocking extensions considered in this study.

People do not directly interact with third-party trackers and are often oblivious of their presence yet they are still susceptible to data collection—so this type of tracking is considered privacy violating [26]. For instance, third-party trackers embedded across websites can *see* people visiting those websites, and link these websites visited to reconstruct peoples' browsing histories, which may contain sensitive websites people visited. Further, by just visiting certain websites people can reveal sensitive information including their interests, demographics, as well as the machines and devices they use. In the previous example, both third-parties on the NYT are tracking in nature, and they collect information about people and their activities as people visit websites where the same third-parties are embedded.

Third-party trackers are able to track people by largely employing *stateful* tracking, which involves the use of HTTP cookies to track website visits. However, some trackers have been shown to also engage in more *persistent* and *stateless* tracking techniques such as re-spawning Flash cookies and fingerprinting respectively—both of which can track people even when they clear HTTP cookies [13, 36]. In fact, when Flash cookies were first discovered [42] in 2009, it led to an FTC lawsuit [41].

2.2 Perceptions of Online Tracking

Previous studies [46, 2, 40, 22, 49, 25, 8, 27, 24, 35, 28] have examined peoples' perceptions of, and preferences towards data collection and advertising. For example, one study [22] explored peoples' mental models of how the Internet works, as well as their online privacy and security attitudes and behaviors. The authors found that people with stronger technical backgrounds were able to more clearly articulate privacy and security threats but took no additional steps to protect their privacy and security than people without a technical background. Another study [35] showed that people reported greater concern about data aggregation through third parties than first parties.

One set of these studies examined peoples' perceptions towards online tracking driven OBA. These studies have shown that peoples' attitudes towards OBA are nuanced. First, people find OBA desirable in certain situations (e.g., when a useful product is shown) but not in others (e.g., seeing negative and embarrassing online advertisements) [46, 2]. Second, peoples' attitudes toward OBA depends on how their data is being used [25, 24, 8]—the sensitivity of the data, how long it was retained, the type of advertisements it

was used to deliver, and whether people had the necessary tools to control the advertising if they desired—to target them. Third, peoples' willingness to be tracked varies by the purpose of the tracking [28]—such as OBA, price discrimination, and customization—the entity tracking them (first party vs. third party), and the type of information being tracked (health, financial, or social).

Researchers have also shown that people often have misconceptions about how OBA and online tracking works. First, people have varying mental models about how their data is collected for targeting [49] and this influences their attitudes towards OBA. For instance, people who believed browsers store information used for targeting (e.g., through cookies) were more comfortable with OBA than those who did not; some people in this latter group believed they could use browser settings to clear that information and therefore, restrict OBA. In another instance [46], some people believed they could stop behavioral targeting by using anti-virus software on their machine, or by just using features in their browsers. Finally, researchers have found that people often confuse privacy and security [40], are unsure how tracking works, and therefore cannot adequately protect themselves.

2.3 Blocking Extensions

Currently, people can protect themselves against such tracking by using various browser-based blocking extensions, which take different approaches to block third-party trackers from loading and executing content. Informally, these extensions can broadly be classified into three types: Ad blockers, Tracker blockers, and Content blockers. Table 1 summarizes the extensions we considered in this paper.

2.3.1 Ad blockers

Ad blockers block advertisements from websites. Popular Ad blockers include AdBlock [1] and AdBlock Plus [33]. Both these extensions function using the EasyList [11] list, which contains several patterns corresponding to known advertisements. Each time a user's browser makes a request that matches a pattern in the list, these extensions block that request from loading.

Because Ad blockers block advertisements, they also block third-party advertisers that serve targeted advertisements, such as Google Publisher Tags on the NYT website. However, Ad blockers such as AdBlock and AdBlock Plus fail to block several other non-advertising third-party trackers unless they are specifically configured to do so. Both these Ad blockers can be augmented to block these non-advertising trackers by enabling other lists (e.g., EasyPrivacy [12]).

2.3.2 Tracker blockers

Tracker blockers block third-party trackers more generally, not just those that serve targeted advertisements. Different Tracker blockers take different approaches to blocking trackers. For instance, rather than using the EasyPrivacy ruleset, extensions such as Ghostery [17] and Disconnect [10] use internal lists maintained by the companies that built these extensions, which contain patterns corresponding to tracking services. Each time a user's browser makes a request that matches a pattern in these lists, these extensions block that request from loading. Other Tracker blockers such as PrivacyBadger [34] use heuristics to determine if a third-party is a tracker.

2.3.3 Content blockers

Some blocking extensions aim to function as general-purpose blockers, and block both advertisements and trackers embedded on websites. We call these extensions Content blockers to distinguish these blockers from those described above. Popular Content blockers include uBlock [44] and uBlock Origin [45]. Both these particular blockers have EasyList and EasyPrivacy enabled by default, along with other malware domain lists.

2.4 Effectiveness of Blocking Extensions

Numerous studies have measured the effectiveness and performance of various Ad, Tracker and Content blockers across websites using standard web automation tools [5, 47, 13, 29, 16, 15]. For instance, research by Balebako and colleagues [5] examined the effectiveness of two different privacy tools—Ghostery and Targeted Advertising Cookie Opt-Out (TACO)—in limiting OBA. They tested how the content of online advertisements varied based on the initial profile they were viewed with and when the browser is/is not configured with the extension in question, and found that both types of blocking extensions limit OBA successfully.

Other studies [13, 29, 16, 15] have examined the effectiveness of Ad blockers and Tracker blockers in limiting the number of third-party requests made by websites. These studies collectively found that extensions are effective to varying degrees. For instance, extensions that work with pre-compiled lists such as Ghostery and Disconnect perform better in limiting third-party content than heuristic-based extensions like PrivacyBadger, but overall many extensions miss less prevalent third-party trackers, i.e., trackers found on fewer websites. While these studies show that these extensions are indeed effective in blocking online tracking, they do not examine whether users consciously adopt these extensions to block online tracking, and how effectively these extensions work from a user point-of-view.

2.5 User Studies of Blocking Extensions

Several industry surveys [32, 18, 6, 3] have examined users' motivations behind adopting Ad blocker browser extensions. Collectively, these surveys found that most users adopt these extensions for user experience reasons such as to remove intrusive advertisements and reduce clutter on websites. However, these report findings do not always agree which is why our work examines these topics in more detail. For instance, PageFair [32] found that nearly one third of all their participants used Ad blockers for security benefits, in contrast to global web index [18] and HubSpot [3], which found that nearly one third of users used Ad blockers for privacy benefits, such as to shield their information from advertisers.

Some studies [23, 37] have conducted lab-based usability research on browser-based blocking extensions. First, in a lab study, Leon and colleagues [23] examined whether first-time users could successfully opt-out of or block OBA using AdBlock Plus and Ghostery. They found that users face several problems when dealing with both extensions—including confusing interfaces and technical jargon—that limit their ability to reduce exposure to OBA. Likewise in a lab study, Schaub et al. [37] found that exposing first-time users to Tracker-blocking extensions heightened their awareness of online privacy; however, users found it difficult to fully understand how they were being tracked and what the conse-

quences of being tracked were.

These studies shed important insights into the usability of these extensions, but they either only considered Ad blocker extension users and were not peer reviewed, or only considered a small sample of first-time users interacting with these extensions for the duration of a lab study. In our study, we examine a much larger sample of *real* users of these extensions, who have adopted and currently use these extensions. We also consider a wider variety of extensions including Ad blockers, Tracker blockers, and Content blockers. Further, understanding whether these users' knowledge of these extensions relates with greater use of these extensions in practice, whether users consciously adopt these extensions to protect themselves against online tracking, and how effectively these extensions protect users still remains unclear. In this paper, we examined these questions using both surveys and actual measurements to help determine how we can improve protections against online tracking.

3. METHOD

We conducted two surveys on MTurk. In our surveys, we studied three categories of blocking extensions: Ad blockers, Tracker blockers, and Content blockers, which are listed in Table 1. Through the first survey, we answered two research questions. First, to better understand whether and how users' mental models about online tracking are related to blocking extension adoption, we asked what users and non-users understand about online tracking. Second, to better understand if users are adequately protected from online tracking and to design better tracking protections, we investigated whether whether users consciously adopt these extensions to prevent online tracking. We administered a second survey to all participants from the first survey who reported using at least one blocking extension to answer our third research question: when these extensions break websites, we asked how and whether users decide to disable their extensions, and consequently accept being tracked.

3.1 Survey Design and Deployment

We describe the design of our two surveys below. The study was approved by the Institutional Review Board of our university. The Appendix contains both of our surveys.

3.1.1 Survey One

Questions: The first survey contained four parts and included both open and closed-ended questions. In the first part of the survey, we asked about participants' general Internet behavior. We asked participants how much time they spent online, what services they used, and how many and which Internet connected devices they had access to. In the second part, we gathered participants' general awareness about *Internet/Web tracking*, whether they had heard of this term, who they thought collected information about them as they browsed the Internet, what information they thought was collected, and if they had taken any steps to limit their tracking. In the third part of the survey, we gathered data about the blocking extensions participants had installed on their current browsers. We asked participants whether they had any of the Ad blockers, Tracker blockers or Content blockers listed in Table 1 installed on their current machines, and for each reported blocking extension, we asked who installed it, how long had they been using it, how they learned about it, and why they used it. To collect participants' reasons for adopting their extensions, we used both open and closed-ended responses. Participants first provided their reasons in an open-ended format, after which we asked them to respond to a set of statements (see Appendix A.18.g)—which we borrowed and edited from related work [23]—on a five-point scale ranging from strongly disagree to strongly agree.

Finally, in the fourth part of the survey, we gathered participants' demographic information, including age, gender, education, and profession.

Measurements: In addition to the survey questions, we conducted several measurements of participants' browser configurations and privacy settings to confirm what they self-reported. We checked whether participants' browsers were blocking third-party cookies from being set, blocking third-party trackers, and blocking advertisements.

To measure whether participants' browsers were blocking third-party cookies, we attempted to set and read back a cookie from a different domain than our survey. This domain was also under our control and resolved to a server hosted at our university.

To measure whether participants' browsers were blocking third-party trackers—indicating the presence of an extension that blocked such trackers (such as by using EasyPrivacy)—we added the Google Analytics tracker to the survey and detected whether its JavaScript objects correctly loaded. We chose the Google Analytics tracker for two reasons. First, it is a common tracker, blocked by the extensions we considered, and therefore a good choice to run measurements. Second, we did not want to cause any harm to participants' by exposing their data to possibly nefarious trackers. The Google Analytics account we used for this purpose was password and two-factor protected, and under our control.

To measure whether participants' browsers were blocking advertisements—indicating the presence of an extension that did so—we injected an image wrapped in a HTML div element tagged with a HTML tag found in EasyList into the survey, and checked whether its element loaded.

3.1.2 Survey Two

Questions: We sent survey invites to participants from the first survey who had reported using at least one of the extensions listed in Table 1. This survey asked participants to report their experiences when they had to disable their extensions in order to access content in two particular situations. First, when websites fail to function correctly as a result of users' extensions, and second, when websites ask users to disable their extensions in order to access content (as others have measured [29]). In the first part, we asked participants whether they had experienced website fail to function correctly as a result of their blocking extensions; if they responded yes, we further asked them to list the name and type of the websites(s) they experienced break, and how frequently they experienced such breakages. We then asked participants how they responded in the past after experiencing such breakages, whether they proceeded to attempt to fix the websites, and what if, any steps they took to fix the websites. The second part of the survey closely mirrored the first; instead of the asking about incorrectly functioning websites, we asked users to recollect whether they had

seen Ad-blocking messages that appeared as a result of their blocking extensions. Both parts appeared in random order. In this paper we do not report results from the Ad-blocking messages section of the survey.

3.1.3 Two-Step Survey Design

We designed and launched the surveys in two phases for two reasons. Since survey one asked participants to identify their reasons for adopting blocking extensions, we did not want these reasons to prime them when they were later asked to describe their experiences when disabling their extensions. Second, we were concerned that merging both the phases would make the survey long enough that it would be difficult for participants to complete in one sitting.

3.1.4 Survey Pilot

Before launching the surveys, we conducted a small-scale pilot data collection to ensure the questions were comprehensible and clear. This practice, called cognitive interviews [43], is common in survey design and development. We launched our survey on UserBob[4], a crowd-sourced usability testing website, and invited 10 participants to complete the survey. Participants were asked to "think-aloud" as they completed the survey, specifically highlighting what each question meant to them and what specific information each question was soliciting. Participants captured their screens in a video while taking the survey and thinking-aloud. We used these results to refine and revise our questions. These screen captures lasted for about 20 minutes, and we paid participants $10 each.

3.1.5 Survey Deployment

We used the MTurk platform to recruit participants. We launched the first survey in May 2017, and paid participants $1.00 for completing the survey. We advertised the survey as a "Tell us about your Internet browsing experience" task to mask the survey's purpose and reduce response bias. We required that Turkers be 18 or older, located in the United States (US), and have an approval rating of 95% or higher in order to qualify to take the survey. The survey took between 10-15 minutes to complete.

Three weeks after the first survey, we launched the second survey in June 2017 as a bonus task to all the participants who took the first survey and had been using a blocking extension. We paid participants $2.00 to complete this survey, which took no longer than 10 minutes to complete.

We specifically chose MTurk since its capabilities allowed us to re-target the same participants for the second survey survey. Further, since MTurk participants are known to be more Internet savvy than other Internet users, we were also likely to find a larger pool of blocking extension users compared to other platforms.

3.2 Participants

We recruited 1000 participants from MTurk; participant demographics are summarized in Table 2. Two-thirds (N = 664) of participants from survey one had at least one Ad blocker, Tracker blocker, or Content blocker installed. Nearly half of all participants were aged between 18-34 and the sample was nearly equally split in terms of gender with a slightly higher male participation. Close to two-thirds of the sample had attained a college degree. Finally, the median annual income ranged between $35,000 and $49,999. A logistic regression modeling users vs non-uses of these extensions revealed age ($O.R. = 0.97$, $p < 0.00001$) and gender [Male] ($O.R. = 2.45$, $p < 0.00001$) as significant predictors, indicating that current users were more likely to be younger and male. We sent the follow-up survey invitation to all participants from Survey One, and 480 ($\sim$ 72.3%) subsequently completed Survey Two.

[4]https://userbob.com/

Demographic	All Participants	Extension Users
Age		
18–24	14.0%	17.8%
25–35	45.1%	48.8%
36–45	21.8%	17.6%
46–55	11.0%	9.0%
>55	8.1%	6.9%
Gender		
Male	53.1%	60.7%
Female	46.2%	38.6%
Other	0.7%	0.8%
Education		
No High School	0.2%	0.3%
High School	10.9%	10.2%
Some College	28.8%	28.0%
Bachelor's	37.8%	40.4%
Associate's	12.4%	12.5%
Master's	7.5%	6.6%
Other	2.4%	2.0%

Table 2: Demographic information of the survey participants (N = 1000) and the browser-based blocking extension users (N = 664).

3.3 Data Analysis

For qualitative analyses of open-ended responses, the first author examined the data and first created a codebook. The research team held regular meetings to discuss the initial codes and arrived at the final set of codes after several iterations of discussions and consensus building. We used the finalized codebook to code the open-ended responses. Next, we grouped and organized these codes into themes [38] where applicable. As an example, grouping participants' responses around how tracking took place resulted in codes *use_cookies*, *use_searches*, *use_online_activities*, and *use_clicks* among others. For quantitative analyses, we provide summary statistics, and using Chi-squared tests of proportions, compared sub-populations (users vs. non-users).

4. FINDINGS

In the following section, we summarize our findings from both surveys.

4.1 Blocking Extension Usage

Figure 1 presents the distribution of the blocking extension categories across the participants. Of the 664 participants who reported using at least one blocking extension, Ad blockers were the most prevalent (512 of 664 $\sim$77%), followed by Content blockers (205 of 664 $\sim$31%), and finally, Tracker blockers (84 of 664 $\sim$13%). Users sometimes had one or more blockers, a pattern which was particularly striking in the context of Tracker Blockers: nearly 90% of all Tracker blocker users additionally used either an Ad blocker

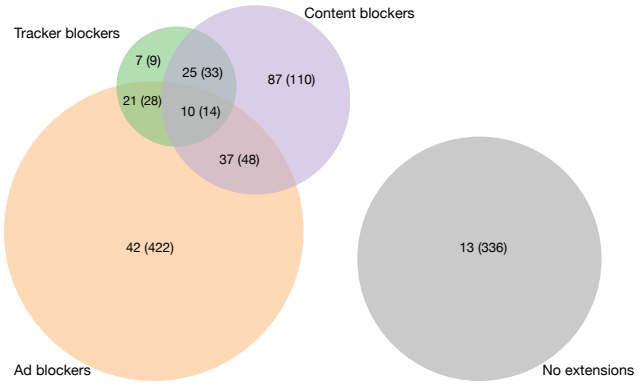


Figure 1: Venn diagram showing the distribution of N = 1000 participants' self reported usage of blocking extensions (within braces) versus those we measured to be blocking third-party trackers (outside braces). For example, only 42 of 422 users who self-reported using *only* Ad blockers were measured to be blocking trackers.

or Content blocker or both.

Using scripts embedded in our survey, we also measured whether participants were blocking third-party trackers and cookies. Across our sample, 9.2% of participants were blocking third-party cookies; a little less than a quarter (242 of 1000) of all participants were blocking third-party trackers. Across extension types, we noted that only about one-fifth of all Ad blocker users (110 of 512 ∼22%), three-quarters of all Tracker blocker users (63 of 84 ∼75%), and three-quarters of all Content blocker users (159 of 205 ∼77%) were blocking third-party trackers.

While our measurements do indicate that most users who reported using these extensions were actually using them, they do not paint a perfect match with the self-reports. We speculate a number of potential reasons for this finding. First, users of Ad blocker extensions such as AdBlock and AdBlock Plus may not have enabled EasyPrivacy, which blocks Google Analytics. Second, users may not have not enabled full protection mode for Ghostery and may not have blocked Google Analytics—the tracker we used to measure tracker blocking. Third, PrivacyBadger does not, by default, block Google Analytics, the tracker we used in our measurements, as it considers it to be a first-party tracker. Fourth, some participants may be using less popular extensions we did not explicitly list. Finally, our measurement script returned incomplete data for certain users due to measurement error: our measurement server was inaccessible momentarily during the survey.

Averaged across the extensions, most users reported learning about these extensions from Internet articles (34.1%) or social media (19.9%). Close to two-thirds (62.5%) of users reported using these extensions on a browser other than the one they took the survey on on their devices, and less than half (40.2%) reported using these extension on a different device than the one they took the survey on, on average. All users had been using them for at least a "A few years" (median across each extension type).

4.2 Mental Models of Online Tracking

To understand participants' mental models of online tracking and whether more developed mental models related with adopting blocking extensions, we analyzed users' (U) and non-users' (NU) mental models together, highlighting instances where these two groups agreed or disagreed. We analyzed the data that emerged from the open-ended question for this section. To compare differences between the groups, we the used chi-square test of proportion. We corrected for multiple testing using the False Discovery Rate method [7], which led to our new significance threshold of 0.025. Table 3 summarizes the themes we list below.

4.2.1 Users & Non-Users Have Like Understanding

Participants' understanding of the online tracking ecosystem could be broken down into four categories: knowing the entities that participated in online tracking, understanding the information that was collected by these entities, recognizing the outcomes of online tracking, and comprehending how online tracking occurred.

Entities that Track. Across our participants, a majority believed advertisers (78.9%) and websites they visited (73.1%) engaged in online tracking. We found no evidence to suggest that the frequency of mention of both entities differed significantly between users and non-users (advertisers: $U = 80.3\%$, $NU = 76.1\%$, $\chi^2 = 2.4$, $p = 0.12$; websites: $U = 74.3\%$, $NU = 70.7\%$, $\chi^2 = 1.5$, $p = 0.23$). This suggests that both users and non-users were equally well-aware of advertisers and websites they visited as entities that tracked them.

Less than 15% of participants mentioned that they were tracked by government agencies ($U = 13.7\%$, $NU = 8.7\%$, $\chi^2 = 5.3$, $p = 0.02$), Internet Service Providers/ISPs ($U = 6.7\%$, $NU = 3.5\%$, $\chi^2 = 4.3$, $p = 0.04$), and third-party companies ($U = 3.9\%$, $NU = 1.1\%$, $\chi^2 = 6.1$, $p = 0.01$). While the frequency of mention of both government agencies and third-party companies differed significantly between users and non-users, these entities were mentioned infrequently by our participants. This suggests that overall far fewer participants were aware of the government, ISPs, and third-party companies as entities that tracked them.

Information Tracked. Only a small fraction of participants (3.7%) did not explicitly list any information that was tracked about them. Well over half of all participants (58.8%) mentioned that basic information was tracked about users, including their demographics, name, sex, email address, location, likes and dislikes, and habits. We found no evidence that users and non-users differed significantly in listing this type of information ($U = 61.2\%$, $NU = 56.3\%$, $\chi^2 = 2.2$, $p = 0.14$), suggesting that both groups were aware that information about them could be tracked.

More than half the participants (54.8%) felt that information about users' online activities such as websites visited, time spent on websites, products looked at and clicked on, search and purchase histories was tracked. We found no evidence that current users and non-users differed significantly in mentioning this type of information ($U = 55.9\%$, $NU = 53.6\%$, $\chi^2 = 0.48$, $p = 0.49$), suggesting that both groups were mostly aware that information about their activities could be tracked.

Themes	Total (%)	Users (%)	Non-Users (%)	Difference (%)	p-value
Entities that Track					
Advertisers	78.9	80.3	76.1	4.2	0.12
Websites Visited	73.1	74.3	70.7	3.6	0.23
Government Agencies	12.0	13.7	8.7	5.0	**0.02**
Internet Service Providers	5.6	6.7	3.5	3.2	0.04
Third-Party Companies	3.0	3.9	1.1	2.8	**0.01**
Information Tracked					
User Attribute Information	59.6	61.2	56.3	4.9	0.14
Behavioral Activities	55.1	55.9	53.6	2.3	0.49
Device Information	26.1	32.9	12.6	20.3	**<0.0001**
Outcomes of Tracking					
Visible Outcomes	44.9	46.7	41.2	5.5	0.10
Invisible Outcomes	23.9	33.2	5.5	27.7	**<0.0001**
Tracking Mechanisms					
Through Activities	56.1	57.7	52.9	4.8	0.4
Through Cookies	23.9	29.7	12.3	17.4	**<0.0001**

Table 3: Summary of the themes that emerged from participants' mental models of online tracking broken down by users and non-users. Bolded p-values are significant at the 0.025 level.

Approximately a quarter (26.1%) of all participants mentioned that information about Internet users' devices, such as their browser name and version, and IP address was tracked. However, current users mentioned this information significantly more often than non-users (U = 32.9%, NU = 12.6%, χ^2 = 47.6, $p < 0.0001$). This suggests that blocking extension users were more aware than non-users about the information that was tracked about their devices. Overall, over half of all participants were aware that tracking occurs but a significant number of participants did not know that online activities and devices could be tracked.

Outcomes of Tracking. A little more than half of all participants (57.4%) were aware of at least one outcome resulting from online tracking. Participants described both "visible" and "invisible" outcomes as others have previously classified [28]. Visible outcomes included those that users could observe in their browsing experience (e.g., targeted advertising). Invisible outcomes included those that users could not directly observe (e.g., price discrimination).

More specifically, less than half of all participants (∼44%) cited visible outcomes of online tracking such as targeted advertisements, customization of websites, and deciding what to sell to users. We found no evidence that current users and non-users differed in how frequently they brought up this outcome (U = 46.7%, NU = 41.2%, χ^2 = 2.7, p = 0.10). This suggests than while both groups were equally aware of tracking outcomes they could directly observe, the majority of participants did not even recognize visible outcomes of tracking as tracking-related.

Even fewer participants (19.4%) reported invisible outcomes of online tracking, including companies maximizing their revenue, offering varying prices, and collecting personally identifiable information. Blocking extension users brought up this outcome significantly more often than non-users (U = 33.2%, NU = 5.5%, χ^2 = 94.0, $p < 0.0001$). This suggests that extension users were more aware of outcomes of online tracking they could not directly observe than non-users. Still, only close to a third of blocking extension users and less than one-fifth of all participants reported knowing these outcomes. Overall, most participants in our study were not able to easily recognize signs of online tracking.

Tracking Mechanisms. Participants varied in how they believed tracking worked. Slightly more than half the participants believed that online tracking occurred on websites through their activities on the websites, the products and advertisements they clicked on, or their search and product history. We found no evidence to suggest that this belief varied significantly between current users and non-users (U = 57.7%, NU = 52.9%, χ^2 = 0.7, p = 0.4). This suggests that both groups were aware that their activities on websites could be tracked.

A smaller fraction of participants (25.5%) stated that cookies were the underlying mechanism through which tracking occurred, and this number varied significantly between current users and non-users. In particular, current users mentioned cookies three times more than non-users (U = 29.7%, NU = 12.3%, χ^2 = 37.2, $p < 0.0001$). This suggests that users were more aware than non-users that cookies can be the underlying mechanism through which tracking works; however only about one-third of users mentioned this overall. The majority of our participants were aware that tracking could occur by collecting information about online activities but three quarters of all participants were not aware that cookies could be used for tracking.

4.2.2 *Comfort with Tracking Depends on Context*

We examined both users' and non-users' responses with respect to how comfortable they were with their data being collected on the Internet. Confirming results from previous work on users' and attitudes towards data collection [46, 2, 28, 39], we found that participants' level of comfort was context dependent: both current users and non-users described situations where they were comfortable and uncomfortable with data collection. The majority of all users were not comfortable with tracking in general. A little over half users (55.4%) and a little less than half non-users (45.9%) were uncomfortable with their data being collected, harboring a general mistrust toward companies that collect data about them, and wanting to keep their information and activities

private. These participants often expressed apathy, saying that data collection was hard to stop, and that if companies really wanted their data, they could acquire it in different ways. These numbers differed significantly between users and non-users ($\chi^2 = 8.1$, $p = 0.005$).

By contrast, a little over a quarter of users (28.5%) compared to more than one-third non-users (36.4%) were comfortable with their data being collected ($\chi^2 = 6.5$, $p = 0.011$). Both sets of participants cited several reasons for being comfortable with tracking such as when the online tracking resulted in positive gains, such as receiving special deals through targeted advertising. For others, tracking was acceptable because they had nothing to hide, and that they believed online services needed users data in order to offer services and function for free.

To summarize, we found that most participants—regardless of whether they used a blocking extension—had only a basic understanding and awareness of online tracking. Our findings support and extend findings from prior work in lab settings that users may know a little, but not significantly more about online tracking after using a browser-based extension [37, 23]. We show that fewer participants were aware of entities that tracked them other than the ones they could explicitly see provide visible modification to content. Across both users and non-users, there existed some differences: users were slightly more able to articulate what data about users' devices is collected, the invisible outcomes of tracking, and how cookies are used in tracking than non-users. However, these differences were spread across only a third of the sample of extension users in each case, indicating that despite these differences, extension users did not present elevated knowledge and understanding about online tracking even after using these extensions for many years.

4.3 Why Use Blocking Extensions?

We examined whether users consciously adopted blocking extensions to block third-party trackers. In the survey, we solicited participants' reasons behind adopting their extensions both in the form of open and closed responses. To analyze the close responses, we binned the Likert scale measurements into agree, not sure, and disagree bins. We compared the open and close ended responses and noted any similarities and differences. We found that current users' responses from the open responses could be grouped into three primary reasons for extension adoption: user experience improvements, security, and privacy—similar to the options we offered them to select from the closed responses.

4.3.1 UX Reasons Drive Ad, Content Blocker Users

In the open responses, the most common reason users cited for adopting Ad blockers and Content blockers was to improve their user experience when browsing the Internet, with the latter finding being unexpected. Close to 89% of participants who used Ad blockers and 84% of participants who used Content blockers said they were motivated by user experience improvements. On the other hand, only a small fraction of users (11.9%) reported using Tracker blockers for user experience improvements. Current extension users' elaborated three main reasons:

Reducing clutter. Nearly half of all current users (50.5%) reported using blocking extensions to block the clutter on webpages. For instance, participant P716, an AdBlock Plus user, stated: *"I hate advertisements that affect my ability to navigate a page without distraction, so I choose to block them in order to have a faster, more streamlined experience."* Often for such users, the extensions were a means to help them block advertising content that obstructed them from viewing desired content on a website.

Blocking Pop-ups. Two-fifths of all current users (40.2%) reported using these extensions to specifically block advertisements that appeared as pop-ups on webpages, which users considered intrusive in nature. For instance, participant P900, an AdBlock Plus user, said: *"The popup advertisements interfere with my online experience. They are annoying and slow down my computer. AdBlock Plus allows me to circumvent unsolicited advertisements."*

Speedup Loading Times. Finally, one-third of all current users (33.1%) reported they used these extensions to speed up the loading of websites, which consequently help them conserve their data and bandwidth. For instance, participant P458, an uBlock Origin user, commented: *"[I use it] to prevent the 100s of advertisements that appear when browsing sites. So many advertisements play or are shown that it slows down browsing performance and uses more bandwidth."*

In agreement with the open responses, ∼95% of both Ad blocker and Content blocker users reported using these extensions for user experience reasons in the close ended responses. We also noticed an additional (∼65%) Tracker blocker users reported using their extensions for user experience reasons.

4.3.2 Privacy Reasons Drive Tracker Blocker Users

Looking at the open responses, 76% of Tracker blocker users said they primarily used the extensions to protect their information from third-parties and advertisers. Participants were concerned that advertisements networks and data mining companies on the Internet collected their data, tracked their browsing history, and showed them targeted advertisements. They believed that they could, using these extensions, block companies that engaged in such practices. For instance, participant P899, a Ghostery user, stated: *"I use Ghostery so advertisers and sites will not track my information or collect info using cookies."* On the other hand, only a small fraction of participants who used Ad blockers (7%) and Content blockers (10%) used them for privacy reasons. In agreement with the open responses, ∼90% of Tracker blocker users reported using these extensions for privacy reasons in the close ended responses. We also noticed an additional Ad blocker (∼76%) and Content blocker (∼71%) users reported using their extensions for the same privacy reasons.

4.3.3 Fewer Security Reasons Across Extensions

From the open responses, only ∼10% of participants—across Ad blocker, Tracker blocker, and Content blocker users—stated they used these extensions for security reasons. Those who did use these extensions for security noted they used it in order to prevent harm to their devices from malicious advertisements and scripts online. For instance, participant P450, an AdBlock user, elaborated: *"I use Adblock because of all the EXCESSIVE advertisements/popups that end up causing me to click on something that I'm not wanting to click on and then a pop-up comes up alerting me that my computer has a Virus, telling me to call some number. Let's just say those people really irritate me."* On in-

specting the close responses, this number increased. We noticed additional users across all extensions—Ad blockers (56%), Tracker blockers (39%), and Content blockers (62%)—reported using their extensions for the same security reasons.

Overall, we noted participants associated each extension type with a primary and secondary reason for adoption, which emerged from the open and close ended responses respectively. That is, users may have mentioned their primary reasons for using the extensions as opposed to including secondary reasons in the open ended responses. Even though users may be aware of other benefits from these extensions, their primary motivation is more focused: Ad blockers and Content blockers primarily for user experience gains, and Tracker blockers primarily for privacy reasons.

4.4 Dealing With Broken Websites

We specifically studied users' experiences when blocking extensions broke the functionality and appearance of websites, as other studies have tried to capture using instrumented measurements [29]. We examined specific changes users reported about their interface and browsing activity, how frequently they experiences these breakages, and users' decision making with respect to disabling their extensions.

4.4.1 Users Report Limited Breakages

Only about two-fifths (180/480) of participants who took the second survey had experienced at least one website that failed to function correctly because of their browser extensions. The majority (94.6%) of those who reported broken website experiences observed them *rarely* or *sometimes* in the span of any given week. Participants reported the following experiences with their extensions in decreasing order of prevalence:

1. Webpages failed to load completely and the content failed to appear (28.7%)
2. Embedded videos failed to play (24.3%)
3. Webpages appeared distorted, and the elements looked out of place (13%)
4. Pop-ups that drove functionality failed to appear (8.1%)
5. Images failed to load completely (7.5%)

Overall, users' self-reported website breakages were lower than expected, which suggests that the blocking extensions were largely effective in distinguishing between trackers and content. However, given that websites failing to appear completely, and videos failing to play, were amongst the most commonly cited website breakages suggests that these extensions often confused trackers and Content Distribution Networks [29].

4.4.2 Content and Trust Drive Disabling Decisions

When websites failed to function correctly, nearly half the users (91/180) who experienced such breakages stated that they never attempted to fix and access the website when they experienced them break, and instead ignored and went on to find alternate content. The other half (89/180) who did access the content on such websites—either sometimes or always—by disabling their extensions based their decisions on the following criteria:

Value of Content. Users who stated they sometimes or always attempted to access the content of such websites, based their judgment on the uniqueness and importance of the content they intended to view; that is, could they gain access to the same content elsewhere? Participant P107 best illustrates this point: *"It depends if I really want to access the content, but I usually just navigate away."*. This suggests breakages can certainly dissuade users from using certain sites if the content is not perceived as unique.

Trust in Website. Similarly, users who stated they sometimes or always attempted to access the content of such websites, reported accessing content if they "trusted" the website and if it was familiar to them; that is, had they accessed it before? Participant P282 explained: *"If it's a site I trust, and understand why they need access to cookies, JavaScript, etc. I will attempt to relax the permissions so the site will work. Otherwise I look for an alternative site (and there's almost always an alternative!)."* This suggests that less popular websites which cause breakages can lose content consumers if blocking extensions do not interact well with their websites.

Overall, most participants reported only limited breakages in the span of a given week, indicating that these blocking extensions largely work effectively from the user point of view. However, when websites did break, nearly half the users attempted to fix the websites by disabling their extensions—and therefore gave up their protection—and based their decisions on how much they valued the content on and the trust they had in the website.

5. DISCUSSION

In this section, we discuss the broader implications of our findings, and outline directions for future work.

5.1 Reducing Privacy Protection Burden

First, our results show that despite having some knowledge about online tracking and how it worked, participants remained mostly uninformed. Having a browser extension did not significantly relate with having a more developed mental model of online tracking. Having adopted these extensions, users remained protected from online tracking to the degree supported by the extensions in their default modes. While these defaults were largely configured correctly for Content blockers and Tracker blockers, they were less so for the largest extension category in our dataset: Ad blockers. Indeed, we saw that only about 10% of all Ad blocker users had enabled EasyPrivacy, which continued to remain disabled by default.

Therefore, we suggest that asking users to take action to protect their privacy may be a sub-optimal suggestion. Instead, an alternate proposal for enhanced privacy protection is to pull users out of the equation completely, and design systems that protect users automatically. Echoing the call of others [22, 31], we suggest that browser designers could more successfully protect users from online tracking through defaults (e.g., by restricting third parties' access to user data), rather than requiring users' to take proactive, intentional steps such as adopting a browser-based blocking extension. In fact, several browser vendors have moved in this direction recently. For example, Mozilla recently incorporated online tracking protection into their private browsing mode, meaning that users who switched to private browsing would be protected from third-party tracking [30]. Apple took this

a step further and implemented intelligent tracking restrictions in Safari 11 [4], where they restricted the lifetime of cookies set by third-party trackers and advertisers, thereby restricting how much data these trackers can collect about users. Future work could examine privacy enhancements that browsers can implement such as contextual situations—e.g. webpages where sensitive information is entered—where third-party trackers should explicitly blocked.

5.2 Reducing Blocking Extension Failure

Second, our results point out that browser-based blocking extensions work largely effectively from a user perspective. When websites did break, users noticed that embedded videos failed to play, or parts of the website failed to load completely. Future work could examine how well users' self-reports of website breakages match with actual website breakages in the wild. Doing so could help determine ways in which extensions can better support feedback from users to improve protection coverage. Out of the extensions we examined in this study, only Ghostery and PrivacyBadger currently collect any feedback at all.

When website breakages occur, users are required to disable their extensions and accept the trackers embedded on the website. Our study reveals that users only disable their blocking extensions when the content they attempted to access is valuable, or if they are familiar with and trust the website (e.g., from a previous engagement). To ensure that users are protected against online tracking—and that non-use is not forced upon them—requires building more efficient blocking tools. For instance, recent approaches to using machine learning to discriminate between JavaScript-based content serving and tracking content has been explored with high accuracy [19]. Improving the status-quo can also be achieved through a broader conversation between the various stakeholders including extension developers and publishers of websites. We encourage the SOUPS and broader privacy community to further investigate how publishers embed content and use third-party services, and the steps that can be taken to design better solutions that do not force users to disable their extensions.

6. LIMITATIONS AND FUTURE WORK

Out study is not without limitations. First, we used Mechanical Turk for data collection, and therefore findings are not generalizable to the full population of Internet users. Recent research has shown that adult Turkers in the U.S. have more privacy concerns than the regular adult US population [21]. Therefore, it is likely that the number of users of these extensions in the general population are much lower. Future research could examine the external validity of these findings in greater detail.

Second, we examined the results in the context of self-reported extension usage by users, but also measured extension usage to ensure users were actually using these extensions; while these measures were mostly in agreement, there were occasions where users reported certain extensions but we did not detect them. However, overall, users have been shown to be able to accurately self-report more deliberate actions, including external browser extension usage [48].

7. CONCLUSION

We studied real world use of blocking extensions to learn how to improve user protections against online tracking. Our results show that Ad blockers and Content blockers are more widely used than Tracker blockers. Furthermore, both users and non-users have limited mental models of online tracking, that they mostly adopt blocking extensions to improve their user experience, and that when extensions break websites, users disable the extensions based on how important the content they are accessing is to them. Based on our findings, we make recommendations to improve blocking tools and provide enhanced privacy by improved extension defaults to better protect users from online tracking.

8. REFERENCES

[1] AdBlock. Adblock. `https://getadblock.com`, 2017.

[2] L. Agarwal, N. Shrivastava, S. Jaiswal, and S. Panjwani. Do not embarrass: Re-examining user concerns for online tracking and advertising. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, pages 8:1–8:13, New York, NY, USA, 2013. ACM.

[3] M. An. Why people block ads (and what it means for marketers and advertisers). `https://research.hubspot.com/why-people-block-ads\-and-what-it-means-for-marketers-and\-advertisers`, 2016.

[4] Apple. Apple. `https://webkit.org/blog/7675/intelligent-tracking-prevention/`, 2017.

[5] R. Balebako, P. Leon, R. Shay, B. Ur, Y. Wang, and L. Cranor. Measuring the effectiveness of privacy tools for limiting behavioral advertising. 2012.

[6] M. Bauman. Six surprising findings about ad block users. `https://www.forbes.com/sites/forbesagencycouncil/2017/07/11/six-surprising-findings-about-ad-block-users`, 2017.

[7] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[8] F. Chanchary and S. Chiasson. User perceptions of sharing, advertising, and tracking. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 53–67, Ottawa, 2015. USENIX Association.

[9] ConsumerReports. Want to protect against websites that spy on you? get an ad blocker. `https://www.consumerreports.org/digital-security/to-protect-against-websites-that-spy-on-you\-get-an-adblocker/`, 2018.

[10] Disconnect. Disconnect. `https://disconnect.me/`, 2017.

[11] EasyList. Easylist. `https://easylist.to/easylist/easylist.txt`, 2017.

[12] EasyPrivacy. Easyprivacy. `https://easylist.to/easylist/easyprivacy.txt`, 2017.

[13] S. Englehardt and A. 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.

[14] FTC. Online tracking. `https://www.consumer.ftc.gov/articles/0042-online-tracking`, 2018.

[15] K. Garimella, O. Kostakis, and M. Mathioudakis. Ad-blocking: A study on performance, privacy and counter-measures. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, pages 259–262, New York, NY, USA, 2017. ACM.

[16] A. Gervais, A. Filios, V. Lenders, and S. Capkun. Quantifying web adblocker privacy. *IACR Cryptology ePrint Archive*, 2016:900, 2016.

[17] Ghostery. Ghostery. `https://www.ghostery.com/`, 2017.

[18] GlobalWebIndex. The state of mobile ad-blocking. `https://www.globalwebindex.com/reports/mobile-ad-blocking-2017`, 2017.

[19] M. Ikram, H. J. Asghar, M. A. Kaafar, A. Mahanti, and B. Krishnamurthy. Towards seamless tracking-free web: Improved detection of trackers via one-class learning. *Proceedings on Privacy Enhancing Technologies*, 2017(1):79–99, 2017.

[20] P. N. Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010.

[21] R. Kang, S. Brown, L. Dabbish, and S. Kiesler. Privacy attitudes of mechanical turk workers and the u.s. public. pages 37–49. USENIX Association, Submitted.

[22] R. Kang, L. Dabbish, N. Fruchter, and S. Kiesler. "my data just goes everywhere:" user mental models of the internet and implications for privacy and security. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 39–52, Ottawa, 2015. USENIX Association.

[23] P. Leon, B. Ur, R. Shay, Y. Wang, R. Balebako, and L. Cranor. Why johnny can't opt out: A usability evaluation of tools to limit online behavioral advertising. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 589–598, New York, NY, USA, 2012. ACM.

[24] P. G. Leon, A. Rao, F. Schaub, A. Marsh, L. F. Cranor, and N. Sadeh. Privacy and behavioral advertising: Towards meeting users' preferences. In *Symposium on usable privacy and security (SOUPS)*, 2015.

[25] P. G. Leon, B. Ur, Y. Wang, M. Sleeper, R. Balebako, R. Shay, L. Bauer, M. Christodorescu, and L. F. 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*, SOUPS '13, pages 7:1–7:12, New York, NY, USA, 2013. ACM.

[26] J. R. Mayer and J. C. Mitchell. Third-party web tracking: Policy and technology. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 413–427. IEEE, 2012.

[27] A. McDonald and L. F. Cranor. Beliefs and behaviors: Internet users' understanding of behavioral advertising. 2010.

[28] W. Melicher, M. Sharif, J. Tan, L. Bauer, M. Christodorescu, and P. G. Leon. (do not) track me sometimes: Users' contextual preferences for web tracking. *Proceedings on Privacy Enhancing Technologies*, 2016(2):135–154, 2016.

[29] 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, NJ, USA, April 2017. IEEE.

[30] Mozilla. Mozilla. `https://support.mozilla.org/en-US/kb/tracking-protection-pbm`, 2017.

[31] A. Narayanan and D. Reisman. The princeton web transparency and accountability project. In *Transparent Data Mining for Big and Small Data*, pages 45–67. Springer, 2017.

[32] PageFair. The state of the blocked web. `https://pagefair.com/downloads/2017/01/PageFair-2017-Adblock-Report.pdf`, 2017.

[33] A. Plus. Adblock plus. `https://adblockplus.org`, 2017.

[34] PrivacyBadger. Privacybadger. `https://www.eff.org/privacybadger`, 2017.

[35] E. Rader. Awareness of behavioral tracking and information privacy concern in facebook and google. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 51–67, Menlo Park, CA, 2014. USENIX Association.

[36] F. Roesner, T. Kohno, and D. 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.

[37] F. Schaub, A. Marella, P. Kalvani, B. Ur, C. Pan, E. Forney, and L. F. Cranor. Watching them watching me: Browser extensions' impact on user privacy awareness and concern. In *NDSS Workshop on Usable Security*, 2016.

[38] I. Seidman. *Interviewing as qualitative research: A guide for researchers in education and the social sciences*. Teachers college press, 2013.

[39] F. Shih, I. Liccardi, and D. Weitzner. Privacy tipping points in smartphones privacy preferences. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 807–816, New York, NY, USA, 2015. ACM.

[40] F. Shirazi and M. Volkamer. What deters jane from preventing identification and tracking on the web? In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, WPES '14, pages 107–116, New York, NY, USA, 2014. ACM.

[41] R. Singel. Online tracking firm settles suit over undeletable cookies. `https://www.wired.com/2010/12/zombie-cookie-settlement/`, 2010.

[42] A. Soltani, S. Canty, Q. Mayo, L. Thomas, and C. J. Hoofnagle. Flash cookies and privacy. In *AAAI spring symposium: intelligent information privacy management*, volume 2010, pages 158–163, 2010.

[43] S. Sudman, N. M. Bradburn, and N. Schwarz. *Thinking about answers: The application of cognitive processes to survey methodology*. Jossey-Bass, 1996.

[44] uBlock. ublock. `https://www.ublock.org/`, 2017.

[45] uBlock Origin. ublock origin. `https://github.com/gorhill/uBlock/`, 2017.

[46] B. Ur, P. G. Leon, L. F. Cranor, R. Shay, and Y. Wang. Smart, useful, scary, creepy: Perceptions of

online behavioral advertising. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, pages 4:1–4:15, New York, NY, USA, 2012. ACM.

[47] R. J. Walls, E. D. Kilmer, N. Lageman, and P. D. McDaniel. Measuring the impact and perception of acceptable advertisements. In *Proceedings of the 2015 Internet Measurement Conference*, IMC '15, pages 107–120, New York, NY, USA, 2015. ACM.

[48] R. Wash, E. Rader, and C. 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.

[49] Y. Yao, D. Lo Re, and Y. Wang. Folk models of online behavioral advertising. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1957–1969, New York, NY, USA, 2017. ACM.

APPENDIX

A. SURVEY ONE

1. How many hours on average do you spend using the Internet each day?
   (a) Less than 1 hour
   (b) 1 - 3 hours
   (c) 4 - 6 hours
   (d) 7 - 9 hours
   (e) More than 9 hours
2. How many Internet connected devices do you own or have access to?
3. Please check all the types of Internet connected devices you own or have access to.
   (a) Personal computers (e.g., desktops, laptops)
   (b) Mobile devices (e.g., smartphones, tablets)
   (c) Activity trackers (e.g., Fitbit)
   (d) "Smart" home-appliances (e.g., Internet connected TV, Refrigerator)
   (e) Other - Write In (Required)
   (f) None of the above
4. Which of the following statements best describe the device you are using to complete this survey.
   (a) Regularly used only by me
   (b) Regularly used by multiple workers at a place of employment
   (c) Regularly used by multiple members of a family
   (d) Regularly used by multiple members who are not members of one family
   (e) Regularly used by many people in a public place (library, Internet cafe, etc.)
   (f) Other - Write In (Required)
5. Do you generally use this device to complete HITs on Mechanical Turk? [Yes / No]
6. Have you heard of the term "Internet/Web tracking"? [Yes / No]
7. (If Yes) In your own words, please describe what "Internet/Web tracking" means to you.
8. (If Yes) In your own words, please describe what comes to your mind when you hear the term "Internet/Web tracking".
9. Please check all the entities that you think collect your information as you browse the Internet.
   (a) The Website you are visiting
   (b) Advertisers and sponsors
   (c) Third-party companies
   (d) Government agencies
   (e) Internet Service Providers
   (f) Browser creators (e.g., Google, Mozilla)
   (g) Other - Write In (Required)
10. In your own words, please list the information you think the entities you checked above collect as you browse the Internet.
11. In your own words, please describe the purposes for which you think the information you listed above is collected.
12. In general, how do you feel about your information being collected as you browse the Internet.
    (a) Extremely Uncomfortable
    (b) Somewhat Uncomfortable
    (c) Not Sure
    (d) Somewhat Comfortable
    (e) Extremely Comfortable
13. In your own words, please explain the reason behind your answer to the above question.
14. Have you taken any steps to prevent your information from being collected as you browse the Internet? [Yes / No / I don't remember]
15. (If Yes) In your own words, please describe the steps you have taken to prevent your information from being collected as you browse the Internet.
16. (If Yes) How confident are you that the steps you describe above prevent your information from being collected?
    (a) Not at all Confident
    (b) Slightly Confident
    (c) Somewhat Confident
    (d) Very Confident
    (e) Extremely Confident
17. Do you use any of the following browser extensions on your current browser?
    (a) AdBlock
    (b) AdBlock Plus
    (c) Ghostery
    (d) PrivacyBadger
    (e) uBlock
    (f) uBlock Origin
    (g) Disconnect
    (h) None of the above
18. For each selected extension (E):

(a) Who installed each of the following browser extensions on your current browser? (Grid)
   i. I installed it myself
   ii. Someone else installed it for me
   iii. I don't remember

(b) How did you learn about extension E?
   i. Friends
   ii. Family
   iii. Social Media
   iv. News
   v. Extension's Website
   vi. Internet Articles
   vii. Other - Write In (Required)
   viii. I don't remember

(c) For how long have you been using each of the following browser extensions? (Grid)
   i. A few days
   ii. A few weeks
   iii. A few months
   iv. A few years
   v. Many years
   vi. I don't remember

(d) Please check all the statements that best describe where you use extension E:
   i. I also use E on a different browser(s) on this device
   ii. I also use E on another device
   iii. Other - Write In (Required)
   iv. None of the above

(e) In your own words, please describe why you use E.

(f) In your own words, please describe how you think E works.

(g) Please state how much each of the following statements indicate your reasons for using E (Strongly Disagree - Strongly Agree):
   i. I use extension *E* in order to block unwanted content.
   ii. I use extension *E* because I do not like seeing advertisements.
   iii. I use extension *E* in order to speed-up the loading of websites.
   iv. I use extension *E* to prevent websites from serving viruses through advertisements.
   v. I use extension *E* because I am concerned websites that I visit collect, share or sell my information to other companies.
   vi. I use extension *E* to prevent online advertising companies from delivering advertisements that are tailored specifically to me.

19. What is your age?
20. What is your annual household income?
    (a) Less than $25,000
    (b) $25,000 to $34,999
    (c) $35,000 to $49,999
    (d) $50,000 to $74,999
    (e) $75,000 to $99,999
    (f) $100,000 to $124,999
    (g) $125,000 to $149,999
    (h) $150,000 or more
    (i) Prefer not to answer
21. What is the highest education level you have completed?
    (a) No High School
    (b) High School Graduate
    (c) Some College
    (d) Bachelor's Degree
    (e) Associate's Degree
    (f) Master's Degree
    (g) Doctoral Degree
    (h) Professional Degree (e.g., MBA, J.D.)
    (i) Prefer not to answer
22. What gender do you most closely identify with?
    (a) Male
    (b) Female
    (c) Other
    (d) Prefer not to answer

B. SURVEY TWO

1. Certain websites "break" or fail to function correctly because of web browser extensions and add-ons such as Ad blockers and Tracker blockers. In the past, has any website(s) failed to function correctly for you as a result of your AdBlocker or Tracker blocker? [Yes / No / I don't remember]
2. (If Yes) In your own words, please describe what functionality or feature of the website(s) failed to function correctly, and list the website(s) on which you experienced this problem.
3. (If Yes) In any given week, how often do you come across websites that fail to function correctly as a result of your AdBlocker or Tracker blocker?
   (a) Never
   (b) Rarely
   (c) Sometimes
   (d) Often
   (e) Always
4. (If Yes) Which of the following best describe the actions you take after you experience a website that fails to function correctly as a result of your Ad blocker or Tracker blocker?
   (a) I ignore the website
   (b) I sometimes attempt to fix the website
   (c) I always attempt to fix the website
5. (If Yes) In your own words, please describe the reason behind your answer to the above question.
6. (If "I sometimes attempt to fix the website" or "I always attempt to fix the website") In your own words, please describe the steps you take to fix the website(s) that fail to function correctly as a result of your Ad blocker or Tracker blocker.
7. (If "I sometimes attempt to fix the website" or "I always attempt to fix the website") In your own words, please describe why you take the steps you describe above.

1. Certain websites detect whether users are running Ad blockers and present them with a message requesting them to disable the Ad blockers in order to continue using the website. In the past, have you come across such messages? [Yes / No / I don't remember]
2. (If Yes) In your own words, please describe the message(s) you observed and list the website(s) you observed these messages on.
3. (If Yes) In any given week, how often do you see messages requesting you to disable your Ad blocker?
   (a) Never
   (b) Rarely
   (c) Sometimes
   (d) Often
   (e) Always
4. (If Yes) Which of the following best describe the action you take after seeing one of these Ad-blocking messages?
   (a) I never proceed to access the content on such websites
   (b) I sometimes proceed to access the content on such websites
   (c) I always proceed to access the content on such websites
5. (If Yes) In your own words, please describe the reason behind your answer to the above question.
6. (If "I sometimes proceed to access the content on such websites" or "I always proceed to access the content on such websites") In your own words, please describe all the steps you take to access the content on websites that ask you to disable your Ad blocker.
7. In your own words, please describe why you take the steps you describe above.

Can Digital Face-Morphs Influence Attitudes and Online Behaviors?

Eyal Peer
Graduate School of Business Administration, Bar-Ilan University, Israel
eyal.peer@biu.ac.il

Sonam Samat
Heinz College, Carnegie Mellon University
Pittsburgh, PA, USA
sonamsamat@gmail.com

Alessandro Acquisti
Heinz College, Carnegie Mellon University
Pittsburgh, PA, USA
acquisti@andrew.cmu.edu

ABSTRACT
Self-images are among the most prevalent forms of content shared on social media streams. Face-morphs are images digitally created by combining facial pictures of different individuals. In the case of self-morphs, a person's own picture is combined with that of another individual. Prior research has shown that even when individuals do not recognize themselves in self-morphs, they tend to trust self-morphed faces more, and judge them more favorably. Thus, self-morphs may be used online as covert forms of targeted marketing – for instance, using consumers' pictures from social media streams to create self-morphs, and inserting the resulting self-morphs in promotional campaigns targeted at those consumers. The usage of this type of personal data for highly targeted influence without individuals' awareness, and the type of opaque effect such artifacts may have on individuals' attitudes and behaviors, raise potential issues of consumer privacy and autonomy. However, no research to date has examined the feasibility of using self-morphs for such applications. Research on self-morphs has focused on artificial laboratory settings, raising questions regarding the practical, in-the-wild applicability of reported self-morph effects. In three experiments, we examine whether self-morphs could affect individuals' attitudes or even promote products/services, using a combination of experimental designs and dependent variables. Across the experiments, we test both designs and variables that had been used in previous research in this area and new ones that had not. Questioning prior research, however, we find no evidence that end-users react more positively to self-morphs than control-morphs composed of unfamiliar facial pictures in either attitudes or actual behaviors.

1. INTRODUCTION

Face composites, or face-morphs, consist of facial images merged together to produce a new, realistic-looking image of a person that contains some of the elements of the comprising facial images [11]. A substantial body of work has shown that individuals sometimes fail to consciously recognize themselves in face composites that contain their own picture [4], but tend to prefer such self-morphs, trusting them more and finding them more attractive [12,13] when compared to morphs that do not contain the individual's own facial image. Facial images are commonly used in advertising (e.g., of models or celebrities), and if morphs are effective in influencing end-users' attitudes and behavior, that could have far reaching implications for marketers [28] but also for consumer privacy. Consider a marketer who has access to a consumer's Facebook profile. That marketer may use a picture of that consumer in an ad for a product. Such use of the consumer's picture may be deemed unethical (or even appalling), and would probably not promote the marketer's goals. However, what if the marketer instead used the consumer's picture to create a digital morph that combined that picture with an unknown face? The consumer might not consciously recognize this self-morph. The morph, however, could still evoke strong and positive emotional responses in the consumer, due to the familiar elements it contains. How would consumers react to this implicit, visceral mode of persuasion? Social media users make many types of personal information publicly available [2]. Firms use that information to learn more about potential customers and target advertisements accordingly [32], sometimes influencing end-users [20] without their explicit consent or awareness - a form of hidden "digital market manipulation" [7]. Leveraging individuals' innate attraction to self-morphs to promote products is an example of a targeted marketing strategy [10] that may influence end-users' actions while operating outside their awareness, raising potential yet significant privacy concerns.

Existing research has examined the impact of celebrity morphs on consumers' behavior [28], but not the potential impact of self-morphs as a covert and visceral forms of targeted marketing. Moreover, research on self-morphs has been limited to artificial laboratory settings, raising questions about the generalizability and applicability of the reported effects. We explore the uncharted territory of the impact of self-morphs on consumers' behavior in settings that more closely model real-life conditions. Unlike prior studies (that relied on taking photos of subjects in a lab, thus raising awareness among subjects about the purposes of the experiments), we examine whether self-morphs could be created using individuals' personal information from their social network profiles, and then used without subjects' awareness. Furthermore, unlike prior studies (that focused on participants' attitudes towards facial morphs, including trust) we examine to what extent self-morphs can affect also behavioral intentions, including purchase intentions. Our work thus ties into the privacy literature in two ways. First, it highlights how, due to the vast self-dissemination of personal information, public yet personal data can be used in interactions with consumers by both services and independent third parties surreptitiously – that is, without the former's awareness. Second, it highlights potential limits on individual autonomy [27] in decision making by examining the effectiveness of technologies that may covertly influence consumer decision making based on their own data – a form of "visceral targeting," so to say.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018. August 12 -- 14, 2018, Baltimore, MD, USA.

In three online and lab experiments, we tested whether self-morphs impact consumers' attitudes and online behavior. We relied on real world data (images posted on social networking sites by experiment participants) and used various dependent variables and a combination of techniques and study designs. Unlike prior research, we found no differences in how consumers judge, or react to, a self-morph vs. a control-morph composed of unfamiliar facial pictures in both realistic settings and in settings that follow previous lab experiments, as well as both when focusing on attitudinal variables and when focusing on behavioral intentions. Indeed, across the experiments, we ended up testing both 1) new designs and variables that had not been used in previous research in this area (as we initially focused on extending prior work) and 2) designs and variables that had already been reported (when we switched to replicating prior research). Our research endeavor did not begin as a replication attempt: building upon the apparent robustness of results in existing literature (see Section 1.2), at the onset we focused on capturing behavioral metrics (such as behavioral intentions in Study 1, and actual self-disclosures in Study 2, as opposed to the attitudinal variables more common in previous studies) to measure to what extents face-morphs derived from social media could affect actual online behaviors. As both initial pilots and main experiments failed to find such an effect, we traced back our efforts to attempt to replicate existing results in the literature, without finding significant results. Nevertheless, such null results are worthy to be reported through the scientific community for several reasons. First, null findings (when backed by appropriate methodologies) can be important and enlightening [17], especially in light of the recent evidence of non-replicability of major findings revealed in many scientific areas [23]. Second, the results suggest that, if self-morphs have any effect on people's judgments and behavior in the lab, that effect may not robustly extend to other settings.

1.2 Related Work

Facial images are an exceptional type of perceptual stimuli. Evidence from neuroscience, in support of the 'face-specificity' hypothesis, suggests that the brain has specialized cognitive and neural mechanisms dedicated to face perception [19]. Further evidence suggests that the brain implicitly and automatically evaluates faces, thus enabling individuals to make social judgments about unfamiliar individuals from facial properties alone [11,14,31]. With the advent of computer graphics, face-morphing technology has made it possible to alter the familiarity of faces. Morphing a familiar face into an unfamiliar one creates a composite that has familiar features but may still be unrecognizable as a whole. Previous research has shown that participants fail to consciously recognize themselves in face composites created by morphing their own face with an unknown face [12,4]. This happens when the unknown face contributes a larger proportion of the composite (e.g., 60%) while the self-face contributes a smaller proportion (e.g., 40%). Despite this lack of conscious recognition, participants tend to prefer self-face composites. Other researchers have studied face composites created with a family member's or a friend's face [6,24].

As has been noted, a substantial amount of prior work has studied the effect of self-morphs on individuals' attitudes. DeBruine found that participants tend to trust self-morphs more than non self-morphs [12]. DeBruine also studied the attractiveness of self vs. non-self-morphs and found that participants find self-morphs more attractive [13]. Bailenson et al., created composites of participants with electoral candidates (with the participant's face contributing the smaller proportion and the candidate's face contributing the larger proportion) and found that participants report higher intentions to vote for self-like candidates than for non-self like candidates [4]. Tanner and Maeng morphed Tiger Woods's face with a stock model's face. They collected data on willingness to buy from this composite versus a control composite before and at the peak of the famous Tiger Woods scandal [28]. They found a significant decline in reported levels of willingness to buy from the Tiger-morph after the scandal. These results have been explained through a "familiarity based valence accessibility" account. This hypothesis assumes that implicit recognition of a familiar individual in a morphed face is sufficient to enable an underlying (and pre-existing) valence judgment of the familiar individual to be automatically perceived [28].

2. THE CURRENT RESEARCH

Although various studies have examined the effects of face composites on various dependent variables, a number of unrequited issues require additional research. From a methodological perspective, most (if not all) of the previous studies have used pictures that were explicitly solicited from the participants – thus, participants may have been (perhaps subconsciously) aware of the research questions or objectives, making a demand effect possible. Furthermore, most previous research on self-morphs used artificial lab environments, as is customary and warranted for basic cognitive and perceptual psychological research. However, the use of such strict settings limits the generalizability of the research findings to actual real-world scenarios, reducing the potential implications of these findings for the HCI and privacy communities, as well as for every day users of online technologies.

In our research, we focused on more realistic and privacy-sensitive settings: we used pictures taken from participants' online social network profiles (specifically, from their Facebook profiles), without their explicit *ex ante* knowledge or awareness (while still ensuring proper experimental consent; all studies were conducted with IRB approval of our institution, and all studies secured inform consent of participants) in order to rule out the possibility that previous findings were, to some degree, confounded by expectation effects. Furthermore, this novel use of pictures from online social networks data also allows us to focus our examination to domains that are of interest to human-computer interactions. Namely, while previous research on self-morphs focuses on people's judgments and attitudes such as trust (e.g. [12,13]), in our studies we mostly focused on behaviors that directly pertain to online consumer behavior (such as purchasing intentions and self-disclosure behaviors) and highlight how consumers' personal data may be not merely accessed, but also used, in manners that are hard for end-users to predict or prevent.

In their 2009 staff report, the Federal Trade Commission (FTC) defines the term 'behavioral advertising' as "the tracking of a consumer's online activities over time – including the searches the consumer has conducted, the web pages visited, and the content viewed – in order to deliver advertising targeted to the individual consumer's interests" [16]. The industry greatly favors the use of such targeted ads because, in comparison to non-targeted ads, targeted ads generate higher click-through rates [15] and higher sales [5]. While personalization of ads can benefit consumers by exposing them to relevant products, the extensive collection and use of personal information also raises consumers' privacy risks and concerns. In fact, consumer surveys about perceptions of targeted advertising suggest that, by and large, people do not like being tracked and do not wish to receive targeted ads [29,21,25]. In this paper, we investigate the effect of individuals' facial images,

in the form of self-morphs, on online consumer behaviors (such as purchasing intentions and self-disclosure behaviors).

2.1 Overview of Studies

The design of our studies builds upon prior research on self-morphs. The set of studies covers an array of experimental setups, participants' pools, and dependent variables. Two studies (Studies 1-2) were conducted online, using pictures obtained from the participants' online social network (Facebook) profiles (thus, we used these pictures without explicit, ex ante participants' awareness, in order to ensure that any observed effects could only be attributed to the implicit exposure to self-morphs)**;** one replication study (Study 3), instead, was conducted in a lab, using photos captured in the lab at the onset of the experiment. For technical reasons (explained further below), Study 1 only included Caucasian males and Study 2 focused on Caucasian females. Study 3 included participants from both genders. Study 1 focused on purchasing or hiring intentions (in addition to measures traditionally captured in morph studies, such as perceived trustworthiness); Study 2 focused on self-disclosure– a variable common in online privacy research, but novel in the context of morph studies; in Study 3, we only focused on replicating previous studies' results using a trustworthiness dependent variable. Studies 1 and 2 were conducted online; although more ecologically appropriate for testing online visceral marketing strategies, Studies 1 and 2 relied on a two-step design (discussed at length below), and therefore required significant per-participant recruitment and retention efforts; in Study 3, we conducted a large-scale laboratory experiment with a larger sample and higher power.

2.2 Morph Preparation

All three studies (and the pilots we ran to test our experimental infrastructure) relied on a two-step design: in a first phase, participants' facial images were collected (either from their publicly available Facebook profiles in Studies 1-2, or by taking a photo of them in the lab in Study 3). The second phase took place either several weeks after (Studies 1-2) or a few minutes after (Study 3) the first phase. Before phase 2, we created morphed images for the experimental and control conditions for each participant using Abrosoft's FantaMorph (www.fantamorph.com, see examples in Figures 1 and 2). During the second phase, morphed images were shown to the subjects as part of the studies' respective experimental designs. In the rest of this section, we describe the process through which we collected images for making the morphs.

Participants in the online studies were invited on Amazon Mechanical Turk to take part in a survey about Facebook activity. The survey took less than 5 minutes and participants were paid 50 cents for their participation. The survey included various questions about Facebook (such as how often and for what purposes participants use Facebook) to establish the study's legitimacy. The last question in the survey was the question of interest to us: participants were told that we were interested in collecting data from their Facebook profile in order to validate whether they would be eligible for future studies, and for this reason we asked them to provide a link to their Facebook profile page. We assured participants that we would only collect publicly available data, and that this question was optional – participants were informed that they could skip the question and still receive full payment. This enabled us to get access to Facebook profiles of our MTurk participants and collect their publicly shared facial images. These images were then used to create morphs to be used in the second phase. The morph-creation process replicated the methodology used in prior research published in this area (specifically, [4]; see also Sections 2.3 and 5.1). Using this approach, we surveyed over 10,000 participants from MTurk and about 50% of them gave us links to their Facebook profiles. About 20% of those had publicly shared facial images which could be used in morphs (images that are well illuminated, good resolution, and where the participant's face is front-facing with neutral expression). These participants comprised our sampling population from which we recruited participants for the second phase of Studies 1 and 2, taking into account participants' ethnicity and gender (which they reported in the first phase survey).

Study 3 followed a similar two-step approach. However, Study 3's participants were invited to a lab, where their photo was taken and used to make morphs that were immediately shown to them. As noted, we also conducted two online pilots to test and hone our technical and experimental two-step procedure.

2.3 A Note on Replication

Our research endeavor did not begin as a replication attempt. Initially we focused on capturing the impact of face-morphs on new dependent variables that had not been the focus of prior research (Studies 1 and 2 and their pilots). We attempted to replicate existing results on previously used dependent variables (in Study 3) only after failing to find effects for our behavioral dependent variables. That noted, across all three studies presented here, we did try to follow as closely as possible all the technical steps in designing face-morphs and in presenting them to participants. While exact replication of methods was made harder by the fact that not all previously published papers comprehensively disclosed their methods, and not all authors were responsive to our requests for their materials, we were able to follow most closely the method used in [4], whose authors were the most responsive to our questions regarding their experimental material. The authors of [4] were responsive to questions and shared with us details of their morphing software (Magic Morph). Furthermore, given the large sample size used in their study, their clear description of the methodology employed, and the magnitude of effects reported, [4] seemed like one of most rigorous approaches and methods to follow. Thus, our morphing strategy was based on [4], although we utilized a different morphing software, Fanta Morph (after [4] was published we found in rounds of tests that new software Fanta Morph produced more realistic morphs). We also followed [4] in criteria for picking suitable images (for instance, images where participants were not wearing glasses).

3. STUDY 1

The scenario and setting chosen for this study was searching and hiring a private instructor online. We aimed to explore whether instructors whose images would be made of self-morphs would be regarded more favorably, giving them an advantage in the hiring process. This scenario illustrates one of the many ways sophisticated online entities could exploit individual's self-images that are publicly available on social network websites.

3. 1 Method

A review of the previous studies on the effects of face composites revealed that the effect sizes (Cohen's d) ranged from 0.39 [28] to 0.70 SDs [6]. Based on that, we estimated an effect size of about 0.4 SD and aimed at a sample that would provide about 80% power to detect such an effect. We were able to recruit 118 Caucasian males (M_{age} = 28.3, SD = 7.5) through our pool of MTurk participants, which completed the study for $1.5 in an average duration of about five minutes. This sample had a power of about

71% to detect the estimated effect size with a two-sided test, or 82% power for a one-sided test (i.e., to show that self-morphs are more attractive than control-morphs).

After reading and agreeing to the consent form, participants were given a list of musical instruments (e.g., guitar, violin, piano, etc.) and were asked to choose one instrument they would most like to learn to play. Then, participants were asked to imagine they are looking to hire an instructor who can teach them how to play the instrument they have chosen. They were then shown two images of two private instructors that they, supposedly, found in their online searches. Instructors were called "A" and "B", both were reported to have had 10 years of experience in playing this musical instrument and both reportedly charged $10 for a lesson. Participants were asked to indicate which instructor (A or B) they would personally choose to hire. One of the instructors' images (randomly selected) was a self-morph, while the other was a morph of two unfamiliar persons. One of these two unfamiliar persons was a randomly selected other participant's face (used at 40% in the morph) and the other face was a second spokesperson, different from the spokesperson used in the self-morph (unfamiliar to participants; used at 60% in the morph). An example is given in Figure 1. We used two spokespersons because using the same spokesperson would result in two very similar looking morphs where the differences would entirely be because of the 40% face used in making the morph. This could prompt the participants to specifically look for subtle differences between the faces and perhaps interpret the goal of the study. We randomly varied whether the self-morph was created with one stock-model or the other and whether it appeared on the left or the right in a split-panel.

For participants randomly assigned to the treatment (or "self") condition, the morph was created by combining the participant's face (obtained from publicly shared images on his Facebook profile) with the stock model's face. The participants randomly assigned to the control (or "other") condition viewed the same ads, but the face shown to them was a morph created by combining a randomly chosen other participant's face from among the participants in the treatment condition with the same stock model's face. This procedure ensured that the participants in the control condition viewed (in aggregate) the same images as the participants in the treatment (i.e., self-morph) condition, and that the only difference between the conditions was that for participants in the treatment condition the morphed image included their own face, whereas for participants in the control condition it did not. Allocation to treatment vs. control condition was done before participants started the study, using a computerized randomizer that assigned each invited participant to either be in the control or treatment condition.

Afterwards, participants were asked to rate each instructor (separately) on how trustworthy, attractive and knowledgeable he seemed to them, how much they liked him, how similar they thought he was to themselves and how strongly they identified with him. They also indicated if they found anything strange or unusual, or familiar, in either of the instructors' images, and if they said they did, then we asked them to elaborate further (six participants said they recognized themselves in the image and were thus dropped from the analysis). In the next section of the study, participants were presented with five facial images, two of them had instructor A and B, and were asked to identify who was instructor A, and who was B. This was a check question included to ensure that participants remembered which face was which, because the questions on whether participants found anything unusual or familiar with the morph were asked on a screen that did not show participants the morphs. Then, participants completed the Narcissistic Personality Inventory (NPI, [3]), which we included to examine whether the impact of a self-morph could be restricted to people who hold a higher, more self-loving, perception of themselves.

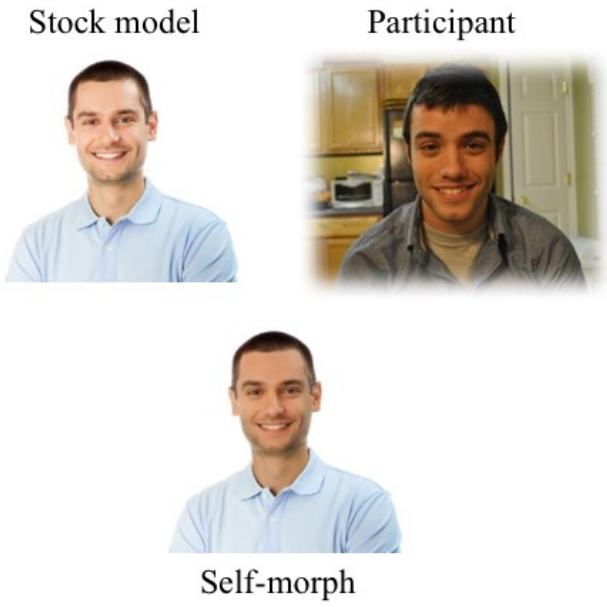


Figure 1. Example stimuli used in Study 1.

Lastly, participants completed the demographic questions and were fully debriefed. In this and in all the following studies, participants were explained that the study was connected to a Facebook survey they had previously completed, and that the researchers may have used publicly available pictures from their Facebook profile for the purposes of the current study. We explained that we did so only for research purposes and that all images collected for this study were kept secure and confidential. We also asked participants to refrain from sharing the details of this study with anyone else in any manner, until the study is completed. Participants' responses to all questions in all studies (including responses to sensitive questions) were always kept separate from their personal or identifiable information. In all studies, participants were given contact details of the researchers and the IRB, and could also to leave comments, concerns or complaints in the survey form itself. Moreover, we provided participants with the option to withdraw their responses from the survey by clicking on a link to a withdrawal form. All participants were thanked and paid, regardless of their final consent. (The option to withdraw responses after being debriefed was provided in all studies, but only two participants chose to use it.) Our IRB approved the procedure of this and all studies reported in the paper. Experimental materials used in Study 1 (as well as Studies 2 and 3) can be found online at https://www.heinz.cmu.edu/~acquisti/SOUPS2018/Study1-2-3.pdf.

3.2 Results and Discussion

Even though our pilot tests confirmed that the two stock-model images we used in creating the morphs were perceived to be equally attractive, in this study the morphs created with these images were not perceived to be equivalent. One was hired more often than the

other (60% vs. 40%; $p = 0.029$). Still, we examined the percent of participants who chose to hire the two different instructors (self vs. other) and found no difference in proportions: 49% chose to hire the self-morph whereas 51% for the other ($p = 0.84$).

Whether people had taken lessons in the instrument before had a significant effect on their decision to hire their own self-morph. Thirty-one participants had taken lessons to learn the musical instrument before, and they were significantly more likely to hire the self-morph (67.74%, $p = 0.048$). However, this result does not hold after we account for multiple comparisons. Eighty participants had never taken lessons to learn the musical instrument before and there were no significant differences there (45%, $p = 0.3173$). One person reported to be currently taking classes to learn the musical instrument. Paired t-tests on participants' ratings of self vs. other morph for trustworthiness, attractiveness, knowledgeable, liking, identifying with self and similarity to self were not statistically significant, as detailed in Table 1. Although the measures showed high internal reliability (Cronbach's alpha = 0.827) overall mean judgments were also not statistically different between the conditions (see Table 1). There were also no significant differences between the percent of participants who found self and other morphs familiar (9% vs. 6.3%) or between the percent of participants who found self and other morphs unusual looking (24% vs. 25%), $p > 0.9$. We ran mixed model analyses of NPI on all DVs. None of the interactions was significant ($p > 0.14$ before correcting for multiple comparisons, $p > 0.7$ after). (Descriptive statistics are presented in Table 1.)

To summarize, we did not find evidence for an effect of a self-morph on individuals' attitudes, intentions or judgments. Given the sampling constraints, the study was fairly powered, with more than 50 participants in each cell yielding a power of about 70% to detect a 0.4 effect size in a two-sided test (or 82% for a one-sided test). Even though the task was novel to most of the participants, and preferences between the two options could (or should) have only depended on the physical appearance of the two prospective service providers, having one of them include subtle aspects of the participant's own face did not seem to impact their decisions.

4. STUDY 2

In our next study, we decided to shift our focus to a different domain, and employ a different type of dependent measure – namely, self-disclosure. We opted for this choice for several reasons. First, in the current era of explosion of information on social networking websites and proliferation of personal information being harvested by online companies, users' propensity to disclose personal information is an important privacy issue [1]. The second reason is that previous research has reportedly found an effect of self-morphs on level of trust [12] so a logical extension of this effect might be that individuals will be more willing to disclose personal information to someone they trust more. Lastly, we thought that perhaps self-disclosure could be a more indirect way of measuring reactions to a self-morph.

In Study 2, we used a method for eliciting self-disclosure that relies on asking participants to respond to sensitive and personal questions, and that had been used successfully in previous self-disclosure studies (e.g., [22]). Participants were asked to imagine they are talking to a therapist who asks them several questions about themselves. The therapist was either a self-morph created by using a stock model's face and the participant's face, or a morph of the same stock model and an unfamiliar person in a between-participants design. We predicted that participants would be more likely to divulge personal, sensitive information when the therapist's image was a self-morph. In this study, we used female participants only to expand our inquiry beyond males.

4. 1 Method

Based on the null results of Study 1, we adjusted our estimated effect size to $d = 0.25$ and aimed at obtaining a larger sample. We calculated that a sample of about 300 would yield a 70% power for a two-sided test, and about 82% for a one-sided test. We thus recruited 310 Caucasian female participants ($M_{age} = 30.76$, SD = 8.7) from MTurk who completed the study for a payment of $1.5 each. Participants were asked to imagine that they are looking for a therapist to discuss something going on in their life and they are referred to a specific therapist whose image is displayed. The image of the therapist was, for half of the participants, a self-morph of their own picture with a female stock model's picture (see Figure 2) and for the other half a morph of one of the other participants' image with the same female model's picture.

Figure 2. Example stimuli used in Study 2.

Participants were asked to imagine that during their meeting with the therapist the therapist asks them several questions about themselves. They could choose to answer these questions or indicate that they would prefer not to answer on a per question basis. The questions referred to engaging in unethical or socially undesirable behaviors that have been used in previous research about online self-disclosure [18]. Participants were asked to indicate, on a scale from 1 (never) to 5 (frequently) have they ever: *Had sex with the current husband, wife, or partner of a friend? Masturbated at work or in a public restroom? Had a fantasy of doing something terrible (e.g., torturing) to someone? Fantasized about having violent non-consensual sex with someone? While an adult, had sexual desires for a minor? Neglected to tell a partner about a sexually transmitted disease from which you were suffering? Had sex with someone who was too drunk to know what they were doing? Stolen anything that did not belong to you? Tried to gain access to someone else's (e.g., a partner, friend, or colleague's) email account? Looked at pornographic material?* Participants could also mark "prefer not to answer" for any of the questions.

Next, participants rated the therapist on how attractive, trustworthy, and knowledgeable they thought she was, how much they liked her, how good they thought she was at her job, how similar to themselves they thought she was and how much they identified with her. Then, we asked participants to rate how intrusive they found the questions asked by the therapist and whether she looked familiar or unusual to them. Participants then completed the NPI scale [3], and entered their demographics. They also indicated whether they found anything strange or unusual, or familiar, in either of the images, and if they said they did, then we asked them to elaborate further. Participants were debriefed as in the previous study.

4.2 Results and Discussion

We examined whether participants disclosed more to the self vs. the control-morph by examining participants *Active Affirmative Responses* (or AARs) that are the instances when participants indicated that they engage in the listed unethical or socially undesirable actions irrespective of the frequency with which they reported engaging in them [18]. In other words, AARs measure the amount of times participants indicated a response that was not "never" or "prefer not to say" to the listed unethical behaviors. Comparing AARs between conditions, we found no statistically significant differences between self vs. other conditions (M = 2.91 vs. 3.08, SD = 1.63, 1.59, t (313) = 0.944, p = 0.346, Cohen's d = 0.11). Similarly, there were no statistically significant differences in perceived ratings (see Table 1; Cronbach's alpha for the measures was 0.889). The NPI scale showed a high internal reliability (Cronbach's alpha = 0.746). Thus, we averaged the items to compute an overall NPI score for each participant and then used that average measure to examine whether an effect of the conditions could be different for different levels of NPI. We found a significant effect of NPI on self-disclosure (AARs): the higher the NPI score the more participants disclosed (*beta* = 0.163, *SE* = 0.642, p = 0.03). However, the effect of self vs. other morph was not statistically significant nor was the interaction between this effect and NPI, p > 0.37. (Descriptive statistics are presented in Table 1.)

To summarize, it appears that in the domain of self-disclosure as well we could not find evidence for the self-morph effect, as a self-morph did not seem to lead participants to disclose more personal information when compared to a non-self-morph. This study was highly powered. Thus, we feel more confident that this null finding does not represent a sampling problem.

One remaining difference we could see between our studies and previously published ones was the fact that we obtained participants' images from real-life services (their profiles on online social networks), whereas the previous researchers either took participants' pictures at the beginning of the study [12,13] or asked participants to submit a high-resolution image of themselves [4]. Therefore, previous research had the advantage of high quality pictures that could ensure high quality morphs, thereby reducing a possible source of noise relative to our experimental design (but with the disadvantage of potentially adding demand effects). In our experiments, pictures were typically of lower resolution and poorer illumination than photos captured in a lab; furthermore, profile photos included several different poses and expressions, whereas

Table 1. Comparisons between "self" and "other" condition on all measures in Studies 1-3.

DV	Mean (SD)						t (p)			Cohen's d		
	Study 1		Study 2		Study 3		Study			Study		
	Other	Self	Other	Self	Other	Self	1	2	3	1	2	3
Trustworthy	5.07 (1)	5.01 (1.1)	5.29 (1.1)	5.25 (1.2)	4.46 (1.2)	4.59 (1.2)	-0.47 (0.64)	0.3 (0.76)	1.202 (0.23)	-0.09	0.03	0.11
Attractive	5.44 (1)	5.35 (1.1)	5.27 (1.2)	5.28 (1.1)	4.16 (1.3)	4.22 (1.4)	-0.93 (0.36)	-0.1 (0.92)	0.457 (0.65)	-0.17	-0.01	0.04
Knowledgeable	4.32 (1.3)	4.55 (1.2)	4.81 (1.2)	4.69 (1.1)	4.40 (1.0)	4.46 (1)	1.74 (0.09)	0.89 (0.37)	0.65 (0.52)	0.32	0.10	0.06
Like	4.81 (1)	4.70 (1)	5.28 (1.2)	5.11 (1.1)	3.96 (1.2)	4.17 (1.3)	-1.12 (0.27)	1.31 (0.19)	1.972 (0.05)	-0.21	0.15	0.18
Identify	4.12 (1.3)	4.14 (1.3)	4.10 (1.4)	4.10 (1.4)	3.16 (1.4)	3.26 (1.5)	0.18 (0.86)	-0.05 (0.96)	0.804 (0.42)	0.03	-0.01	0.07
Similar	4.28 (1.2)	4.19 (1.3)	4.20 (1.3)	4.10 (1.4)	3.41 (1.5)	3.51 (1.4)	-0.65 (0.52)	0.61 (0.54)	0.835 (0.41)	-0.12	0.07	0.08
Good (Study 2 only)			5.22 (1.2)	5.04 (1.3)				1.29 (0.20)			0.15	
Overall judgments	4.66 (0.8)	4.67 (0.8)	4.88 (0.9)	4.80 (0.9)	3.92 (1.0)	4.03 (1.0)	-0.21 (0.83)	0.77 (0.44)	1.279 (0.2)	-0.04	0.09	0.12

photos taken in the lab were always taken frontally and with neutral expressions.

5. STUDY 3

Thus, for our third study, we decided to employ a design similar to the previous researchers' by having participants come to a lab where their pictures would be taken and used to create self-morphs. This allowed us to ensure that our participants' images were of high resolution and well illuminated, with consistent expressions and poses across images, thus maximizing the possibility of detecting the self-morph's effect if such one truly exists. Also, we increased our sample size considerably, aiming to get at least 200 respondents in each condition, that would ensure a minimum of 80% power to detect a $d = 0.25$ effect size. We preceded this study with other pilot studies that pre-tested the stimuli and questions used in this research. Lastly, Study 3 only measured trustworthiness (a measure that had produced significant results in prior research [4]) without involving additional measures of hiring intentions, self-disclosure, or others. At this point, we predicted that a self-morph would *not* be judged as more trustworthy compared to a control condition's morph. Participants were recruited either using an online participants pool at our university, which included both students and non-students, and also using a mobile "Data Truck" that was parked at several common intersections during rush hours of a large U.S. city. Participants were invited for a "study about images" that took approximately 15-20 minutes and were paid $10 for their participation. The sample included 495 Caucasian participants, 250 of them males, with an average age of 30.92 years (SD = 14.57). Study 3 took place in a lab. As participants arrived to the lab, the experimenter explained that this was a study about images and in order to take part in the study we need to take their picture, which may be used in future studies for future participants. One experimenter took a picture of the participant and uploaded it to a shared folder, while another experimenter sat in an adjacent room and prepared the morph by accessing the shared file, so participants could not see that their picture was actually being used at that time.

Participants were then seated at a computer and asked to complete an "Image task." In this task they were shown three images and asked to describe, in an open-ended manner, their thoughts and feelings about what they saw in the images. The three images were a scenery picture, a picture of several team members working together and a stock photo of a person. Next, participants completed a "Video task" that involved viewing a short video and answering some questions about it. The purpose of these tasks was both to give the experimenter time to create and insert the morph into the survey and to convince participants that the study was about image perceptions.

The third task was also called an "Image task" in which participants were asked to look at the picture of a person. This image was either a self-morph, created by morphing the participant's own picture with a stock model's face or the morph of another participant with the same stock model's face (at a 40:60 percent ratio, as in [4]). The randomized assignment was done in the following way: the first participant in a session was assigned to the "self" condition, and the next participant received the same morph as the first participant did (putting the second participant in the control or "other" condition), and so on for the following participants.

After viewing the person's image, participants were asked to rate how trustworthy they thought that person was from 1 (not at all) to 7 (very much). On the next page, participants rated how attractive and knowledgeable they thought the person was, how much they liked him/her, how similar they thought he/she was to them, and how strongly they identified with the person. Then, participants completed the Rosenberg's Self-Esteem Scale [26] followed by five questions about their own appearances by indicating how much they agree (from 1 – strongly disagree to 5 – strongly agree) with each of the following statements: *I think I am more attractive than the average person of my age; All in all, I like the way I look; I typically dislike my own pictures; I am very critical about my own looks; I like being photographed; I do not like some of my facial features*. Then, participants were asked if they found anything unusual or odd about the person whose image they just saw or if they thought this person looked familiar. Then they provided their demographics and were debriefed as in previous studies.

5.1 Replication Notes

As noted earlier, Study 3 was the closest to a replication attempt of [4]. We deviated from [4] in the following ways and for the following reasons: a) participants' pictures were shot on site (and not delivered ahead of time) to ensure high quality and standards; b) we only used a between-subjects design, as our pilot study showed no advantage to a within-subject design; c) we focused on a simple trustworthiness dependent variable, as we had already captured other variables in previous studies (and pilots), whereas for this final study we aimed at testing a straightforward and broadly applicable metrics of face-morph's impact (trustworthiness is one of original and most common metrics in self-morphs studies: see, e.g., [12] and [28]).

5.2 Results and Discussion

As detailed in Table 1, no statistically significant differences were found between the conditions on the ratings ($p > 0.05$, except for liking, $p = 0.049$) or the overall judgment score (Cronbach's alpha = 0.85).

We then examined whether self-esteem, or liking of personal appearances, could help detect an effect of the self-morph. The RSES showed high reliability (0.882), and so did the questions of "self-looks" (0.734). Thus, we averaged the questions to form two composite scores: self-esteem and self-looks. We then conducted a regression analysis on the overall judgments score with condition, self-esteem, self-looks, the interactions of condition with self-esteem and self-looks, as well as gender and age as independent variables. We found that self-looks significantly predicted overall judgments, *beta* = .73, *SE* = .35, .07, $t = 2.09$, $p = .037$. However, the condition variable (self vs. other morph) was not statistically significant, nor were any of the interactions, $p > .12$. Even when we excluded participants who reported seeing something unusual or familiar in the morphed picture, there was no significant effect of condition or the interaction of condition with self-looks and self-esteem on overall judgments, $p > .25$. In all our analyses, we could not find any support for a significant effect of the self-morph on people's judgments.

6. DISCUSSION

Previous lab research has suggested that people evaluate self-morphs differently than they evaluate face-morphs of unfamiliar people, and that self-morphs are judged as more trustworthy and attractive [12,13]. In the real world, self-morphs may be created using, for example, people's photos on social networks, and then employed to covertly influence consumers and individuals in a form of highly personalized "visceral" targeting - thus raising potential yet significant privacy concerns. Whether they might still exert influence on people's attitudes, however, was an important and open question that warranted direct research. In this paper, we examined the potential effect of self-morphs on people's online

behavior, only to realize that we could not find evidence of an effect: if such an effect does exist, it could not be captured in our studies under a variety of different settings. In the online studies that tried to find the basic self-morph effect using participants' images shared on a social networking site, as well as a highly powered third lab study, we could not replicate the effects of self-morphs that were reported in the past: we found no evidence that self-morphs impact judgments or choices regarding the purchase of products or services.

As is the case with any null result, there may be various reasons why we did not discover an effect of self-morphs in our studies that do not necessarily disprove the existence of an effect. Although most of our samples (especially Study 3) were relatively large, a bigger sample could have provided the ability to test whether the effect might still occur under some specific moderating conditions that could have explained the discrepancy between our results to previous studies. Our results may have also been due to other factors that pertain to the design and procedure of the studies. For example, while Study 2 focused on actual disclosure behavior, Studies 1 and 3 used hypothetical measures of attitudes, judgments, and behavioral intentions. It is possible that self-morphs may not affect attitudes and intentions, but could still influence people's behavior in an implicit and covert manner. Indeed, past research has shown self-morphs to affect outcomes of trust games, for one [13]. However, we still expected self-morphs to show the effect, if it does exist. Furthermore, the fact that we could not even replicate the effect on the same measure – trustworthiness – that was used in prior studies (e.g., [12]), should be regarded as problematic as well. Restricting (due to technical limitation of the morphing process) Study 1 to males and Study 2 to females might have also played a role, although usually making the sample more homogenous should increase, rather than decrease, statistical power (and Study 3 used both genders). An additional possible concern is that survey-based scales may have low fidelity in measurement, making it harder to detect small effects. However, the vast majority of morph studies, to our knowledge, also used survey answers as their main dependent variables. Across our experiments, while Study 3 ended up using survey scales similar to those employed in previous morph research, Study 1 actually leveraged scales from a different stream of literature (privacy and self-disclosure research) and Study 1 used a behavioral intention dependent variable. Finally, the morphing procedure may also play a significant role in their likelihood of affecting consumers. Images in Study 1 and Study 2 came from social media profiles; thus, the quality of resulting morphed images may have been different from morphs based on photos taken under controlled conditions in a lab. However, and importantly, quality of images was *not* different between conditions; furthermore, lower photo quality may not necessarily mean lower effect size, and Study 3 did use lab photos (as in comparable prior studies). As noted, we closely followed the methods used by [4]. We also contacted that research team and verified that we are indeed following the same procedure. In fact, we followed previous studies to extent made possible by published information and details shared with us both in terms of experimental design and morphing technique, and deviated from those in narrow details for hard-thought reasons.

While no single study was an exact replication (Study 3, for instance, focused on a simple trustworthiness metrics rather than on voting intentions), if the effects of self-morphs disappear even with relatively minor design changes, this does suggest that the effects of self-morphs on individuals' behavior may not be robust. One of our contributions therefore is that influence of face-morphs may be restricted to stringent lab conditions: while we cannot refute whether it has *internal* validity, we show it may not have significant *ecological* validity.

Future research may endeavor to chase the effect of self-morphs on individuals' perceptions and behavior - it is of course possible that, in the future, other and perhaps more sophisticated and advanced morphing procedures may overcome the limitations of our studies and discover that self-morphs can be effective at influencing individuals' judgment and behavior. More broadly, it is also possible that, outside the realm of face-morphs, other types of personal information (such as an individual's preferences for certain colors or sounds) may be used in covert manners to target and personalize messages, invitations, or suggestions. If such a visceral effect does exist, and if online firms were able to take advantage of these technologies to collect consumer information and use it to subtly and nearly undetectably target messages to influence people's behavior, it would raise important theoretical, practical, and legal issues. Policy makers would then have to consider whether current online safeguards meant to protect individuals' privacy and autonomy need to be re-evaluated in order to prevent covert third parties from exerting undue influence in such forms. Firms, on the other hand, may have to consider whether or not to engage in such strategies, given their ethical and legal implications. A broader implication arising from this manuscript, therefore, is to highlight how, due to the vast self-dissemination of personal information, public yet personal data might be used in interactions with individuals by both services and independent third parties surreptitiously and covertly – that is, without the individuals' awareness of the data being used, and of the effect it may have on their decision making.

7. ACKNOWLEDGMENTS

This work was supported by the NSF Grant Award Number 1012763 (Nudging Users Towards Privacy) and NSF Grant Award Number 1327992 (Societal, Economic, Technological, and Legal Implications of Personalized Face Composites). The authors thank the reviewers for the insightful comments, and Mr. Jeffrey Flagg for excellent research assistance.

REFERENCES

1. Acquisti, Alessandro, Brandimarte, Laura, and Loewenstein, George (2015), "Privacy and Human Behavior in The Age Of Information," *Science*, 347(6221), 509-514.
2. Acquisti, Alessandro and Gross, Ralph (2006), "Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook," In *Privacy Enhancing Technologies: Lecture Notes in Computer Science*, 4258, 36-58.
3. Ames, Daniel R., Rose, Paul and Anderson, Cameron. P. (2006), "The NPI-16 as A Short Measure of Narcissism," *Journal of Research in Personality*, 40(4), 440-450.
4. Bailenson, Jeremy N., Iyengar, Shanto, Yee, Nick and Collins, Nathan A. (2008), "Facial Similarity Between Voters and Candidates Causes Influence," *Public Opinion Quarterly*, 72(5), 935–961.
5. Beales, Howard. (2010), "The Value of Behavioral Targeting," *Network Advertising Initiative*, 1.
6. Bressan, Paola and Zucchi, Guendalina (2009), "Human Kin Recognition is Self- Rather Than Family-Referential," *Biology Letters*, 5(3), 336-338.
7. Calo, Ryan (2014), "Digital Market Manipulation," *George Washington Law Review*, 82(4), 995-1051.
8. Chen, Pei-Yu, Wu, Shin-Yi, and Yoon, Jungsun (2004), "The Impact of Online Recommendations and Consumer Feedback on Sales," *ICIS 2004 Proceedings*, 58.

9. Chevalier, Judith A., and Mayzlin, Dina (2006), "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, 43(3), 345-354.
10. Cohen, Julie E. (2012), "What is Privacy For," *Harvard Law Review,* 126, 1904.
11. Critchley, Hugo, Daly, Eileen, Phillips, Mary, Brammer, Michael, Bullmore, Edward, Williams, Steven, Van Amelsvoort, Therese, Robertson, Dene, David, Anthony and Murphy, Declan (2000), "Explicit and Implicit Neural Mechanisms For Processing of Social Information From Facial Expressions: A Functional Magnetic Resonance Imaging Study," *Human Brain Mapping*, 9(2), 93–105.
12. DeBruine, Lisa M. (2002), "Facial Resemblance Enhances Trust," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269(1498), 1307–1312.
13. DeBruine, Lisa M. (2004), "Facial Resemblance Increases the Attractiveness of Same-Sex Faces More Than Other-Sex Faces," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1552), 2085–2090.
14. Engell, Andrew D., Haxby, James V. and Todorov, Alexander (2007), "Implicit Trustworthiness Decisions: Automatic Coding of Face Properties in Human Amygdala," *Journal of Cognitive Neuroscience*, 19(9), 1508–1519.
15. Farahat, Ayman and Bailey, Michael (2012), "How Effective is Targeted Advertising?," *Proceedings of the 21st International Conference on World Wide Web, ACM*, 111-120.
16. Federal Trade Commission (2009), "Self-Regulatory Principles for Online Behavioral Advertising," *Federal Trade Commission Staff Report, Washington DC,* Feb 12.
17. Greenwald, Anthony G. (1975), "Consequences of Prejudice Against The Null Hypothesis," *Psychological Bulletin*, 82(1), 1-20.
18. John, Leslie, Acquisti, Alessandro and Loewenstein, George (2009), "Strangers on a Plane: Context-Dependent Willingness to Divulge Personal Information," *Journal of Consumer Research*, 37(5), 858-873.
19. Kanwisher, Nancy and Yovel, Galit (2006), "The Fusiform Face Area: A Cortical Region Specialized for the Perception of Faces," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1476), 2109–2128.
20. Kramer, Adam D., Guillory, Jamie E. and Hancock, Jeffrey T. (2014), "Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks," *Proceedings of the National Academy of Sciences*, 111(24), 8788-8790.
21. McDonald, Aleecia and Cranor, Lorrie (2010), "Beliefs and Behaviors: Internet Users' Understanding of Behavioral Advertising."
22. Moon, Youngme (2000), "Intimate Exchanges: Using Computers to Elicit Self-Disclosure From Consumers," *Journal of Consumer Research*, 26(4), 323-339.
23. Open Science Collaboration (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349(6251), aac4716.
24. Platek, Steven M. and Kemp, Shelly M. (2009), "Is Family Special To The Brain? An Event-Related FMRI Study of Familiar, Familial, and Self-Face Recognition," *Neuropsychologia*, 47(3), 849–858.
25. Purcell, Kristen, Brenner, Joanna, and Rainie, Lee (2012), "Search Engine Use 2012," *The Pew Research Center's Internet & American Life Project, Washington DC*.
26. Rosenberg, Morris (1965), *Society and the Adolescent Self-Image*, Princeton, NJ: Princeton University Press.
27. Samat, Sonam, Acquisti, Alessandro and Babcock, Linda (2017), "Raise the Curtains: The Effect of Awareness About Targeting on Consumer Attitudes and Purchase Intentions," In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)* (pp. 299-319). USENIX Association.
28. Tanner, Robin J. and Maeng, Ahreum (2012), "A Tiger and a President: Imperceptible Celebrity Facial Cues Influence Trust and Preference," *Journal of Consumer Research,* 39(4), 769-783.
29. Turow, Joseph, King, Jennifer, Hoofnagle, Chris, Bleakley, Amy, and Hennessy, Michael (2009), "Americans Reject Tailored Advertising and Three Activities that Enable It," *In SSRN: http://ssrn.com,* Vol. 1478214.
30. Verosky, Sara C. and Todorov, Alexander (2010), "Generalization of Affective Learning About Faces to Perceptually Similar Faces," *Psychological Science*, 21(6), 779–785.
31. Winston, Joel S., Strange, Bryan A., O'Doherty, John and Dolan, Raymond J. (2002), "Automatic and Intentional Brain Responses During Evaluation of Trustworthiness Of Faces," *Nature Neuroscience*, 5(3), 277–283.
32. Yan, Jun, Liu, Ning, Wang, Gang, Zhang, Wen, Jiang, Yun, and Chen, Zheng (2009), "How Much Can Behavioral Targeting Help Online Advertising?" In *Proceedings of the 18th International Conference on World Wide Web* (pp. 261-270). ACM.

"Privacy is not for me, it's for those rich women": Performative Privacy Practices on Mobile Phones by Women in South Asia

Nithya Sambasivan, Garen Checkley, Amna Batool[#], Nova Ahmed[*], David Nemer[+], Laura Sanely Gaytán-Lugo[Ω], Tara Matthews[Ψ], Sunny Consolvo, and Elizabeth Churchill

Google Inc., Mountain View, CA, USA,
{nithyasamba, garen, sconsolvo}@google.com, churchill@acm.org
[#] Information Technology University, Pakistan
batool.amna@itu.edu.pk
[*] North South University, Bangladesh
nova.ahmed@northsouth.edu
[+] University of Kentucky, USA
david.nemer@uky.edu
[Ω] Universidad de Colima, Mexico
laura@ucol.mx
[Ψ] Independent Researcher
taramatthews@gmail.com

ABSTRACT

Women in South Asian own fewer personal devices like laptops and phones than women elsewhere in the world. Further, cultural expectations dictate that they should share mobile phones with family members and that their digital activities be open to scrutiny by family members. In this paper, we report on a qualitative study conducted in India, Pakistan, and Bangladesh about how women perceive, manage, and control their personal privacy on shared phones. We describe a set of five performative practices our participants employed to maintain individuality and privacy, despite frequent borrowing and monitoring of their devices by family and social relations. These practices involved management of phone and app locks, content deletion, technology avoidance, and use of private modes. We present design opportunities for maintaining privacy on shared devices that are mindful of the social norms and values in the South Asian countries studied, including to improve discovery of privacy controls, offer content hiding, and provide algorithmic understanding of multiple-user use cases. Our suggestions have implications for enhancing the agency of user populations whose social norms shape their phone use.

1. INTRODUCTION

A large and growing population of nearly 760 million women live in India, Bangladesh, and Pakistan [55–57]. One of the highest worldwide gaps in phone ownership is among women in South Asia (that is, the sub-Himalayan region of eight southern Asian countries including India, Pakistan and Bangladesh). Here, women are 26% less likely to own a mobile phone compared to men [17]. Twenty-nine percent of South Asian women regularly borrow a phone [16]. Even when phones are individually owned—*i.e.*, in the possession of a user for a majority of the time—women in many South Asian contexts face cultural expectations to share their devices and digital activities. For example, in a survey conducted by GSMA, men, and sometimes women too, found it acceptable for a husband to check his wife's digital activity on her phone [16].

However, in the design and development of mobile devices and services, user privacy is predominantly modeled on the "one account, one user" paradigm, despite the fact that shared device usage of devices challenges the definition, architecture, and presentation of privacy controls developed on this assumption [3,7,23,36,37,40,44].

Prior work in various cultural contexts has focused on shared device practices among families, co-workers, friends and strangers, identifying factors such as economic constraints and social values that drive shared use [7,23,36,37,40,44]. Fewer studies have focused on social power relations as drivers for shared use and the resulting privacy practices and challenges, for example in settings where cultural expectations shape mobile phones that are shared and digital activities that are scrutinized by family relations.

In this paper, we examine the ways in which current technology designs could better support the privacy challenges of women in South Asia We explore two main questions:

- How do women in South Asia perceive, manage and control their privacy on shared mobile phones?
- How are social expectations of women fulfilled through technological and social affordances?

We report results from a study with 199 women in India, Pakistan, and Bangladesh who were owners of phones (167 of them owned smart phones, 22 had feature phones). Among our key findings

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018. August 12 -- 14, 2018, Baltimore, MD, USA.

were that our participants' digital activities on their phones were carefully monitored by close social relations, and participants' mobile phones were highly shared between people in their household.

In addition to describing the social context in which women's phones were shared and their use was monitored, we describe five performative practices used by our participants that allowed them to maintain some individual privacy on their mobile phones while adhering to cultural values of transparency and sharing that were expected of women's gender roles. These practices were: *phone locks* for prevention of misuse by strangers; *app locks* for securing content and applications from weak ties and children; *aggregate* and *entity deletions* of content, queries, recommendations, and history to remove content traces from everyone's view; *private modes* to enter private experiences; and *avoidance* of certain digital activities situationally or permanently. These ordinary privacy practices became performative when they enabled the women in our study to balance their gender role expectations of openness, manifested in openly shared phones and apps, with their own desire for privacy on devices. The repertoire of mostly covert and sometimes overt privacy practices was employed as needed in various social situations. For example, by selectively deleting search queries, our participants maintained covert privacy while not signaling that they were hiding anything from those who shared or monitored their devices. We also recommend several design suggestions that we hope will better support the needs of user groups that face cultural expectations to share mobile phones, such as offering content hiding, transience, and improved algorithmic feedback for shared use.

The paper is structured as follows. We begin by situating our research in related work. We then describe our study methodology and follow with our results. Finally, we make recommendations for designing technology for contexts in which device sharing is common and expected.

2. RELATED WORK

To situate our research contributions, we discuss related work on device sharing and access controls with social relations, device monitoring by social relations, and research at the intersection of gender and privacy.

2.1 Device sharing & access controls

Several research studies document the social practices around device sharing. We focus primarily on the literature on device sharing in South Asia, because of cultural similarities with our study. Definitionally, Matthews *et al.* define 'device sharing' as the action of using a device or an account by two or more people, simultaneously or one after another [29]. Studies in the West have documented a range of concerns with shared devices: for example fears of data being deleted [23]; desire to use profiles to personalize content instead of achieving privacy over content [7]; and how children act as trusted adversaries in households [41].

Prior work from South Asia has focused on the prevalence of shared phone use, exploring both economic constraints and cultural values. Cultures of shared technologies are so prevalent in these regions that James and Versteeg argue that subscriptions and accounts are not a reliable measure of mobile phone access; rather phone usage remains the best measure [21].

Others have examined the motivations and practices around phone sharing in South Asia. Steenson and Donner describe how mobile phones were shared in Indian households along two dimensions: proximity and socio-spatial contexts [44]. They observed that phone sharing may occur informally due to co-presence or stealthily without the owner's knowledge. Phones were also shared when they were used to call someone known to be near the phone, or when the phone was used as a family landline. Sambasivan *et al.* describe how devices are shared in low-resource communities due to the presence of fewer devices, leading to 'intermediated usage,' where technologically aware members may use technologies on behalf of those with lower technical literacy; thus intermediation vastly expanded access to devices, especially for women [40]. Rangaswamy and Sambasivan describe how technologies were fluidly shared in slum communities in India, deriving more value out of less money [36]. They invoke a local term, *cutting chai*, used to share a cup of tea among many members, as a metaphor for how a device is divided among many users.

Access controls can help users cope with device sharing. Little has been said about access controls in shared device environments in South Asia, though. A notable exception is a study by Ahmed *et al.* [3] on privacy challenges with shared mobiles. They showed that device sharing is a cultural practice that can be affected by power relations. The authors briefly discuss gender dynamics, but this was not the study's focus.

The literature on access controls from the West is extensive; for example, user profiles, locks, and logins have been well-researched in Western contexts [6,7,11,12,18,19,23]. Across these studies, common themes include the importance of appropriate access controls, flexibility, and customization for various social contexts.

Family profiles have been reported to be a good middle ground between individual profiles and a single shared account for all users, in environments where privacy and security requirements across users is less stringent [11]. Guest profiles with discrete switching have been recommended to avoid misunderstanding [31]. Karlson *et al.* showed that the binary access models on phones do not address the social discomfort users experience when sharing phones [23]. Transparency of access controls to avoid social implications has been suggested by Harbach *et al.* and Mazurek *et al.* [18,31]. Logins [6] and locks [12,19] have been studied, for example, showing that all-or-nothing lock models do not fit the needs of users [19], and that users may not make a connection between sensitive data in apps and the need for locking [12].

Most prior studies on device sharing and access controls have been conducted in the West, where adult users are typically not socially obligated to share their phones to the same extent as women in the South Asian countries we studied. In contrast to prior studies in South Asia, which have focused on women as borrowers or recipients of sharing [36,40,44], all our participants owned their Internet-enabled phones but were still culturally obligated to share with others. We describe privacy techniques employed by our participants, showing how they fulfill cultural expectations of sharing and yet maintain some privacy on their devices. We further show how many of these practices are unanticipated workarounds due to poor app usability.

2.2 Device monitoring by social relations

Our study revealed that women's devices in South Asia are monitored by their social relations (including husbands, brothers,

parents, and children). While some device sharing research touches upon device monitoring by social relations, a series of other studies, with a range of specialized populations, focus on these issues. As an overview, Marques *et al*. [28] showed that snooping on other's phones is something that an estimated 1 in 5 U.S. adults had done in the year prior to their survey. Monitoring of device use by social relations has been studied in multiple more focused contexts, including (but not limited to) parents monitoring children (*e.g*., [8,15,51,52]), snooping in romantic relationships (*e.g*., [9,28]), and intimate partner abuse (*e.g*., [10,13,14,30,43,53]. Much of this monitoring research has occurred in Western contexts, barring the exception of a study of Bangladeshis' shared phone use [3] and a high level overview of this team's research on gender equity [39], which both allude to monitoring of women's phones.

A common theme in this literature is that monitoring *does* occur, but it is often *not* a socially desirable behavior in the West. Monitoring is generally more accepted in parent-child relationships, but it still not necessarily welcome by the person being monitored [15]. Another theme is that study subjects try to maintain privacy from social relations but face various challenges. For example, abusers go to great lengths to monitor and control survivors (such as coercing survivors into physically sharing a device, or covertly installing spyware on the survivor's device), leading some survivors to take drastic actions like deleting accounts or abandoning devices [13,14,30,43]. In studies with a general population, willing device sharers have expressed an obligation to give open access to close relations to communicate trust, which opened them up to snooping [29].

While our study also discusses monitoring by social relations, the cultural context—especially the acceptability of social monitoring—is very different from prior work. Our study explores an under-studied population and describes a variety of cultural factors that result in the commonplace and sometimes accepted practice of social relations regularly monitoring women's devices in South Asia, and how the women perceive and react to this reality.

2.3 Gender and privacy

A growing body of research observes that women's use of technology in South Asia is limited and controlled by cultural norms in a variety of ways. Privacy is sometimes discussed, but the focus is primarily on the ways in which women's use of technology is limited. For example, technology needs and perceptions are different depending on a person's gender [34]. An emerging area of research is concerned with women living in gender-unequal contexts [2,3,22,24,39,40,46–48]. Restrictive gender norms limit the impact of information technologies for women in practice [3,39,47]. For example, Abokhodair and Vieweg, in their research study in Saudi Arabia and Qatar, reported that women preferred to keep their online presence private and restricted to same-gender interactions [1]. Meanwhile, Sultana *et al*. detail how some women depended on their husbands, even in emergency situations, as they were required to wait until their husbands returned home in order to make phone calls [47]. Murphy and Priebe present a well-rounded literature review on how class, race and sex shape women's attitudes towards mobile phones, by discussing cases from India that reveal how gendered perceptions of modesty conflict with phone ownership [32]. Sambasivan *et al*. briefly describe device sharing and privacy practices employed by women in South Asia, in a broader research overview of gender and technology [39].

However, technology can be empowering to women. For example, research by Alghamdi *et al*. showed how online banking enabled Saudi women to perform banking transaction from home, giving them new financial autonomy. When the task had to be completed in public, their male family members had to transact on their behalf due to Islamic principles of gender segregation [4]. In another example in Morocco, where unrelated women and men engaging on phones was culturally taboo, SMS codes helped women communicate with water managers [46].

While these examples touch upon some implications for how women in South Asia experience privacy, it is not their focus and so we do not have a full understanding of the privacy issues women face and how they cope with them. Our work contributes to this body of work by focusing on women's privacy challenges and practices in a cultural setting where device and account sharing is typical. Distinctively, women in our study *had* access to phones and were not reliant on borrowed phones.

3. METHODOLOGY

Our research inquiry was focused on understanding mobile phones in daily life, as part of a larger project on studying how women in South Asia encountered technology. We conducted focus groups with a total of 199 women. The research was conducted from May to December of 2017. In total, we conducted over 500 hours of fieldwork across India, Pakistan, and Bangladesh (see Table 1 in appendix for a breakdown of participants and sites). We conducted focus groups of three participants per group (triad focus groups) with 199 participants who identified as women. Focus groups were chosen because it was easier to break the ice and share common experiences on the sensitive nature of the topics covered. Each focus group session lasted about 2 hours on average. The focus group discussions were semi-structured in nature and organized around aspirations, phone and Internet use, device sharing, privacy practices, identity models, and safety concerns. We ended every focus group by asking the participants what topics or issues they would like to highlight the most in our research reports, giving them a chance to reflect upon the conversations and represent their voices in their own terms. Interview questions are provided in the appendix.

The study followed a comparative fieldwork format [33]; rather than a thick description of behavior and context, comparative fieldwork helped us understand points of transition where phenomena break, continue, or transform.

Here we describe participant recruitment, data collected, analysis, and ethical considerations in reporting this research.

3.1 Participant recruitment

Participants were recruited through a combination of non-governmental organizations (NGOs), personal contacts, and recruitment agencies, using snowball and purposive sampling that was iterative until saturation. Prior to the sessions, recruitment contacts and NGO staff verbally mentioned the purpose of the study, the categories of questions (access, information & content, privacy and safety), and the affiliation of the researchers, providing potential participants an opportunity to decline participation prior to any contact with our team.

Focus group participants were already known to each other, like friends and neighbors, in order to help with rapport and trust. Incentives varied depending on the country, demographic and

format of session. Sample size was determined based on ensuring representative coverage, balanced with recruitment resources available in each country. In order to obtain a well-balanced sample, participant recruitment was divided such that roughly a third of participants each were of high, medium, and low socioeconomic status (SES) as determined by SES definitions per country and verified through income, education and material possessions [27]. Participants were from 18 to 65 years old. All were Internet-enabled phone owners. See the appendix for more detail.

3.2 Moderation and incentives

In all three countries, focus group moderators were native female researchers and regional language speakers, to leverage common cultural ground [25]. Another researcher took notes. Due to the mixed gender nature of our design-research team, male designers were observers during interviews. Country-specific incentives are noted below. The incentives were ethically determined to not be coercive, based on socio-economic segments. Incentives were determined via research experts from and specialized in the countries. All participants were verbally thanked for their time at the end of interviews.

3.3 Analysis

Interviews were conducted in local languages and translated to English in transcription (see country sub-sections below). Inductive analysis was conducted on the raw interview data [49]. We focused on stories about (1) access to devices and software; (2) technology usage by women; (3) privacy considerations in shared spaces; (4) management of uncomfortable or sensitive information on shared devices; (5) identity and account handling in shared use situations; and (6) aspirations for a different social order around device usage. From a close reading of transcripts, we developed categories and clustered excerpts together, conveying key themes from the data. Three team members created a code book based on the themes, with four top-level categories (identity, co-located privacy, access, online privacy) and several sub-categories *e.g.*, micro-deletions, public environments, and technology literacy). Codes were iterated in the order of conducting research: India codes were developed first, then iterated with Bangladesh and Pakistan. The five practices that are the focus of our results were then developed and applied iteratively to the codes (see appendix).

3.4 Research ethics

To protect our participants and to create neutral and non-judgmental spaces, we invited them to coffee shops, restaurants, university campuses, and NGO locations where they felt safe and comfortable. Having a neutral, safe space was important as contextual interviews in the home or work posed the possibility of other co-located members like in-laws and children overhearing, which could open up possibilities of participant harm and compromise accuracy of responses. Same-gender and same-ethnicity moderation was employed to leverage common cultural ground and build trust. Note-takers were men on our research team, who positioned themselves to sit in the background to not obstruct the rapport between the participants and moderator. For sensitive topics, such as privacy and surveillance, male research members pro-actively left the room to give participants space.

Verbal informed consent was translated by a native speaker into local languages, explained and obtained from all participants. Fifteen-to-twenty minutes were spent explaining the purpose of the interviews, answering any questions, and building rapport. Participants were made aware that they had the right to terminate the study at any point without forfeiting the incentive. Methods of recording, *i.e.*, audio, video, notes, or none of the above, were explained to participants, who chose the most comfortable technique. In a few interviews, we stopped recording and taking notes when participants became emotional; we retroactively wrote textual notes after the interview. All data were stored on a locked Google Drive folder, with access limited to the research team.

Only pseudonyms are used in this paper. Any identifying information has been redacted. Age ranges are reported to protect participant privacy. Locations are only specified if the population is larger than 100,000.

3.5 Country-specific details

3.5.1 India (n=103)

In India, our 103 female participants included college students, housewives, domestic maids, village farm workers, IT professionals, bankers, small business owners, teachers, and two women with physical and visual disabilities (banker and microenterprise owner). Focus groups were conducted in Chennai and Bangalore (south India); and Delhi, Kanpur, and villages in the state of Uttar Pradesh (north India).

Focus groups were conducted in rented conference rooms, community centers, cafes and restaurants, universities, and quiet public spaces like communal seating areas. The first author conducted each interview in Hindi, Tamil, and English, depending on the participants' language preference. Recordings were transcribed into English by the research team. Each participant received $10-15 USD for participation, depending on urban versus rural locations.

3.5.2 Pakistan (n=52)

In Pakistan, our 52 female participants included working women, housewives, and students. Occupations of working women included gym trainers, janitors, beauticians, school teachers, security personnel, corporate employees, university instructors, and home tutors. Focus groups were conducted in Lahore, Multan, and Rawalpindi (central Pakistan); Peshawar (northwest Pakistan, bordering Afghanistan); Karachi (south Pakistan); and Hunza (north east Pakistan, bordering India). We chose places like community centers, schools and facilitators' homes for conducting the focus groups according to the comfort levels of participants.

Participants were recruited with the help of local facilitators. We visited Muslim, Christian and Ismailee communities with facilitators to recruit participants and to conduct the focus groups in their communities. Goody bags consisting of food items worth up to $5 USD were distributed among the participants who showed up for interviews. Cash incentives worth $50 USD were given to facilitators in each city. All focus groups were conducted in Urdu and responses were audio or video recorded after obtaining verbal consent from participants. Recordings were transcribed into English by the research team.

3.5.3 Bangladesh (n=44)

In Bangladesh, our 44 female participants included garment workers, housewives, teachers, medical doctors, engineers, and day laborers. Focus groups were conducted in Dhaka (central

Bangladesh), Chittagong (southeast Bangladesh, bordering India), and Sylhet (northeast Bangladesh, bordering India). Participants were recruited by contacting each group through a known contact, such as through members of the research team, university staff, and known professional and personal contacts, in order to gain the trust of participants.

Focus groups were conducted in Bengali, and recordings were later transcribed to English. Incentives of warm food along with monetary incentives of $12 USD, or the gift equivalent, were provided for each participant.

3.6 Gender in South Asia

We picked these three countries for the study since they share a great amount of cultural and economic similarities. The three countries used to be one unified country, India, before the British partitioned India into three free countries when they left in 1947: India, East Pakistan, and West Pakistan. In 1971, West Pakistan became Pakistan and East Pakistan became Bangladesh. In all three countries, women occupy a tenuous position between empowerment and disempowerment. All three countries have had female Prime Ministers, CEOs, and public intellectuals since independence. Yet, women face gender inequality in multiple areas, including health, education, and the economy, due to complex cultural beliefs and practices.

4. FINDINGS

An overarching theme in our results is that participants had to cope with an expectation that they allow their phones and accounts to be frequently monitored by a variety of social relations. In the first section below, we describe device sharing as a cultural expectation and how this led to mediated and monitored technology use for participants. Since participants were embedded in this cultural context where their technology use was monitored, they were generally accepting of it; we discuss these perceptions of privacy in the second section below. However, participants experienced situations when they wanted to avoid having others learn about their digital activities. In the third section below, we describe the practices they adopted to maintain some privacy when device sharing was expected.

4.1 Device sharing as a cultural expectation

In our study, cultural norms for women were one of the major factors that led to phone sharing (also seen in [3,44]). Participants experienced a cultural expectation that they, as women, would share their devices with social relations. In practice, this could involve multiple onlookers as they used their device, having their device passed between multiple people, or using a device that was primarily shared in nature. Since women are typically viewed as the caregivers, participants often reported that their children used their phones to play games or watch videos. Note that this cultural expectation of sharing did not end with phones; participants were also expected to share personal belongings like jewelry, savings, and *saris* (clothing) with other family members.

Other factors also motivated device sharing. While access to a phone was not a barrier in our study (phone ownership was a criterion for participation), the high cost of mobile data sometimes led to shared use. In most of the cases, sharing was reported to be a voluntary act, including in some cases where it may have been considered (by the participant) a man's or elder's right to monitor the woman's devices. Regardless of the perception of sharing among our participants, all of them created practices to maintain a sense of privacy.

Below, we highlight various contexts in which our participants' device use was shared, mediated and monitored.

4.1.1 Shared usage (IN (India): 83; PK (Pakistan): 31; BG (Bangladesh): 11)

Many participants reported sharing mobile phones in the household (83 out of 103 participants in India, 31 out of 52 in Pakistan, and 11 out of 44 in Bangladesh stated experiencing this theme). In Peshawar and Hunza in Pakistan, some participants noted that they were not able to own their own phones until they were married (they did own at the time of the interview). When women had mobile phones, their devices were often viewed as 'family' devices. Several mothers in our study reported that their phone became the default shared phone of the family. A mother's loss of identity in possessions and space has been well documented, *e.g.*, [26,45]; however, this generally gendered issue takes on a specific locational nature in South Asia, discussed here around mobile phones. Some women in Bangladesh reported that their children would immediately grab their phone when the women returned home from work but left the father alone or asked to use his phone much less. As Shaina, (a 20 to 25-year-old young mother of two from Chittagong, Bangladesh) noted:

"*My kids don't touch the father's phone. They only use mine all the time. They are scared of him....my daughter broke my husband's phone and got a lot of beatings. She only uses mine. So I have an app lock on my phone.*"

4.1.2 Mediated usage (IN: 33; PK: 20; BG: 4)

Mediated usage refers to one person setting up or enabling a digital experience for a less tech-savvy user (*e.g.*, a daughter might search for and play a video for her mother). Some participants from all three countries described that it was common for a man in the family to load content that she desired.

As documented elsewhere, mediated usage builds upon the social infrastructure and enables women, especially those with lower technical literacy, to make use of tools they find challenging to use [40].

While some men enabled female relatives to access technology, this practice was also restrictive in that it required women to rely on others for access. As Zeenat (a small business owner, 30 to 35-year old in Lahore, Pakistan) described, she depended on her husband each time she logged in to social media:

"My husband created my Facebook account and he didn't enter my complete information. Because I didn't know how to create a profile, I asked him to create it for me. Now every time I want to use it, he logs in for me."

4.1.3 Monitoring (IN: 43; PK: 17; BG: 2)

Monitoring refers to situations in which someone other than the primary user examines the phone, without otherwise having a need to use the phone. Among the 62 participants who experienced monitoring, their reactions to it were mixed. Roughly half of these participants viewed it as being acceptable for men, elders and in-laws to monitor their devices, and they did not usually reciprocate

by examining the device of the person doing the monitoring (although, in a few cases, participants reported checking on their husband's devices secretly). Some of these participants reported that they appreciated when male members checked their phones to ward off unwanted calls and attention on social media, or to check for viruses, as these participants perceived that their technological skills were lower than those of the person monitoring.

Being open to monitoring was performative, in that in enabled participants to show and feel they fit the role of a good family member. As Sujata (a 20 to 25-year old receptionist in New Delhi, India) noted, enabling her parents to check her device fit with upholding the image of being a 'good daughter.'

"My parents can pick up my phone and check whenever they want because they have the right to. They give us freedom all the time. We have nothing to hide from them."

In another case in old Dhaka, Bangladesh, Nilima (a 50 to 55-year old school teacher) told us how her husband had the right to check her messages, and she felt that it was acceptable.

In some cases, monitoring was viewed as coercive. In Bangladesh, two (out of 5 whose husbands worked abroad) participants reported that their husbands installed tracking tools on their phones to monitor their phone activities. Aysha (a 25 to 30-year old domestic worker) reported feeling upset when her husband first mentioned putting spyware on her phone (it is unclear if he actually did so), but she has now found ways to deal with the monitoring. She explained,

"When I call my mother or make personal calls, I borrow my employer's phone."

Small spaces and multi-generational households also led to over-the-shoulder looking. Participants reported that content was accidentally viewed by family members around the home, especially large content, visual content, and the applications women used.

4.2 Participants' notions of privacy

In this section, we describe what 'privacy' meant to our participants. At the outset, it should be noted that the term 'privacy' carried a variety of connotations and implications for the women we interviewed. Across the three countries, it was often challenging to discuss privacy. The term itself was sometimes considered objectionable, particularly among the participants with lower and middle SES backgrounds. Many participants described that 'privacy' was for upper class families, where boundaries in personal and social settings were acceptable, but it was not a part of their cultural ethos that emphasized openness.

'Privacy' was often viewed as a Western concept, imported along with cultural goods like *"jeans and dating,"* as Bhanu (a 30 to 35-year old housewife from Delhi, India) described. A direct analogy offered by Raahat (a 25 to 30-year old office clerk in Dhaka, Bangladesh) was that of *"closing doors...we don't allow it in our family unless there is a special situation. Privacy is like that, it is against our values."* To contextualize the analogy, in some socio-economic segments, the idea of closing a door is considered unacceptable or even unheard of, especially among lower-income families that inhabit one-room homes.

Conversely, the more educated and wealthier participants did not associate a stigma with the term 'privacy', which prior work attributes to their education in liberalized institutions and association of higher social classes with westernized values of individuality [5].

In contrast to the verbal dissociation of the concept of privacy, *all* participants in our study—no matter their SES background—employed strategies and techniques that the usable security and privacy community would likely call safeguarding and controlling their 'privacy' on their devices. While many of the lower to middle SES participants did not think of these practices as privacy-related, the practices were intentional steps taken to protect device activities and content from being revealed to co-located household members. The higher SES participants did associate these practices with the concept of privacy.

4.3 Privacy practices in device sharing

Despite the wide range of views on what 'privacy' meant and how applicable it was to them; our participants used an assortment of practices to keep others from learning about some of their digital activities. In the cultural contexts of our study, the outright refusal for a woman to hand over her phone to men or elders was considered disrespectful or impolite. Thus, our participants used several privacy practices—phone and app locks, content deletion, private modes, and technology avoidance—to maintain individual privacy (see Figure 1 for a summary). These privacy practices were *dynamic* and *situated* in the social setting, in that they varied according to the social relationship, space, and device activity. They were also *performative* in that they enabled participants to uphold the impression of openness that was culturally expected of them, while maintaining some privacy. The level of sophistication of the privacy practices varied based on the participants' familiarity with technology.

4.3.1 Phone locks (IN: 83; PK: 27; BG: 6)

Participants regularly locked their phones with pins or patterns to prevent misuse by strangers or in cases of theft. Such *phone locks* can be an effective strategy in many contexts [11,18]; however, they were almost never effective in preventing proximate family members or friends from accessing the mobile phone. Many participants reported living in small spaces and maintaining open social environments, such as spending time in the living room and not necessarily having one's own room, which led to over-the-shoulder device onlooking. As Jyoti, a 40 to 45-year old housewife in Kanpur, India noted,

"Since I want to prevent my kids from using my phone, I use phone locks. But my kids open it each time I change it. They are too smart. I have to change my app lock pin every week. I have done it so many times that I often forget my pin."

Phone locks were most effective in providing peace of mind in theft and unmonitored scenarios. As Yasmin, a 30 to 35-year old garments worker in Dhaka, Bangladesh described, phone locks brought comfort when strangers may have accidental access to the phone.

"I am extra careful about phones as it is confidential. My phone was misplaced once and I panicked. Later when it was returned I remembered that it had a phone lock."

4.3.2 App locks (IN: 43; PK: 9; BG: 6)

Following phone locks, the second most commonly used strategy reported by our participants was that of *app locks*. App locks, such

as Do Mobile app lock, Security Master, and Cheetah Mobile, provide users with the ability to password- or pin-protect specific applications, content, or folders. In comparison to the largely ineffective strategy (as reported by participants) of using phone locks in co-located groups of families or friends, app locks were reported to provide more control to participants, although not always.

App locks provided privacy protection to participants who shared their phones but wanted to maintain privacy over certain applications or folders. In many cases, app locks were enabled after a privacy violation had occurred among co-located others. Sanaa (an 18 to 25-year-old beautician in Lahore, Pakistan) noted how she had to turn over her phone to her employer during work hours per work rules. In her case, a prior incident of monitoring by the receptionist staff motivated her to install an app lock.

"My friend introduced me to app lock. As I work in a beauty salon. I have to submit my phone to the receptionist when I go to work. Other staff also do this. I found out that some of my messages were read by someone when I submitted my phone to the receptionist. I told my friend about this, and she said that the receptionist did this to her too when we were not looking, and she asked me to install app lock. Now I feel secure. Anyone can borrow my phone for calling."

As Sanaa notes above, app locks allow users to share their devices, instead of blanket refusal, by providing granular control over specific apps or content. Most of our participants hid social media applications, photo and video folders created by social applications, and Gallery (a photo editor and storage folder). A few participants reported hiding other applications like menstrual period trackers, banking applications, and adult content folders. As Gulbagh (a 20 to 25-year old college student from Multan, Pakistan) described:

"I have enabled app locks in addition to the phone lock. I have it on WhatsApp, Messenger, and Gallery because sometimes friends share some pictures and videos with you that are only meant for you [smile]. My brother is never interested in my phone but it is my younger sister who is a threat [laughs]. So I have an extra shield of protection."

App locks also prevented friends or children from accessing data-intensive applications. In a context where the cost of mobile data is relatively high as a proportion of monthly expenses, many moms in our study reported locking apps (*e.g.*, video apps) to prevent children from spending too much mobile data. Another common concern was that children would accidentally delete an application. However, app lock passwords were sometimes easily known to co-located household members, similar to phone locks.

"Both my elder daughters use my phone. I have enabled an app lock on my phone but my kids learn the lock pin easily. Even if I change, they learn it."

The design of most app locks enables privacy, without consideration for secrecy. Five participants mentioned that the visible app lock password or PIN screens when invoking certain applications, or the very presence of the application on the phone led to questions such as, "what are you hiding?" Some app lock applications were reported to enable invisibility, but that often costs extra. As noted in [36,38], there is a general reluctance among technology users in many emerging markets to pay for online applications, services, or content due to freely available pirated content and lower affordability.

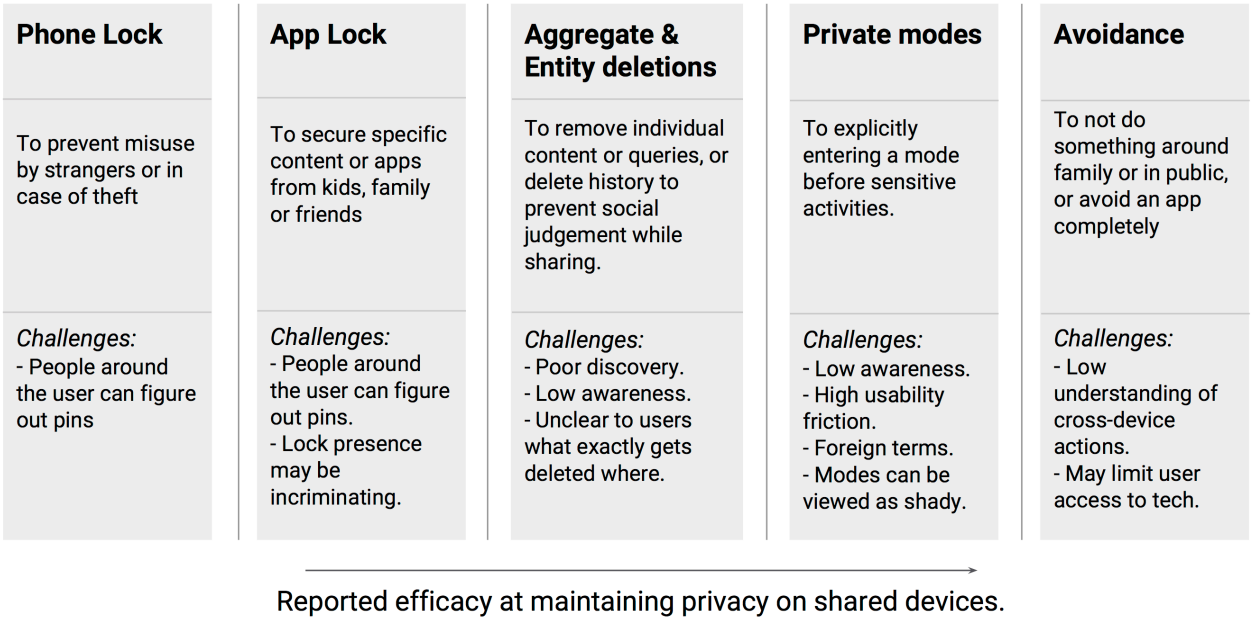


Figure 1: Reported efficacy by participants in achieving their privacy goals on shared phones

As Rupa, 30 to 35 years old, from Chennai, India, explains,

"If you hold a button on Vault, you can see a screen where it allows you to hide the app lock. But you have to pay to use it. I heard there is another app where if you press a button five times it becomes visible."

To summarize, app locks were popular among our participants since they enabled a degree of privacy among co-located others. However, two challenges were (1) that passwords and PINs were often discovered by others in close proximity, and (2) if others found the app lock, that might suggest the participant was trying to hide something which could lead to tensions.

4.3.3 Aggregate and entity deletions

Phone locks and app locks were used by our participants to prevent others—acquaintances, strangers, and children—from accessing personal applications and content. Participants deleted information in situations where devices traveled freely across various users. While locks make visible the refusal to share certain applications, information deletion was used to remove sensitive content without a detectable trace. Two types of practices were observed in content deletion: (1) *aggregate deletions*, where participants deleted entire threads or histories of content, and (2) *entity deletions*, where participants deleted specific chats, media, or queries. However, many participants were not aware of these aggregate and entity deletion controls, so they often resorted to avoiding applications entirely.

4.3.3.1 Aggregate deletions (IN: 17; PK: 5; BG: 9)

Participants reported using aggregate deletions when (1) they were not able to find a mechanism to delete a specific piece of content, or (2) they wanted a large amount of their content deleted, for example browsing history, search history, or message history. Confusion appeared when participants reported wanting to delete specific content, but ended up deleting all content history, because they were not able to discover the affordances to delete specific content. In a few cases, search history was also deleted because participants perceived that it would speed up phone performance (most participants owned low-end mobile phones in the $50-$100 range). Janaki, (a 35 to 40-year-old clerk in Chennai, India), explained:

"See when I search for something, it shows what else I have searched before. Sometimes it can be a little cheap for other people to see. I like to see medical videos on ladies' topics or 'those' type of videos. But others will get doubts on my character. When my son uses my phone, he will think why is amma [mother] *seeing all this. So I just clear my search history every week to be safe."*

Among more technologically aware participants, concerns over cross-platform privacy leaks and complex strategies emerged. Chitra (a 20 to 25-year old engineering student in Bangalore, India) recounted how she deleted her search history on occasion. Recently, she had searched and shopped online for gifts for her boyfriend. Chitra was wary of ads popping up on other platforms and awkward questions from her relations, like "who are you buying a men's t-shirt for?". So she deleted her search history. Chitra's friend and classmate, Chrissie, had sophisticated practices to negotiate device privacy using pause-and-resume functionality. She noted:

"I like to watch Game of Thrones. When I see clips or highlights, I first pause my viewing history and resume after I have finished watching the clip. If I am too worried, I just delete the entire history. But sometimes I forget."

To summarize, aggregate deletions were commonly employed to achieve peace-of-mind regarding the privacy of all browsing, searching, and viewing habits. A common assumption made by our participants was that deleting history would delete all records on that original platform and other platforms that communicate with it. However, this may not be true in most cases, where deleting history may not delete personalized recommendations already trained on the user's habits and does not delete browser cookies and data exchanges to other cross-linked platforms. Private modes, discussed below, may have been more helpful to participants in accomplishing their goals.

4.3.3.2 Entity deletions (IN: 89; PK: 29; BG: 9)

Entity deletions were used to remove individual items or actions—such as texts, photos, previously searched terms, etc. While aggregate deletions were more commonly used by participants for web content and specific applications like video and shopping platforms, social media content was predominantly managed through entity deletions.

The prevailing use case for entity deletions was to remove sent and received media and messages, to control what others who used or monitored their phones would see. Photos, videos, and texts were deleted from chats and folders. Maheen, (a 20 to 25-year old housewife from Lahore, Pakistan) described her rationale for deleting specific photos and videos.

"When I open [social media] chat, sometimes my friends send inappropriate videos. Sometimes they send boyfriend photos. Then that will lead to questions from elders like 'where did you go? Who have you been with? Who is that man?' So it is better to delete the chats and avoid misunderstanding."

Families often needed to manage their content histories when sharing with children. Sahana, (a 40 to 45-year old accountant in Delhi, India) described:

"Actually, I don't have any lock on my phone since my son uses my phone. I would never want my son to watch anything that is inappropriate. Sometimes, I receive videos from friends that are vulgar for children, then I immediately delete such videos."

Entity deletion was not isolated to situations where individual integrity or ethics came into question. With the constant possibility of someone examining a phone, entity deletions offered great freedom and agency. Bushra, (a 40 to 45-year old bank employee from Peshawar, Pakistan) explained,

"I have some glamorous photos of myself on my phone. Sometimes I wear a sleeveless top and take photos. If God forbid someone checks my phone then what will happen? So as soon as I take pictures, I save them in my PC and delete from my phone. I don't rely on my phone."

Entity deletions in personalized systems were particularly challenging for our participants to discover and manage, even though they typically are available. Entity deletions in personalized systems were typically invoked through prolonged presses or hidden behind settings that required multiple clicks to find, limiting reach and value to those less familiar with technology. Take, for

example, Shaina (a 35 to 40-year old medical representative in Kanpur, India) who manages how her application's personalized home page looked to co-located others with indirect techniques. She described:

"When I watch a video that is little bit not nice, then I search for 5-6 other videos on different topics to remove it."

Shaina understood that the algorithm learned from prior history and presented personalized recommendations but was not aware of how to signal to the system to remove or dismiss recommendations through the user interface.

For one participant, the inability to control specific content presented by platforms in a public context led to unfortunate circumstances. Nafisa (a 40 to 45-year old faculty member in Dhaka, Bangladesh) recounted how she liked to show video tutorials to engage students, who in turn listened with rapt attention to her lectures. In one such class, when Nafisa opened videos for a lecture, unexpected content was displayed, which led to ridicule and laughter from the students. Not knowing how to immediately dismiss or hide it, Nafisa felt confused, left the class crying, and took the day off work. Better feedback mechanisms over content platforms and user education may positively impact such unexpected loss-of-control situations.

4.3.4 Private modes (IN: 8; PK: 0; BG: 3)

Use of *private modes*, such as private browsing, were restricted to the (1) technology-savvy and (2) censorship-conscious participants. Participants explicitly chose to use private modes for privacy. As Mary (an 18 to 25-year old engineering college student in Bangalore) described:

"I use hidden mode a few times, like when reading the 50 shades of Grey e-book on my phone…."

A majority of our total participants were not aware of what the private modes in their web browsers did or where to find them. One issue was that terms used to refer to private modes were hard to understand among our participants. (Note that in India, only 10% of the population speaks English[11]). Even when advertised, private modes are often associated with 'secret' activities, threatening participants' values of openness as they performed culturally appropriate gender roles. These design issues might help explain why only 11 out of 199 participants used private modes, despite their potential usefulness. When explained as 'a button you press to temporarily browse anything you like, without affecting your recommendations or history,' participants positively appreciated the concept. Our participants foresaw the need for a private mode for a broad spectrum of informational activities, such as medical and sexuality searches, planning activities like birthday surprises, and content activities like watching adult content or intimate chats.

4.3.5 Avoidance (IN: 43; PK: 33; BG: 6)

Certain applications were *avoided* on the phone to prevent questioning or incrimination by co-located household members. For example, 24 participants described that they had a bank account hidden from their husbands, built up over time from small monthly budget remains and salary leftovers. Many participants avoided installing a banking app on their devices, due to low trust in their ability to control the app's visibility[2].

As another example, certain types of digital content or applications were entirely avoided in households with children, like watching gynecological videos, for fear that they would eventually figure out the app lock passwords or pins. As a third example, participants preferred in-person meetings or phone calls for sensitive communications (*e.g.*, about spousal issues or abortion advice), to prevent others from later seeing the conversation (*e.g.*, in chat history), similar to Tibetans in [7]. As Lathika (a 45 to 50-years old, banking professional in Bangalore) noted,

"We just call and talk to each other. Everyone in the [social media] *group knows that the phone is in the midst of the family. So we don't send anything to each other awkward or secretive at any time of the day."*

Exits were a specific type of avoidance described by participants, in which they suddenly closed an application due to contextual sensitivities (*i.e.*, who was around). Participants reported some vivid exits from apps when they unexpectedly saw embarrassing or sensitive content and wanted to avoid social judgment. In one case, Asma (a 40 to 45-year old housewife from Lahore, Pakistan) described that she threw the phone battery out when an inappropriate ad was presented to her, to ensure no one else could see the content or question her morals (she later reassembled it). Sonia (an 18 to 25-year old arts student in Chennai, India) described how she exited an app by locking the screen and closing the app privately later:

"Quite often I am watching something on Internet and suddenly a porn ad or video pops up. I immediately lock my screen in that case and look around if anybody has seen this or not. I then open it again when nobody is around, view it and then delete or close it. My brother and parents would definitely not like the idea of me watching porn."

Such exits do not remove the recorded history of content presented, even though some participants believed they did.

5. DISCUSSION

We summarize key results and discuss design suggestions, open questions, and privacy challenges for technologists to consider for our participants, and which might be relevant in other contexts where device sharing is common and expected.

[1] English or Hinglish, BBC, 27 Nov 2012. http://www.bbc.com/news/magazine-20500312

[2] In the case of India, demonetization in 2017 led to devaluation of 86% of the high-value currency overnight. Six participants were distrustful of installing banking applications, worried about the loss of hard-earned money, both from government decisions like demonetization and from their husbands discovering their balance. Women were among the most affected by the initiative, since they often had cash bills saved for personal and family expenses that was hidden from men in the family (participants reported that men may squander the money if discovered, sometimes for drinking), which was de-valued [54].

5.1 When device sharing is cultural

Our participants experienced culturally-shaped autonomy in their daily lives, which led to specific performative practices around device privacy. While smartphones are often designed to offer individual user experiences, close social relations are often a part of this assumed personal space. Whether it was husbands, fathers, brothers, bosses, colleagues, children, or in-laws demanding access, women often socially cherished or were expected to share access to their devices. Since device sharing is a cultural expectation and value in this region, we expect this phenomenon to continue even as the number of devices increases in the region (indeed, device sharing in India has been documented in HCI and ICTD research from 2009-10 [40,44], when phone penetration was 36% of the population [40].)

While it might be tempting to conclude that the lack of autonomy is problematic when viewing our results from outside the cultural context, our participants had a range of views regarding their limited privacy. Some felt it was acceptable or even welcome for their husbands and brothers to monitor their phones. This occurred when they felt technologically challenged or wanted protection from untoward admirers on social media (see similar views in [17]). Some others felt that they were non-consensually being monitored.

While there were divergent views on how relevant the concept of 'privacy' was to them, all participants developed privacy-related practices. Some practices were more effective in achieving their goals than others, and their sophistication varied based on their technology literacy. The five types of privacy practices they employed —1) *phone locks*, 2) *app locks*, 3) *aggregate and entity deletions*, 4) *private modes*, and 5) *avoidance*—helped them maintain privacy while adhering to the cultural expectation that they should share their mobile phones with their social relations.

Aggregate and *entity deletions* were often perceived as being useful. Participants believed that deletion, if used when no one was looking, enabled them to remove content without anyone else knowing it had been on the phone. Thus, it enabled them to perform openness (a cultural value for many South Asian women) while keeping select information private. This was unlike *phone* or *app locks*, which signaled to borrowers and co-located social relations that something suspicious might be hidden behind the authentication screen. (Note that we were not able to determine if participants had achieved their goal of deleting the content to the point of it being truly undetectable.)

However, having alternatives to deletion were valuable, since deletion was not the best way to achieve all of their goals. Participants had content they wanted to preserve and access on their phones. Moreover, aggregate and entity deletion controls were not always fully discovered by participants with lower technical literacy. In the future, we anticipate that the growing presence of cross-device, cloud-based interactions could pose new challenges for users seeking to understand the impact of content deletions performed on a device.

Phone and *app locks* were used, but participants did not always consider them to be appropriate in intimate settings. The affordances of app locks sometimes led to tensions with social relations, such as "what are you hiding" questions. Phone and app locks were viewed as effective against strangers who might use one's phone temporarily or in the event of a lost or stolen device. App locks also seemed reasonable for keeping nosy colleagues and acquaintances from snooping, and children from accidentally deleting an app or using too much data (provided that they don't learn the app's pin).

We offer the following considerations to help technologists make design choices that empower users who commonly experience device sharing, mediated usage, or monitoring.

5.2 Supporting privacy

5.2.1 Awareness and education

Our research highlights the rich variance in our participants' mental models and adaptations of device and app privacy controls. As Wash writes in his description of security folks models, *"whether the folk models are correct or not, technology should be designed to work well with the folk models actually employed by users"* [50]. Participants in our study would benefit from a better match between their mental models and the functioning of several technologies they use, especially personalized systems and private modes. Results from our study provide a basis for improving awareness and understanding of these features among South Asian women.

Our research also points to an opportunity to improve user education around available privacy features. For example, participants liked the idea of private modes, yet such modes were rarely discovered or used. Promoting such modes in a culturally appealing way could help more users benefit from them.

5.2.2 Content trail management

Most women we spoke to indicated that the ability to delete content (*e.g.*, downloaded images) and behavioral traces (*e.g.*, browsing history) was the most commonly used, powerful, and effective tool for managing their privacy. Participants described how deleting traces offered peace of mind to browse desired content, while avoiding awkward explanations to social relations. However, deletion was often hard for our participants to understand; for example, some did not realize deletion was a two-step process that required emptying the trash (a finding also reported in research on the technology experiences of survivors of intimate partner abuse [30]). Design explorations aimed to improve the discovery of deletion affordances could be valuable, especially for less tech-savvy users. We found that visual affordances for deletions (such as 'X's') worked best with our participants, since some of them had lower literacy. Technologists may consider increasing the power of tools by ensuring they provide both aggregate and entity removal for all user data (see also [30]).

Technologists should also consider the fact that many new technology users may not be aware of the concept of the cloud or that browsing actions are not just one-time actions, but train personalization models that may present recommendations in the future. Software design should consider communicating to users how cloud backups can be pushed as recommendations to users, so they do not implicate them in situations of device sharing.

Lastly, there are interesting future work directions to explore in offering the ability to transfer content from one device to another, which can help users like our participants maintain privacy on their primary device while storing content on a secondary device. It should be noted that many mobile South Asian users are familiar with downloading, storing and transferring media content between devices and memory cards, and much more comfortable with offline media than cloud-based media (offline media are often used for content consumption in low-bandwidth environments, see

[35,42]). Migration and deletion tools could consider the user's desire to keep content backed up in a safe and private place.

5.2.3 Account switching opportunities

For South Asian women, the one-device, one-user model breaks down, encouraging technologists to challenge the assumption that a single application should have a single account (also noted by [20,29] for users in other regions). Yet none of our participants had multiple user profiles on the phone, possibly due to added friction or poor discovery. Specifically, participants in our study (and in prior work [23,29]) noted that account switching in applications is laborious and time-consuming, deterring them from using this functionality unless absolutely necessary. Also, some of our participants had low literacy and most were accessing apps primarily on mobile devices, making easy account switching harder to use. These challenges are promising areas for future work. While account and profile switching can provide private spaces, they present a fairly heavy cognitive task to users. Automation holds promise for more accurate personalization and recommendations in shared use. Machine learning models to classify and differentiate multiple user activities and invoke different experiences may reduce the cognitive load on the user's part to switch accounts or profiles. On a cautionary note, automated learning should take care to avoid misprediction, in order to avoid accidental disclosure to unintended recipients.

5.2.4 Private mode opportunities

Future work may explore the idea of providing private modes within applications or at the device level, to prevent history being left behind. Private modes could ease deletion-related confusion. As an example in India, Hike Messenger, a popular social media application, allows a private mode to hide specific chats that a user wants to keep private (based on their research that Indian young adults live with parents and want to maintain privacy).[23] To improve discovery, private modes may be shown prominently where the feature is more likely to be used, such as in apps that display culturally-taboo content. Alternatively, a single device-level private mode could simplify the experience.

5.2.5 Content hiding opportunities

While a powerful way for our participants to maintain privacy was to delete content or traces (or access content from a private mode), it was often important for participants to keep content on their devices, such as motivational videos, medical documents or emotional messages. In order to support this need, technologists may consider ways to hide content within the user's device ecosystem (content hiding has also been reported to be useful to other sensitive populations [30]).

App locks allow users to protect the content, but increase the risk of incrimination, since locked apps were often viewed as obviously private. Moreover, our study points to how app locks are not reliable in preventing access by people with power over a user (*e.g.*, elders, spouses, and in-laws). They can be useful in preventing children from accessing content, but children are often quick to figure out passwords and pins, which can leave users with no choice but to keep changing passwords (which increases cognitive load).

Regardless of the method used to hide content, a visible indication of hidden content (*e.g.*, a visual lock icon) may cause more harm than good—at least for this population. Users hiding content are often aware of how their behavior could be perceived as incriminating, leading to reduced usage of the feature. We suggest that designs for hiding content carefully consider the value of making it obvious that content is hidden. Additionally, it is important to consider making such valuable invisibility features available free-of-charge.

5.2.6 Algorithm-related opportunities

While many users have become accustomed to personalized content experiences based on prior activities, many are not aware of how to control them. We recommend that technologies employing algorithms provide or continue to provide clear, easy-to-find settings for novice users to control personalized recommendations. Additionally, improving opportunities for females (and other under-represented groups) to provide algorithmic feedback may be useful in making machine learning datasets more inclusive (as the Internet has disproportionately more male users than female users in many South Asian contexts [17]).

We encourage technology designers to consider the social dynamics and implications discussed above for women in South Asia, which could alter gender power imbalances in unexpected and positively transformative ways.

5.3 Culturally appropriate text

Care should be taken to evaluate privacy controls across various cultural contexts of deployment. Technology is often designed with normative assumptions based on Western cultural values suggesting that online privacy and safety is a right. In contrast, some participants did not identify as having "privacy needs", saying *"Privacy is not for me, it's for those rich women,"* or that 'privacy' was a Western value, even if privacy-related practices were prevalent. This perspective should be considered when writing the actual text that is used to discuss privacy and safety experiences in apps and devices that will be used by women in South Asia. For example, invitations to modify privacy settings may be well intended, but may not seem as inviting to women in South Asia. We suggest that the technology community explore how to adequately explain the use cases and value of privacy-related features to audiences around the world, using terminology that is appropriate to them.

6. LIMITATIONS

This paper presents findings from a study we conducted on how women in South Asia who come from a range of occupations and socioeconomic backgrounds perceive, manage and control their individual privacy on shared mobile devices. Future research studies may examine other populations, such as teenagers, families or women in other parts of the world.

[3] In India, an App for Chats and for Keeping Secrets. New York Times. Aug 2014. https://www.nytimes.com/2014/08/26/world/asia/in-india-an-app-for-chats-and-for-keeping-secrets.html

Our approach was qualitative, hence inductive in nature. Common limitations with qualitative studies include recall bias, observer bias, participants self-censoring on sensitive topics, and limitations in the generalizability of results. We are currently deploying a large-scale survey to measure privacy attitudes in South Asia. Another limitation is the triad focus group format, which may have limited participants from opening up on certain topics in the presence of others.

A possible limitation is the cross-comparisons of countries undertaken in this paper. While we are not aware of any other research studies that focus on all the three countries sampled in this study, our claims are comparative and are likely to miss city- or country-specific nuances or depth.

7. CONCLUSION

We presented a qualitative study of how 199 female participants from India, Pakistan, and Bangladesh perceived, managed and controlled their individual privacy when social relations frequently borrowed and monitored their mobile phones. We examined the ways their social expectations were fulfilled through technological and social affordances. We described how participants used five types of practices to maintain their privacy while navigating cultural expectations to share their phones: 1) *phone locks*, 2) *app locks*, 3) *aggregate and entity deletions*, 4) *private modes*, and 5) *avoidance*. We also discussed some suggestions, open questions, and privacy challenges for technologists to consider when designing for contexts where device sharing might be common. We hope that by sharing our participants' experiences and proposing several opportunities for future work, that technologists have new insight regarding how to make privacy more usable for women in South Asia. Such improvements could, in turn, help others, especially in contexts where device sharing occurs and usage is scrutinized.

8. ACKNOWLEDGMENTS

First and foremost, we thank our participants. We also thank Taylor Marable, Asif Baki, Lauren Johnson, Dave Shapiro, Aaron Stein, Ali Lange, Cary Bassin, Francesca Ginexi, Lawrence You, Michael Falgoust, Miguel Guevara, and Thomas Roessler. We thank our SOUPS reviewers for providing feedback on this paper.

9. REFERENCES

1. Norah Abokhodair and Sarah Vieweg. 2016. Privacy & social media in the context of the Arab Gulf. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 672–683.
2. Syed Ishtiaque Ahmed, Nova Ahmed, Faheem Hussain, and Neha Kumar. 2016. Computing beyond gender-imposed limits. In *Proceedings of the Second Workshop on Computing within Limits*, 6.
3. Syed Ishtiaque Ahmed, MD Haque, Jay Chen, and Nicola Dell. 2017. Digital Privacy Challenges with Shared Mobile Phone Use in Bangladesh. *CSCW*.
4. Deena Alghamdi, Ivan Flechais, and Marina Jirotka. 2015. Security Practices for Households Bank Customers in the Kingdom of Saudi Arabia. In *SOUPS*, 297–308.
5. Arjun Appadurai. 1996. *Modernity al large: cultural dimensions of globalization*. U of Minnesota Press.
6. Jakob E Bardram. 2005. The trouble with login: on usability and computer security in ubiquitous computing. *Personal and Ubiquitous Computing* 9, 6: 357–367.
7. AJ Bernheim Brush and Kori M Inkpen. 2007. Yours, mine and ours? Sharing and use of technology in domestic environments. In *UbiComp*, 109–126.
8. Lorrie Faith Cranor, Adam L Durity, Abigail Marsh, and Blase Ur. 2014. Parents' and teens' perspectives on privacy in a technology-filled world. In *Proc. SOUPS*.
9. Kelly Derby, BETH EASTERLING, and others. 2012. Snooping in Romantic Relationships. *College Student Journal* 46, 2.
10. Jill P Dimond, Casey Fiesler, and Amy S Bruckman. 2011. Domestic violence and information communication technologies. *Interacting with Computers* 23, 5: 413–421.
11. Serge Egelman, AJ Brush, and Kori M Inkpen. 2008. Family accounts: a new paradigm for user accounts within the home environment. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 669–678.
12. Serge Egelman, Sakshi Jain, Rebecca S. Portnoff, Kerwell Liao, Sunny Consolvo, and David Wagner. 2014. Are You Ready to Lock? In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (CCS '14), 750–761. https://doi.org/10.1145/2660267.2660273
13. Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. 2017. Digital technologies and intimate partner violence: a qualitative analysis with multiple stakeholders. *PACM: Human-Computer Interaction: Computer-Supported Cooperative Work and Social Computing (CSCW) Vol* 1.
14. Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. 2018. "A Stalker's Paradise": How Intimate Partner Abusers Exploit Technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 667.
15. Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J LaViola Jr, and Pamela J Wisniewski. 2018. Safety vs. Surveillance: What Children Have to Say about Mobile Apps for Parental Control. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 124.
16. GSMA. 2015. Bridging the gender gap: Mobile access and usage in low-and middle-income countries.
17. GSMA. 2018. *The Mobile Gender Gap Report 2018*.
18. Marian Harbach, Emanuel Von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. 2014. It's a hard lock life: A field study of smartphone (un) locking behavior and risk perception. In *Symposium on usable privacy and security (SOUPS)*, 213–230.
19. Eiji Hayashi, Oriana Riva, Karin Strauss, AJ Brush, and Stuart Schechter. 2012. Goldilocks and the two mobile devices: going beyond all-or-nothing access to a device's applications. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, 2.
20. Jeffrey James. 2011. Sharing mobile phones in developing countries: Implications for the digital divide. *Technological Forecasting and Social Change* 78, 4: 729–735.
21. Jeffrey James. 2016. Mobile phone use in Africa: Implications for inequality and the digital divide. In *The Impact of Mobile Phones on Poverty and Inequality in Developing Countries*. Springer, 89–93.

22. Kelby Johnson. 2003. Telecenters and the gender dimension: an examination of how engendered telecenters are diffused in Africa.
23. Amy K Karlson, AJ Brush, and Stuart Schechter. 2009. Can i borrow your phone?: understanding concerns when sharing mobile phones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1647–1650.
24. Neha Kumar. 2015. The gender-technology divide or perceptions of non-use? *First Monday* 20, 11.
25. Ann Light, Ilda Ladeira, Jahmeilah Roberson, N Bidwell, Nimmi Rangaswamy, Nithya Sambasivan, and Shikoh Gitau. 2010. Gender matters: Female perspectives in ICT4D research.
26. Karen Danna Lynch. 2005. Advertising motherhood: Image, ideology, and consumption. *Berkeley journal of sociology*: 32–57.
27. Market Research Society of India. 2011. The New SEC system. Retrieved from http://mruc.net/uploads/posts/b17695616c422ec8d9dadafc1c3eec26.pdf
28. Diogo Marques, Ildar Muslukhov, Tiago Guerreiro, Luís Carriço, and Konstantin Beznosov. 2016. Snooping on mobile phones: Prevalence and trends. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*.
29. Tara Matthews, Kerwell Liao, Anna Turner, Marianne Berkovich, Robert Reeder, and Sunny Consolvo. 2016. 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*, 5921–5932.
30. Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F Churchill, and Sunny Consolvo. 2017. 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*, 2189–2201.
31. Michelle L Mazurek, JP Arsenault, Joanna Bresee, Nitin Gupta, Iulia Ion, Christina Johns, Daniel Lee, Yuan Liang, Jenny Olsen, Brandon Salmon, and others. 2010. Access control for home data sharing: Attitudes, needs and practices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 645–654.
32. Laura L Murphy and Alexandra E Priebe. 2011. "My co-wife can borrow my mobile phone!" Gendered Geographies of Cell Phone Usage and Significance for Rural Kenyans. *Gender, Technology and Development* 15, 1: 1–23.
33. Bonnie Nardi, Ravi Vatrapu, and Torkil Clemmensen. 2011. Comparative informatics. *interactions* 18, 2: 28–33.
34. David Nemer. 2016. "LAN Houses Are for Boys and Telecenters Are for Girls:" CTCs As Gendered Spaces. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development* (ICTD '16), 54:1–54:4. https://doi.org/10.1145/2909609.2909638
35. Jacki O'Neill, Kentaro Toyama, Jay Chen, Berthel Tate, and Aysha Siddique. 2016. The increasing sophistication of mobile media sharing in lower-middle-class Bangalore. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*, 17.
36. Nimmi Rangaswamy and Nithya Sambasivan. 2011. Cutting Chai, Jugaad, and Here Pheri: towards UbiComp for a global community. *Personal and Ubiquitous Computing* 15, 6: 553–564.
37. Nimmi Rangaswamy and Supriya Singh. 2009. Personalizing the shared mobile phone. *Internationalization, design and global development*: 395–403.
38. Nithya Sambasivan and Paul M Aoki. 2017. Imagined Connectivities: Synthesized Conceptions of Public Wi-Fi in Urban India. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 5917–5928.
39. Nithya Sambasivan, Garen Checkley, Nova Ahmed, and Amna Batool. 2017. Gender equity in technologies: considerations for design in the global south. *interactions* 25, 1: 58–61.
40. Nithya Sambasivan, Ed Cutrell, Kentaro Toyama, and Bonnie Nardi. 2010. Intermediated technology use in developing communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2583–2592.
41. Stuart Schechter. 2013. The user is the enemy, and (s) he keeps reaching for that bright shiny power button. In *Workshop on Home Usable Privacy and Security (HUPS)*.
42. Thomas N Smyth, Satish Kumar, Indrani Medhi, and Kentaro Toyama. 2010. Where there's a will there's a way: mobile media sharing in urban india. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 753–762.
43. Cynthia Southworth, Jerry Finn, Shawndell Dawson, Cynthia Fraser, and Sarah Tucker. 2007. Intimate partner violence, technology, and stalking. *Violence against women* 13, 8: 842–856.
44. Molly Steenson and Jonathan Donner. 2009. Beyond the personal and private: Modes of mobile phone sharing in urban India. *The reconstruction of space and time: Mobile communication practices* 1: 231–250.
45. Julie Stephens and others. 2004. Beyond binaries in motherhood research. *Family matters*, 69: 88.
46. Sarah Revi Sterling, Leslie Dodson, and Hawra Al-Rabaan. 2014. The fog phone: water, women, and HCID. *interactions* 21, 6: 42–45.
47. Sharifa Sultana, François Guimbretière, Phoebe Sengers, and Nicola Dell. 2018. Design Within a Patriarchal Society: Opportunities and Challenges in Designing for Rural Women in Bangladesh.
48. Divy Thakkar, Nithya Sambasivan, Purva Kulkarni, Pratap Kalenahalli Sudarshan, and Kentaro Toyama. 2018. The Unexpected Entry and Exodus of Women in Computing and HCI in India. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 352.
49. David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation* 27, 2: 237–246.
50. Rick Wash. 2010. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, 11.
51. Pamela Wisniewski. 2018. The Privacy Paradox of Adolescent Online Safety: A Matter of Risk Prevention or Risk Resilience? *IEEE Security & Privacy* 16, 2: 86–90.
52. Pamela Wisniewski, Arup Kumar Ghosh, Heng Xu, Mary Beth Rosson, and John M Carroll. 2017. Parental Control vs. Teen Self-Regulation: Is there a middle ground for mobile online safety? In *Proceedings of the 2017 ACM Conference*

on Computer Supported Cooperative Work and Social Computing, 51–69.
53. Delanie Woodlock. 2017. The abuse of technology in domestic violence and stalking. *Violence against women* 23, 5: 584–602.
54. 2016. What of the women who hide cash to feed their children or to escape abuse? *Scroll.in*. Retrieved from https://scroll.in/article/821255/note-demonetisation-what-of-the-women-who-hide-cash-to-feed-their-children-or-to-escape-abuse
55. *Census of India, 2011*. Retrieved from http://www.censusindia.gov.in/2011census/F-series/F-1.html
56. *Pakistan Bureau of Statistics*.
57. *Bangladesh Bureau of Statistics*. Retrieved from http://www.bbs.gov.bd/

10. APPENDIX

A. Interview script

Moderator instructions

Work on building a strong rapport, be personable.

Approach intimate topics with care. If the participant is uncomfortable, leave the topic. Offer some examples to help them open up from your own stories

When topics get sensitive, please use your judgement to ask other Googlers to leave. *e.g.*, ask them to check on something, buy batteries etc. so they can leave.

Find private and neutral spaces to chat.

After the core topics of the interview, please ask Googlers to leave so that you can talk about intimate topics freely.

Always ask for consent before the interview. Ask for permission before recording.

Interview script

Hi, thank you for coming here. My name is X and these are Y and Z. We are here from Google.

Today we are conducting research on what it means to be a Pakistani/Indian/Bangladeshi woman and use Internet, smart phones, apps. Everything you know and use daily. This is not an exam, everything you say is going to be helpful to us.

The purpose is to help us understand how to improve technology for women like you. We encourage you to be frank and open, so we can really learn how you are using phones and improve the experience overall. Some of the topics may be a little intimate or personal because we are talking about women. If you are uncomfortable, just let us know.

Everything we discuss today is confidential. Please do not discuss with your friends or family. Anything we discuss today can be used to improve or build new Google products and features.

Could I get your permission to record this interview (video, audio and photos)? It will be stored confidentially and be used for research purposes only. If you feel uncomfortable, just let us know. Any questions?

Grand tour

Intent: to understand their background, life situation, stresses, and context in which they live.
Could you introduce yourself? Name, profession, age.

Whom do you live with? What do they do?

What's a typical day like in your life?

What are your pastime activities?

What do you look forward to doing each day? What do you dislike doing everyday?

What do you wake up worrying about?

Device and Internet mapping

Intent: What is their device/internet technology landscape like, and why? Why did they choose some devices over others? What struggles do they have with technology and internet? What trade offs did they make and why? What is important to them and why?

I'd like to learn about your devices at home.

What devices do you access?

When did you buy your phone?

Why did you buy this model?

How were you able to finance it? / Who funds it?

When did you buy your first ever smartphone? Do you remember why?

What data and Internet plans do you have?

What do you do on your phone?

Access issues

Intent: how does being a woman affect their access to technology and information? Why? What are the barriers they face with getting access to a phone/Internet? What are barriers with using a phone/Internet?

Are there things you want to do with your phone that you are unable to do for any reason?

How much control do you feel you have over your phone?

How do you fund your data plan?

How about credit money (for Internet or calls), time allowed to spend on phone, time allowed to spend online, apps considered acceptable for women...how are these different from the men in your lives?

What apps do you have on your phone?

Which ones do you use the most?

Which ones do you use the least?

How often do you install new apps?

What motivates you to try out a new app?

Who makes the choice on which app to install, *e.g.,* you, husband, friend?

Device and account sharing

Intent: understand what extra technology demands are placed on women (vs men) and how they handle it. What privacy implications does this have, and how do women accomodate (or not) this?

In a given day or week, does your phone get shared with other people? Tell me more (who, why, how long, what access)

Do you borrow other devices in your family or from your friends?

Tell me how you deal with shared use.

Do you have any privacy considerations with leaving traces online?

Do you ever hide stuff from people around you on your phone or online, say parents, in-laws, husbands or children?

Have you ever tried to delete or remove some browsing history, search queries or recommendations so it does not show to anyone else?

Are there times when you wished you could erase what you have done? Or do stuff without leaving a trace? Tell me more (situations, how, when).

Do you use app locker? Tell me more (which app, which apps hidden, situations, instances)

Which apps do you lock? Why?

Do you ever see a need to view history of what you have browsed or done?

What aspects of identity are private and not to be relieved in a closed circle vs. open circle?

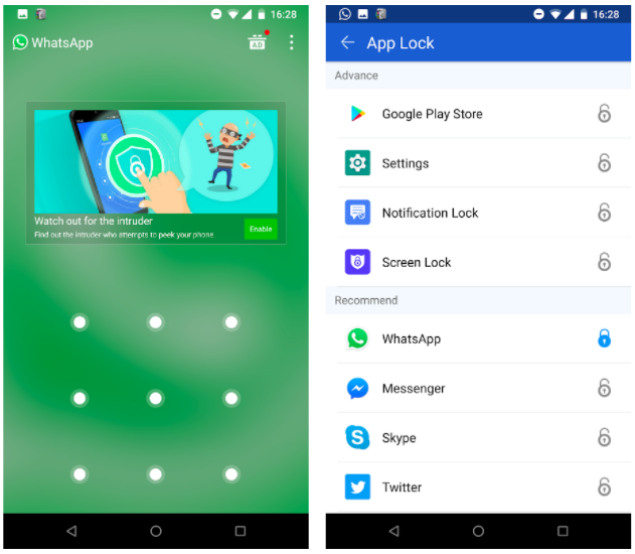
Conclusion

If we write a report based on this interview, what should we highlight?

That brings us to the end of this interview. Do you have any questions for us?

Thank you very much for your time and patience! We learned a lot from you!

Figure 2: Do Mobile's App Lock with over 100 million installs. (Left) Password screen when opening a protected app, (Right) Settings to invoke passwords on apps and folders.



B. Participant table

Table 1. Research sites and locations

Country	N	Locations	SES	Ages	Professions	Education	Tech access
India	103	Chennai (42) Bangalore (16) Kanpur (9) UP villages (15) Delhi (21)	Low (39) Mid (52) High (12)	18-25 (33) 26-35 (24) 36-45 (27) 46-55 (11) >56 (8)	Informal sector (24) Salaried (34) Business owner (6) IT/CS (9) Not employed (12) Student (12) Retired (6)	School dropout (7) High school (31) Undergraduate (45) Postgraduate (14) PhD. (6)	Mobile phone (103) Laptop (37) Tablet (17) PC (15)
Pakistan	52	Lahore (17) Peshawar (8) Karachi (6) Hunza (9) Multan (6) Rawalpindi (6)	Low (12) Mid (32) High (8)	15-20 (9) 20-25 (10) 25-30 (20) 30-35 (10) >40 (3)	Students (11) Not Employed (11) Self-Employed (3) IT/CS (1) Business owner (1)	School dropout (15) High school (7) Undergraduate (16) Postgraduate (13) PhD. (1)	Mobile phone (51) Laptop (30) PC (6)
Bangladesh	44	Dhaka (24) Sylhet (11) Chittagong (9)	Low (9) Mid (21) High (14)	18-25 (1) 26-35 (12) 36-45 (5) 46-55 (3) >56 (6)	Informal sector (6) Salaried (13) Business owner (1) Not employed (3) Student (18) Retired (3)	School dropout (9) High school (22) Undergraduate (6) PhD. (7)	Mobile (41) Laptop (24)

C. Codebook

Top-level category	Definition	Codes
Identity	Identity management of the user, including profiles, self-presentation and settings	Identity management Reputation management Family feedback loops
Co-located privacy	Considerations and management of privacy on shared, mediated or monitored devices	Family sharing Mediation Monitoring Views on privacy Phone locks App locks Micro deletions History deletions Private modes VPN Avoidance Subversion Sensitive content or activities
Access	Ability to use a technology at will, including time, location and social factors	Onboarding Motivations for access Pressures and concerns Money Time Mobility Social perception of access Online activities Fears Mitigation practices Myths of Internet
Online privacy	Considerations and management of privacy on apps, websites and services	Safety concerns Safety practices Information disclosure App-specific privacy models Privacy settings Privacy affordances Privacy violations

"You don't want to be the next meme": College Students' Workarounds to Manage Privacy in the Era of Pervasive Photography

Yasmeen Rashidi Tousif Ahmed Felicia Patel Emily Fath
Apu Kapadia Christena Nippert-Eng Norman Makoto Su
School of Informatics, Computing, and Engineering
Indiana University
Bloomington, IN, USA
{yrashidi, touahmed, fjpatel, ecfath, kapadia, cnippert, normsu}@indiana.edu

ABSTRACT

Pervasive photography and the sharing of photos on social media pose a significant challenge to undergraduates' ability to manage their privacy. Drawing from an interview-based study, we find undergraduates feel a heightened state of being surveilled by their peers and rely on innovative workarounds – negotiating the terms and ways in which they will and will not be recorded by technology-wielding others – to address these challenges. We present our findings through an experience model of the life span of a photo, including an analysis of college students' workarounds to deal with the technological challenges they encounter as they manage potential threats to privacy at each of our proposed four stages. We further propose a set of design directions that address our users' current workarounds at each stage. We argue for a holistic perspective on privacy management that considers workarounds across all these stages. In particular, designs for privacy need to more equitably distribute the technical power of determining what happens with and to a photo among all the stakeholders of the photo, including subjects and bystanders, rather than the photographer alone.

1. INTRODUCTION

In the United States, individuals view privacy as a largely personal managerial task including the selective concealment and disclosure of information about the self to manage relationships with others [54]. According to Altman, people engage in a dynamic 'boundary regulation' process to control access to one's self, which may change depending on the time and circumstance [3]. Through what Goffman calls 'impression management' [28], we try to control the ways others think of us by also managing our 'self presentation'. Individuals are members of multiple groups, and such impression management tends to vary based on the audience and the place [28, 45], e.g., managing one's work versus home personas [53]. Managing privacy thus encompasses a variety of activities both online and offline, utilizing personal socio-technical systems to try to control the accessibility and use of information about us by others [54, 52].

The rise of digital photography and the sharing of high-resolution imagery on social media is not only blurring the line separating the face-to-face and online worlds, it is also forcing us to grapple with a face-to-face world that is, in effect, losing its ephemerality. The implications for impression management are staggering, including an increasing threat to what Nissenbaum calls the 'contextual integrity' of personal information. Existing norms guiding the appropriate collection and dissemination of information are at an ever-greater risk of being broken [55]. The possibilities of what boyd calls 'context collapse' and its associated violations of privacy [55, 8, 9] loom as the captured actions associated with one's social, temporal, and physical context (e.g., photos from a party) are able to be viewed and judged from another – and very different – social context (e.g., an internet search by a potential employer) [73].

Young adults, still in their formative and exploratory years, are often subjected to and impacted by digital photography where smartphone cameras are now integrated with 'one-click' sharing onto social media. Growing up in a world where cameras augmented with seamless social sharing functions are pervasive, perhaps no population has had their privacy more impacted by digital photography than today's young adults. Face-to-face interactions – once a safe, impermanent place for exploratory thought and expression – may now be recorded, altered, reframed, and turned into a persistent online record capable of going viral in seconds, often without the subject's knowledge. Young adults today may thus feel they are being constantly surveilled by their peers. Such pervasive photography raises important questions: What does it mean to be a young adult living in such an environment? What does this mean for an individual's privacy, and for the challenges of trying to control the impression others have of one, now and in the future?

To understand the relationship between privacy, pervasive photography, and social media, we conducted interviews with 23 undergraduates. We focus on three research questions: (1) *What are the everyday privacy concerns of undergraduates with regard to photography and social media?*; (2) *How are undergraduates responding to these concerns?*; and (3) *What privacy enhanced designs might help support*

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

their more immediate and longer term goals given their concerns and current responses to them?

Based on our interviews, we constructed an 'experience model' [37] that represents our participants' experiences with photography on social media, focusing on the 'workarounds' young adults enact to better respond to and manage challenges posed by the current technology. This experience model presents four key stages in the life of a potentially social-media bound photo – when the photo is in its 'potential', 'imminent', 'existent', and 'shared' states. This model expresses the intersubjective nature of what happens to a photo in each of these stages as its fate is negotiated by the relevant actors: photographers, subjects, and bystanders. We build on past research that discusses the privacy concerns of our participants at each of these corresponding stages (e.g., [7, 1, 29]) while adding to existing research by presenting a holistic perspective on digital photography and privacy across time, space, and people. This model highlights, for instance, how the threat of even the potential of being photographed leads to various forms of self-discipline among our participants; how the perception of imminent photography involves split-second reactions from our participants and their friends; how the continued existence and uses of photos may be negotiated at the point of capture; and how the equivalent of 'neighborhood watches' work to mitigate the consequences of shared photos.

This holistic approach to workarounds makes two contributions. First, we provide a model of the constant state of watchfulness that undergraduate students are engaged in to manage their privacy in the face of surveillance from ubiquitous photography and social media. Second, we outline a map from our experience model to designs (a design opportunity map) that highlights how extant and future designs can address users' privacy concerns, both on individual stages and across stages. Taken as a whole, for instance, our model suggests that the power to determine what happens with a (potential) photo at any given stage should be spread more widely across all the stakeholders – not just those who control the taking, altering, and posting of photos.

2. RELATED WORK

In this section, we first describe concerns of surveillance in everyday life and social networking sites followed by a description of privacy management and workaround practices.

2.1 Surveillance Concerns in Everyday Life and in SNS

Due to the pervasiveness of technology and social media, digital records of our daily activities are now a common aspect of everyday life. We are, for example, physically surrounded by closed-circuit television cameras (CCTV) that operate 24/7. We also regularly use digital technologies (e.g., smartphones and digital cameras) to create and preserve fragments of digital information about ourselves, our friends and family members, and even strangers, permanently retrievable by anyone, anywhere, anytime, as long as they have access to the internet (e.g., Facebook, YouTube, and Flickr). Since the technology's inception, privacy advocates have investigated the intersection of privacy and technology. In direct response to the rise of print photography on the society pages of the daily newspaper, Warren and Brandeis wrote their celebrated 1890 article on privacy as the "right to be let alone" to enforce definite boundaries between public and private life [83].

Photography has become a "modern tool of choice for constructing one's identity and conveying it to others" [84, p. 4]. Photos (i.e. film or digital), unlike other media types, are seen as highly context-dependent [24, 77, 35, 34, 66]. Reading and interpreting a photo's context depends mainly on the viewer and could differ significantly from the photographer's intention/context in the moment of taking the photo [77, 78]. Due to our interest in the pervasiveness of personal photography, we will focus our discussion for surveillance concerns around photo recording technology, such as digital and wearable cameras and smartphones.

Various researchers have focused on privacy concerns about pervasive photo recording technology in both online (e.g., Facebook, Snapchat) and offline environments (e.g., public arenas, shared spaces, and private spaces) in different, *specified* stages of a photo's lifespan. Concerns about recording activities and behaviors were usually tied to the location where the activity is being recorded [17, 21, 74, 66]. For example, Choe *et al.* investigated activities that people do not want recorded in their home or shared with other stakeholders with whom they share the home [17]. They found that the most reported activities fell into the categories of self-appearance, intimacy, cooking and eating, media use, and oral expressions. They also found that bedrooms and living rooms were thought to be more private than other locations in the home. Denning *et al.* studied bystanders who may be captured by augmented reality glasses [21] and found various factors affected bystanders' comfort levels and behaviors. For instance, participants were not comfortable when glasses were used during certain activities (e.g., withdrawing money from the ATM), places (e.g., bedroom and bathroom), or if the recorded image conflicted with their desired self-presentation. Such *et al.* [74] found that co-owners (i.e., photographer and subjects) of online shared photos had privacy conflicts around photos of drinking or at parties. Photographers (uploaders) often did not ask for approval before sharing.

Research has also investigated the control, access, awareness, and consent of photo records [66, 6, 35, 34, 48, 51, 11]. For example, Besmer and Lipford [6] investigated users' concerns regarding photo tagging on SNS. Subjects worried about the negative consequences of a photo being seen by a specific social group or presenting them in an unfavorable light. Rashidi *et al.* investigated privacy concerns in mobile instant messaging application and found that some users had concerns about their profile photos being seen by others [61]. People anticipated no covert recording in their home and referred to others (e.g., friends, roommates, and family members) who might surreptitiously try to record them as "interlopers" [48]. Hoyle *et al.* studied the privacy of lifelogging cameras [35, 34] and found that camera wearers were concerned about impression management when managing the sharing of their lifelogs. They also found that sometimes camera wearers chose not to share photos because of objects in the image, activities in the image, and someone in photo (i.e., self or another bystander(s)). Nguyen *et al.* investigated the use of wearable cameras in everyday life [51] and reported on how bystanders wanted to be informed and provide their consent before recording, but felt at the same time

that they have no power and cannot even rely on their social relationships to enforce their preferences, such as asking for deletion or requesting not to share. In another study, although bystanders reported expecting and tolerating recording in public settings, they felt "helplessness" because of the absence of any tools, power, or knowledge necessary to effect a change [48]. Caine *et al.* described how older adults desired control over the collection and transmission of activity data from home monitoring systems [11].

2.2 Privacy Management and Workarounds

In today's networked world, privacy can be conceptualized as a 'dialectic' and dynamic 'boundary regulation' process according to Altman [3], as individuals alter their behavior to disclose or not disclose information to manage their identity and allegiances with others over time [57]. Managing one's personal information, privacy, and identity, specifically within social media, is no longer an individualistic process and is increasingly being seen as a collective process [82, 60, 46, 50, 36, 69, 55] – especially in collaborative settings (such as hospitals) [50] where information is co-owned by the original co-owners (e.g., photographer and subjects in the photo) and/or other extended co-owners (e.g., people who are granted access to the shared content by the original co-owners) [50, 74].

Among various strategies to manage privacy in today's socio-technical systems [44, 71, 85, 18, 10, 72, 6, 47, 19, 79, 16], we focus our study on students' use of 'workarounds', which are behaviors adopted in order to 'get the job done', manage gaps, and enact strategies [49] to maintain their privacy in today's era of pervasive photography. A workaround includes the "work patterns an individual or a group of individuals create to accomplish a crucial work goal within a system of dysfunctional work processes that prohibits the accomplishment of that goal or makes it difficult" [49, p. 52]. For Koppel *et al.* workarounds are "actions that do not follow explicit or implicit rules, assumptions, workflow regulations, or intentions of system designers. They are non-standard procedures typically used because of deficiencies in system or workflow design" [40, p. 409].

Workarounds are mentioned in several different research areas, especially those related to health information technology and organizations. Here, to 'workaround' is to "use computing in ways for which it was not designed or avoid its use and rely on an alternative means of accomplishing work" [27, p. 12]. We are not aware of studies that focus on workarounds related to privacy management in everyday life, certainly in the context of sharing photos on social networking sites. Yet, studying workarounds can provide insight into future improvements to computing systems [2]. Student attempts to manage their individual and collective privacy, and photographs can be categorized into two groups of workarounds: online and offline strategies. Online workarounds are the use of technology in unexpected ways to complete a task. For example, although users could create different 'Friend Lists' to control the visibility of individual posts, the associated costs of doing so (e.g., time consuming and tedious) has instead led many users to create multiple targeted profiles on the same site (e.g., Facebook) [85, 44, 81]. Offline workarounds are used when individuals cannot find a technical tool to support their needs [10, 6, 43, 85, 50, 80]. Besmer and Lipford [6] note that Facebook users modified their behavior both online and offline to cope with the use and popularity of Facebook photo sharing. Users self-censored their physical activities to prevent unwanted photos from being captured and to avoid physical confrontation with photographers for deletion of unwanted images.

Lampinen *et al.* propose another way to categorize workaround strategies [43] which are overlapping strategies to manage privacy and publicness on SNS (i.e., mental, behavioral, preventive, corrective, individual, and collaborative). Although the strategies do not necessarily have to all be workarounds – some of them include the straightforward, intended uses of technology – we can still build upon the workaround strategies in Lampinen *et al.*'s framework. 'Mental workarounds', for instance, include developing interpersonal arrangements to manage disclosure, trusting others to be considerate to one's boundary regulation, and becoming more responsible when posting material on social networking sites [44]. 'Behavioral workarounds' can be further divided into preventive workarounds to avoid unwanted outcomes and corrective workarounds to eliminate or reduce the threat after such an outcome has already occurred. Self-censorship and device (e.g., smartphone) avoidance are 'preventive workarounds'. Interpreting a potentially problematic issue to be non-serious and asking peers to remove content are 'corrective workarounds'. Because of the lack of SNS controls to support collaboration to manage privacy boundary [72, 43, 85, 79, 16, 50, 36, 69], we can consider most of the collaborative workaround strategies as 'offline workarounds' (e.g., asking another person to delete content, asking for approval before disclosing content, and negotiating what is appropriate to share on social networking sites). Murphy *et al.* found that emergency department staff, which are highly collaborative, use workarounds when privacy policies or security mechanisms interfered with their actual work practices. They raised the awareness of the need to improve design to facilitate collaboration endeavors and manage privacy in such environments.

We focus on how students work around technology because of the lack of satisfactory tools to manage privacy, especially for fine-grained tasks, with the goal of better understanding their needs and, therefore, providing design recommendations to suit these needs. Our research confirms many of the aforementioned privacy concerns and workaround strategies, but builds upon these findings by taking a holistic, bird's eye view of the ways a potentially social media-bound photo comes into being and garners the attention of various actors through the photo's 'lifecycle'.

3. METHODOLOGY

We conducted semi-structured interviews with 23 undergraduate students on a large, US college campus from March 2016 to August 2016. Undergraduates are a rich information source [59]: (1) they are likely to use social media and new technology; (2) having just transitioned from high school and simultaneously transitioning to professional life, students are aware of their social and professional images, and are grappling with the management of their individual and collective privacy; and (3) the environment of students (i.e., living together in dorms, social events) create rich contexts within which they navigate such concerns. Students were recruited through flyers placed in common areas, online university classifieds, and emails sent to campus orga-

nizations. In total, 14 students lived in dorms, 4 lived by themselves, 3 with family, and 2 with friends. 15 participants have used Facebook, 17 have used Snapchat, 16 have used Instagram, 13 have used Twitter, and 4 have used Yik Yak. Each participant was compensated $15 USD at the end of the study. This study was deemed exempt by the Indiana University IRB (#1510531315). Screened participants completed an informed consent form at the beginning of each interview. All interviews were audio recorded, transcribed, and de-identified. We employed critical incident techniques [15]; once participants told us stories/incidents, we probed for specific details, allowing participants to control the narrative and help us understand what occurred, from their perspectives. Interviews lasted 43–74 minutes ($M = 59.5, SD = 7.7$), including 12 women and 11 men, aged 18–24, and spanning diverse fields of study.

After the first 18 interviews, we met multiple times to analyze the first third of the transcripts in an iterative approach using open and axial coding [70]. We then discussed the identified themes and developed a draft codebook. Dedoose [20], a web application tool, was used to code and organize data. Using the draft codebook and working in pairs, we coded the remaining transcripts to identify new themes, which were then discussed with the entire team and, if appropriate, added to the codebook. Emergent themes led us to iterate on the interview protocol used in the first 18 interviews to investigate topics discussed by earlier participants. The updated protocol was then used with the last five participants. We then analyzed the newly collected data using the same process and reached theoretical saturation – no new themes were identified in this stage.

Our initial protocol investigated privacy concerns and behaviors associated with the ubiquitous presence of smartphones and social media technology in general. The overwhelming focus on shared photographs in these interviews led us to emphasize privacy concerns and photography for the final five students. In addition to demographic questions, our protocol focused on privacy- and technology-related events that happened face-to-face or online, and which then impacted interactions in the opposite realm as well as evolving attitudes and behaviors around digital photography.

4. FINDINGS: PHOTOS AS THREATS TO PRIVACY

The concern over the long-term effects of digital photography on one's privacy is timely – across SNSs, there is a dizzying array of default settings on how photos persist. Participants expressed how such default archiving with photos have a big impact on one's reputation "because [this photo is] there forever, and it's written [which] can be used as evidence against you" (P21). The persistence of photos on platforms such as Instagram creates a bigger, far more permanent, and less controllable audience than for others like Snapchat:

My friend on her 21st birthday [had] this picture that was entirely too ratchet [slang for crazy] ... Her friends sent her a Snapchat of it ... and [she] didn't realize it was on another site [Instagram] ... [Later] she did see it on [Instagram] and was like, "Come on guys, that's not okay." (P6)

Although P6's friend did not mind sharing the photo on Snapchat, which 'disappears' after being viewed twice, she was shocked to see her photo being posted on Instagram, where persistence was the default. This persistence had direct ramifications for her reputation.

The daily routines of undergraduates involve the creation and sharing of photos by themselves or others (e.g., friends or strangers). Consciously or unconsciously, our participants archived a timeline of their daily life events and activities via the sharing of these photos on different SNS. Photos are open to interpretation yet, due to their seemingly objective nature, provide an *evidentiary chain* to potentially invade one's personal and groups' (e.g., one's sorority) privacy.

4.1 Personal Privacy

Participants were aware of the power photos had over their viewers and felt that captured and shared photos could have serious consequences on their privacy and self-presentation. For example, participants expected others would *judge* them based on these images in a potentially negative and persistent manner. P14 notes below that photos can become a permanent stain on her friend's "record", providing misleading evidence that her friend is a "drunk" girl:

[My friend] came home and was drunk. Somebody was taking a video of her. She was really upset about this, because she didn't wanna be recorded and was really really embarrassed about it ... She asked the person to delete it ... [but] she found out the video wasn't completely deleted. That somebody sent it to somebody else ... She didn't want a video leaked on Twitter ... She didn't want her parents or any older friends, like adults, to see it ... She doesn't want that to go on her record. I don't think it's the general reputation she wants to have is this silly sloppy drunk girl. (P14)

When students felt their actions were not inappropriate, they still worried how others would *misinterpret* photos, taking them out of context and harming their image. P21 mentioned one such negative impression from a photo that might be seen out of its actual context:

[People] just sit there and judge you ... They don't want to understand why you are doing anything you do. The fact is that the picture [of you doing a shot is] there. You look happy in the moment. Whatever you're [doing] nothing else about you matters besides that picture to them because they can't see anything if it's not evident. (P21)

This was a sentiment reiterated by many participants: viewers of a posted photo will not exert the extra effort to truly understand its context. For instance, students cynically expected people to misinterpret photos taken in bars and parties. In these environments, the opportunity arises for misinterpretation of one's drinking behavior (e.g., excessive versus moderate drinking). Objects in a photo (e.g., alcohol bottles and cups) were vulnerable to misinterpretation. Participants were especially worried how photos would be interpreted by their professional peers; they knew that misinterpreted photos could effect future job prospects and curated their social media accounts appropriately.

Surprisingly, participants shared a concern for strangers or acquaintances stealthily capturing and altering original, shared photos with captions or framing them as part of a specific scene to create *memes*. Participants described a shaming trend in which people take photos of strangers,

craft them into memes, and "send them around with a rude caption" (P7), maybe because someone was "dressed not differently, but [in] something really radical" (P18) or because the "kind of things they were doing [was] out of the ordinary" (P18): "One day my jeans ripped really bad and I'm like, 'What am I gonna do?' ... That's like where memes come from. You don't want to be the next meme!" (P6) Participants complained that photos have become a means to deride or ridicule activities that the photographer deems as against current social norms. P4 witnessed such an incident; only in hindsight did she realize what was happening:

[I saw a] person taking a picture of a guy at the library, and ... they were laughing around 'cause he was a heavier-set guy. I didn't think anything of it at that time but after ... there was a trend going around social media of people taking pictures of each other and giving rude comments about it ... I was upset 'cause that person [is] just walking doing their normal stuff. They had no idea what was going on. I feel that's an invasion of privacy on their part. (P4)

The spread of such memes, especially in a college circle, can impact undergraduates' privacy and undermine their reputations [67]. Over half of our participants (N=13) were concerned that recipients would share photos with a wider, 'unintended' audience:

My roommate just went to Mardi Gras so she was dressed up really crazy and probably had too much to drink ... and would Snap individual people, but her friends would take a screenshot of it and upload it to Facebook and would be like, "Oh my gosh my friend's so funny." She was kinda like, "Why are they posting these? I know they're funny, but I send them privately to you for a reason" ... She contacted them and asked them to take it down, but I remember her dad called her and was like, "What is this picture?" He just didn't like what she was doing in the picture and people were commenting, "You're so drunk." (P10)

P10's roommate expected her shared photos over Snapchat to 'disappear' soon after being viewed by the specified receivers, but the unexpected sharing on other social media violated her privacy. This dissemination of privately shared photos by a friend to an unintended audience put P10's roommate in an embarrassing situation and opened a door for others, including her father, to judge her ways of celebrating. P15 explained the difference between a few reshares versus going viral: "[I]f one of my friends post a photo of me doing something stupid and it gets 10 retweets, that's ... not enough to truly hurt my reputation. But if it goes viral then people are knowing me as that guy that did whatever."

4.2 Group Privacy

Students were not just concerned about their personal privacy. Some participants (N=6) sought to maintain their privacy in order to maintain their groups' privacy (e.g., sorority or fraternity, IT department in university, and family) as they see themselves as "an extension" of the group (P6). Lampinen *et al.* [44] calls this 'mutual consideration' – one trusts others to be considerate of their privacy boundary-regulation efforts and puts in the effort to be deemed trustworthy in return. Four participants who were all members of a sorority described how they were required to provide their chapter with all their social media accounts for monitoring, and how particular members of their chapter were responsible for overall monitoring of social media for photos that would harm their organization's image:

We have people that, like, watch all our accounts so if you're ever drinking in your letters [in clothes with the sorority's name on them], that's a big no-no because our nationals can see it, and our chapter will be in trouble. So if you ever post a Snapchat at a bar or a party and you have your letters on ... you'll be asked to remove it ... [T]hat's a position in our house, to look at social media. (P11)

Sororities also created house rules to prevent context collapse as described earlier. For instance, P11 noted that her sorority does not allow red Solo cups in any pictures because "people will automatically think 'alcohol'. "

Aside from more formalized rules at sororities, P19 described being aware that any activities in his photos could be interpreted as being condoned by his organization:

There are times I totally forgot what I'm wearing [my work uniform], and I'm drinking and smoking weed. I'd rather that when people start taking pictures that I changed or something ... I'd rather not [make] people directly tie drugs to the place that I work at. (P19)

P19 knew that wearing his work clothes might impact the organization's image – he would not want his actions interpreted as the organization condoning or encouraging drinking or smoking marijuana.

Even family reputation can be impacted if family members shared a risky photo, as P2 recalled in this incident with her sister: "[My sister] sent an inappropriate [photo] to someone, and we were afraid that it was gonna get posted. Our family doesn't really have that reputation" (P2).

5. FINDINGS: WORKAROUNDS TO MANAGE PRIVACY

Previously, we articulated why undergraduate students worry about their individual and group privacy with digital photography and SNS. This sets the scene for our main focus: when they found technology lacking, participants had to create various *workarounds* (WAs), both individually and collaboratively, in various stages of a photo's lifecycle to ensure safe sharing that would not harm one's privacy.

Our results are framed through an experience model (see Fig. 1) that describes the workarounds through the photo lifecycle in college students' lives, which sheds light on the unique design opportunities for each negotiation point in the model. In this section, we first explain the concept of experience models. Then, drawing from our analysis, we describe four stages of a photo, starting with its potentiality to exist and to the phone when it is shared on social media.

5.1 Experience Models

Due to the integral role of photos in everyday life, researchers from different disciplines have examined the *lifecycle* of photos [39, 65, 14, 12, 13]. These models examine photos from the perspective of understanding the activities (e.g., reviewing and organizing) people perform with their digital photos after capturing but prior to their end use (e.g., sharing) [39] and how the assignment of phases in the mobile photo lifecycle to different platforms affects social discourse around

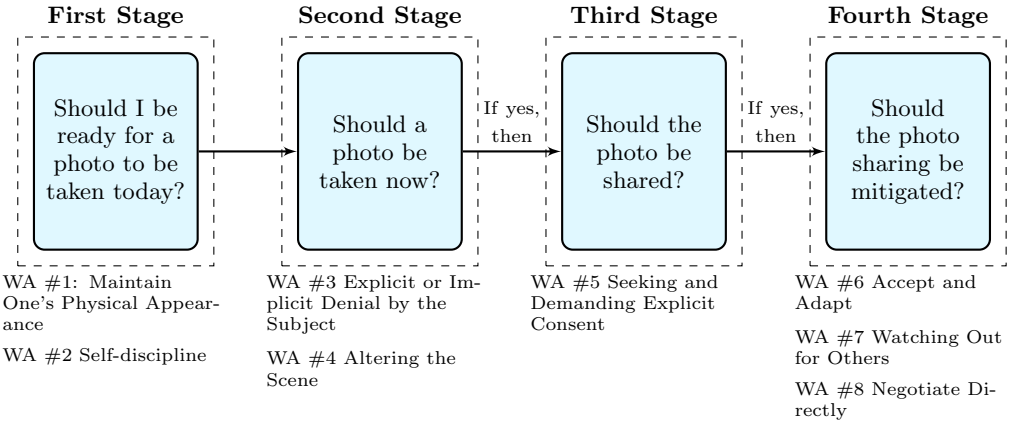


Figure 1: Experience Model: Privacy Workarounds for Surveillance from Everyday Photography

shared photos [65]. Chalfen [14, 12, 13] introduced a 'sociovidistic' framework for the communication activities of films, snapshots, and their artifacts in terms of events (e.g., filming and editing) and components (e.g., participants and topics). Although these models have given us an important temporal understanding of photography and its social use, they have not focused on how we might see the life of photos as intersecting with our privacy in everyday life.

In this paper, we introduce an *experience model* to examine this intersection. Experience models – visual models that facilitate design insights based on ethnographic research – have seen widespread use by ethnographers, especially in the industry [38, 5, 37]. They address the "gap between ethnographic description and the design of technology" [37, p. 2], and are particularly suited to our goal of deriving design insights through qualitative interviews.

An experience model is also "explanatory and developed in a way that has implications for strategic action" [37, p. 2]. Researchers can build upon and refine our model, and can also build other maps on top of our model as they wish. We will present one opportunity map to identify intersections with existing and future designs [37]. Thus, the experience model is both representative and generative.

5.2 First Stage: The Potential to Be Captured in a Photo

All participants remarked upon the pervasiveness of photography. Particularly in public spaces, there is now a constant awareness of the ever-present potential for photos to be (sometimes covertly) taken of anyone, by anyone, in everyday life. This sentiment parallels the feeling of constant surveillance in a modern 'panopticon' [26]. Most participants reported being alert to their environments and actions.

Workarounds at this stage address the following question, "**Should I be ready for a photo to be taken today?**" Our results reinforce Besmer et al.'s findings that people are most concerned with how audiences from their *own* social circles would perceive photos in which they were tagged [6]. However, our participants were also concerned with those to whom they had weaker ties – especially strangers and acquaintances who might perform 'secret captures' (i.e., surreptitiously-taken photos). Participants accepted that they would occasionally appear in strangers' photos as someone in the background; however, many (N=14) specifically feared secret captures (i.e., surreptitiously-taken photos) featuring themselves because of its content (e.g., something embarrassing), what would be done with it (e.g., a caption added to make a meme), and where it might end up (e.g., on popular social media). In other words, the life of the photo would be entirely out of their control:

It's . . . the fact that you don't know where [the photos are] . . . gonna end up. You would hope people won't post them anywhere else. Just that uncomfortable feeling, which is weird to think about because we're . . . photographed every day, without our knowledge, without being aware of it. (P9)

Thus, from the beginning, participants were worried about the potential impact of photos on their privacy.

Our participants sought to enact their views of what constitutes appropriate digital photography to evaluate the potentiality of someone taking a photo of them without their knowledge. Most participants (N=17) were keenly aware that their location (e.g., bar, party, living room, and bedroom) could put them at greater or lesser risk of inappropriate photography to be taken for them. Public places were especially worrisome because of strangers and the difficultly to scan for potential photographers effectively and completely:

[N]ow people don't care about other people's feelings, so they'll just whip their phones out and take pictures of them and make fun of them. I think that's crazy, but that's kind of affected the way I look at other people, or I don't want to eat in public or do other things in public because I'm like, "Oh my gosh . . . That could happen [to me]." I think that has affected my behavior in that aspect. (P4)

For instance, bars were prime locations for cameras to capture people whose guard were down and whose appearance might cause later regret. P14 described always being worried about being watched in bars: "I'm thinking about hanging out with my friends, dancing, having a good time, getting another drink. It's very subliminal, that thought . . . that people are watching you or taking photos" (P14). P18 reiterated that sentiment, saying, "I know that, for the future, I will not want to get really intoxicated in public because I do not want my picture taken" (P18). Small, private spaces,

like parties, were also of concern: "[S]o many people are just holding their phone and prepared to Snapchat some embarrassing moments" (P23). Thus, concerns of photography and space centered around the potential for photos to be especially vulnerable to judgment and misinterpretation.

Participants also told us that vulnerable photography was equally likely in spaces where people expect a higher level of privacy (e.g., living rooms and bedrooms). The very comfort and familiarity one expects in these private, domestic spaces makes the potential of compromising photos even greater. Roommates might document casual conversations, when one is "being kind of goofy" (P19), or general messiness:

When my roommate likes using her phone and [it] is facing toward me sometimes, I would think that she might be taking a picture. Like I don't really care, but there's like a teeny tiny sense of nervousness to it ... I'm not facing a mirror every day; I don't really see myself. Maybe, you know, I'm dressed funny or something and she sees it but I don't. (P5)

The resulting prospect of context collapse in any location and the lack of technical solutions to help with that led students to adopt various WAs to minimize the risk of inappropriate photos, regardless of where they were taken.

5.2.1 WA #1 Maintain One's Physical Appearance

Maintaining good physical appearances (e.g., good taste in clothes, neat hair styles, natural facial expressions, and proper eating habits) was a key (N=14) WA to protecting the propriety of any photos captured of participants. One's appearance was a concern most often mentioned by our women participants (9 out of 12). Men (3 out of 11) also expressed this concern: "[I]f I'm drinking, that's really the only time I'm concerned with [a shared photo], or if I don't think I look good, or if I'm in my pajamas" (P10).

Participants expected friends and family to record and share their activities in events with large attendance, like parties, and they planned accordingly: "I know if I'm going to a social gathering, like a lot of people are gonna take pictures so I make sure to do my makeup and do my hair and I'll wear a new top or whatever" (P6).

5.2.2 WA #2 Self-discipline

To better control their privacy and identity, people engage in self-discipline, modifying their behaviors in both their online [44, 85, 18] and offline worlds [10, 6]. Participants accepted the fact that they would be unable to directly prevent photos taken of them. The majority of our participants (N=15) reported being highly conscious of their surroundings and engaged in *self-censoring* behavior, omitting or curbing their activities to prevent context-dependent photos (i.e., activities open to interpretations). Even in social events where they were surrounded by their friends, some individuals reported carefully regulating their actions to avoid compromising situations that could be captured by others: "I'm very careful about what I say and do around people because I don't want them to share that information with future employers or something like that" (P2). P10 finds herself "just trying to avoid doing things I wouldn't want other people to see unless I was in my apartment or by myself or with my roommates." This decision is often based on having previously witnessed negative consequences for other people who did not censor their behavior:

There's been a time I saw someone really intoxicated, you know, making a scene and yelling, and they probably could have been videoed or something ... That just changed me because I don't really want to be like that ... I will not want to get really intoxicated in public because I do not want my picture taken. (P18)

Participants also reduced the amount of selfies or photos taken of one's friends while drinking or dancing. By self-censoring and continuously monitoring their behaviors, these individual WAs allowed participants to regain some control over their privacy, preventing their unknowingly captured actions from being judged and harming their privacy.

We now turn to the second stage in the photo lifespan, when there is direct evidence that a digital photo is imminent.

5.3 Second Stage: Imminent Photography and Altering the Scene

This stage covers the brief period from when someone is about to take a photo until the photo is actually taken. Our findings show this is a key point that asks both the subject and bystander, "**Should a photo be taken now?**" Of special interest are workarounds that involve implicit and explicit denial of consent to the photo taker as well as the alteration, or arrangement, of the physical scene to be photographed to protect privacy.

5.3.1 WA #3 Explicit/Implicit Denial by the Subject

When subjects felt that a photo shot is imminent, they had to react directly. Some participants (N=6) mentioned explicitly prohibiting photos from being taken, often because they would be depicted unflatteringly. Current appearance was one such reason for not wanting to be photographed: "I was at my friend's apartment the other day, and my hair was a mess, and I didn't want to be in their Snaps so I was kinda like, 'Hey, can you not?'" (P10).

Participants did not always try to explicitly prevent a photo from being taken. Instead, when participants did not want to be in a photo, they (N=4) employed a WA to physically step out of the frame: "I would just tell them if I didn't want to be in [the pictures] or avoid the area where they are taking pictures ... I would either get out of the frame or ... if I wanna be in it, I'll be hyped up and be like, 'Yeah!'" (P10). P9 and P2 also tried to avoid being in strangers' photos because they did not know where these photos would end up:

There's this party I went to where the guy kept taking pictures of his phone. He had like a professional light and everything, and I think there were a few concerns of where the pictures would end up. I mean, you'd think probably just some Facebook, but ... I don't really like my picture being taken if I'm not aware of how it looks. I kind of try not to be in pictures. (P9)

5.3.2 WA #4 Altering the Scene

Drinking (N=18) and dancing (N=7) were the most frequently mentioned activities of concern, and both subjects and photo takers spoke of ways of altering their behaviors when a photo was imminent. With drinking, undergraduates' main concerns were less about underage drinking; instead, they were worried such photos would be posted on social media and affect their relationships:

I wouldn't want people sharing photos of me or posting that I was at a party and drinking alcohol, because, first of all that's illegal for me, and I wouldn't want to be tagged in [it, or] my mom to see it. That would obviously cause a lot of tension between us [me and mom] and would just ruin my reputation 'cause ... I work hard at school. (P2)

Most of the time, a collaborative workaround was needed to ensure a 'clean' shot that would not compromise anyone's privacy in the future. Interestingly, participants (in the role of photo taker) described altering the scene before taking a photo to make a safe photo for a social media post. P4, in the role of photographer, said she always tries to keep alcohol bottles, glasses, or cans out of her pictures:

If I'm out with friends and I do take pictures, I make sure even if they're drinking then I'm not taking pictures with that ... I'm very self-conscious because I don't want [my parents] to think I'm doing those things. If I'm hanging out at a friend's apartment and they do have alcohol, I make sure not to include that in the picture. (P4)

Participants, as the subjects of photos, reported changing or stopping certain actions as soon as they noticed a photo was about to be taken. P1, for instance, explained how she stopped dancing as soon as her friend started to record them: "[W]e were dancing around and acting stupid, and my friend started recording us ... so I stopped dancing ... People would poke fun at me ... I'm just a terrible dancer." (P1)

The questions of whether a photo had the consent of subjects and whether the shot was 'clean' leads to the next stage where the photo exists but has not been posted online.

5.4 Third Stage: The Taken Photo

In this brief stage (before a photo is potentially shared on social media), a photo has now been taken by someone of a subject with (possible) bystanders. With the photo now being a more viable object for putting one's privacy at risk, workarounds turn to the actions that might be taken before possibly posting the photo on social media. Our data reveals one key point: **"Should the photo be shared?"** – actors must decide whether to disseminate the photo. For our college students, sharing usually meant posting the photo on social media. The lack of collaborative tools to facilitate privacy and sharing negotiations force participants to adopt workarounds to enable the safer sharing of photos. Although in most cases participants allowed photos to be shared, they also engaged in workarounds to ensure safer sharing by manually and mutually (with the photo taker) evaluating activities depicted in the photo before sharing the photo.

5.4.1 WA #5 Seeking & Demanding Explicit Consent

When asked how individuals should share photos about others, more than half of our participants (N=16) mentioned the necessity of proactively seeking *consent* from people who are involved in a photo *before sharing* – and sometimes before even taking the photo. Although some social media provide tools to facilitate such consent *after* sharing (e.g., tagging requests), most of them do not offer any tools to approve a sharing request before sharing a photo, which forces participants to adopt WAs. Asking for approval demonstrates responsibility and respect towards others, and the way participants would like others to treat them. Students often do not want to put their friends in harm's way: "If I'm taking Snapchat of someone I will always show them ... I don't want to post anything ... or send it to somebody they're not okay with" (P17).

Mutually negotiated approval allows everyone to be confident that photos will not invade one's privacy. This strategy is not seen as onerous since participants felt many peers were accommodating and reciprocal with this preventive strategy. P3 describes one instance of how this strategy works:

[W]e were finally moved in! Finally roomies! And we took a picture and she was like, "Oh my god my hair looks bad!" So we took a picture like 5 times to get a picture perfect so we could post it on social media. (P3)

Similarly, P10 explains her regular routine: "Before I post pictures I say, 'Hey look at my photos from the night before,' and we'll pass each others phones around, see if there's any good pictures or bad ones and say, 'Hey don't put this up' or 'Let's delete this'" (P10). Here, her friends help her reach an informed decision on what photos would be appropriate and not prone to negatively affecting their reputations.

Collectively approving a photo before posting is not always an option, nor is it the last step of preventing privacy violations. There remains a final stage of a photo open to negotiations: after the photo is shared online.

5.5 Fourth Stage: Photo as Shared Object

In this stage, a photo has already been shared online through social media. Students here asked, **"Should the sharing of the photo be mitigated?"** – how can one reduce the risk (i.e., possible impact on privacy) of a photo that has already been posted online? Interviews showed that students actively negotiate to decide the appropriate WAs to mitigate the risk of an undesirable photo posted on social media.

5.5.1 WA #6 Accept and Adapt

A few participants (N=6) simply accepted the risks associated with their photo being shared online when considering the downsides of confronting social contacts, since that might affect their relationships. Instead, participants *accepted* and *adapted* to the reality of posted photos that did not meet their approval. They mentally brushed off the effect these photos might have on their privacy, or convinced themselves they were overreacting:

I'll send videos of myself being kind of goofy or putting on really silly voices for my girlfriend. I'll ... tend to assume what I'm saying to her is in confidence, and sometimes she'll post one of those silly videos I made just for her onto Instagram, so I'm a bit embarrassed and feel vulnerable ... I try to work on being less self-conscious, so I let it slide pretty much, and she's less self-conscious than I am. (P19)

Avoiding conflict was preferred because participants felt that the negative impact usually is limited, and they would confront others only if there were serious threats to their privacy through misrepresentation that can cause context collapse. P1 explained that most of the time, it is not "worth putting a fuss about. If it was something ... that misrepresented me as a person, I'd probably say something."

5.5.2 WA #7 Watching Out for Others

Students rely heavily on their friends to remain vigilant about any undesirable shared photos that could put their

privacy at risk. They trust their friends to watch each others' backs on social media and warn them of any risks:

It was snowing and [my friend] was wearing open-back moccasins. One of her friends saw on a guy friend's Snapstory ... like, "Who is this girl wearing moccasins?" and she was like, "I know who that girl is." She took a screenshot and sent it to her. She was so weirded out. Cause it was a random guy who made fun of her. (P10)

As we mentioned before, social media accounts of students who live in Greek-letter organizations may be monitored to ensure clean sharing that does not harm the chapter's image. To achieve that, students reported collaborating together to get this mission done and not solely depending on the person formerly assigned to this mission. In some cases (N=4), vigilance translates into direct confrontation with transgressors on behalf of friends, especially with salacious pictures:

I knew a guy who got mad at this girl, and he had a picture of her butt. And on her butt he had written his name in marker. She thought she sent it just to him, but he got mad and posted it on social media ... I actually called the guy 'cause I knew him and I was like, "Look just take that down because that's pretty embarrassing," and it was a really shy girl ... Well, he took it down immediately. (P15)

5.5.3 WA #8 Negotiate Directly

When a shared photo might compromise someone's privacy, participants worked around by engaging in *direct negotiation* to mitigate risk and reclaim context. Most of our participants (N=12) had asked at least one person they knew to delete a posted photo. Since SNS do not provide built-in functionality to facilitate such negotiations, students used workarounds to take matters into their own hands. Negotiation often took place offline, mainly face-to-face. Participants recognized that once someone took ownership of a photo, they had no control over it and could not remove their digital footprint by themselves. They must instead negotiate with the account owner: "On Facebook it's easier ... You can just untag yourself. The concern is that it's still on someone's page, there's not much you can do besides convince them to delete it" (P9).

As mentioned earlier, students are selective in making their requests. They are tactful to both save face and preserve the friendship; they avoid ordering their friends to remove the disputed photo: "I'm just like, 'I look weird take that off' because that's the best way to approach people about stuff like that. Not like coming in really mad and stuff like that" (P22). Students may indirectly and jokingly warn their friends about their concerns: "My friends and I always joke, if we just send each other a stupid face, and we screen shot it we always call each other out like, 'Oh my gosh, I trusted you. It's Snapchat, that's not what you're supposed to do'" (P14). Underlying this humor is serious intent to negotiate a different ending to the photo's shared existence.

6. A DESIGN OPPORTUNITY MAP FOR PHOTO SURVEILLANCE

Our experience model suggests that, in response to the ubiquity of digital photography and SNS, undergraduates now enact a constant, low-level state of watchfulness. It is directed outwardly, toward those in physical and virtual proximity, and, inwardly – toward the self – as these students try to provide as little opportunity as possible for others to subject them to negative, long-term social sanction. This watchfulness is a logical consequence or "harm" of surveillance [63, 68, 56], but not of surveillance by the government or intelligence agencies or even corporations. Rather, it is the consequence of an informal social network of average citizens, including one's own friends and family, all of whom are armed with smart phones and social media accounts. In lieu of technological solutions, students perform workarounds incorporating a dynamic collection of people, practices, rules, devices, apps, and services to manage their privacy. This has an impact on their daily lives, adding to their individual burden of privacy [54]. These students worry that information about themselves and their intimate groups, often captured by themselves or others on social media, might be misunderstood or misused by others when appearing out of its context [55, 9, 73], resulting in negative consequences [4] for them or even losing their reputation [67]. Young adults' watchfulness adds nuance to previous work (cf. Section 2) on individual and collaborative strategies to manage privacy through both online and offline channels.

Moreover, our experience model describes watchfulness as not simply a response to the physical ubiquity of cameras but to the temporal persistence that surveillance via digital photography entails. Just as they did in high school [9], our students manage their boundaries and "presentations of self" [28]. Now, however, the stakes for maintaining students' privacy are indelible and include the loss of scholarships, jobs, and leadership opportunities in addition to the kinds of relationships they might want to have with others. In such a world, where anyone may be instantly, permanently spotlighted, everyone starts to look like the paparazzi.

The necessity of workarounds highlight that pervasive photography and the lack of technology to facilitate their needs forces students to form and respond to models of informational norms of collection and dissemination, both in face-to-face and online interactions. Students predict the vulnerability of photos by examining factors that might likely become misinterpreted (e.g., places). They describe an increasing awareness of invasion of privacy from secret capture by both friends and strangers. They engage in self-censoring behaviors in face-to-face interactions, enact intentional boundary work across different social media platforms, negotiate with each other over shared expectations and practices, and adopt positions of personal and social vigilance to prevent and respond to cases privacy invasion.

We will describe an *opportunity map* – a mapping from our experience model to design directions for researchers and practitioners to pursue [37] – that provides two design approaches to our experience map. First, it provides a way to identify and organize design requirements by temporal stages in the lifespan of a digital photo. With this map, we can readily survey what aspects of privacy management in a stage current designs address and do not address; in particular, it points to the need for designs to address the underlying causes of WAs. These design opportunities are grounded in our findings but are nonetheless speculative and sometimes future-oriented; they are not fully fleshed out solutions to the concerns of photographic surveillance. Second, our map provides designers ways to envision boundary management as not an isolated series of actions but as in-

terconnected. This perspective, for example, suggests new design avenues for dealing with surveillance from shared digital photos that address multiple stages.

6.1 Designs for Individual Stages

Since the emergence of social platforms and mobile cameras, researchers have proposed various mitigation strategies that are applicable to different stages of our experience model; we have arranged these solutions onto our model in Table 1. Social media platforms such as Facebook, Snapchat, and Instagram already support some privacy protection techniques where a subject can, for instance, 'untag' themselves from, or report, an offensive photo (Stage 4). Yet, such photos can still persist and continue to be shared. Much of the existing work has focused on Stages 1 and 3, where the user can specify regions, places, objects, and attires to be identified; these solutions protect their privacy by blurring or tagging photos that have these attributes. Some prototypes allow the user to alter a photo by marking or specifying sensitive areas [62, 76, 41] or setting up a privacy policy [64]. Other prototypes propose wearing additional accessories such as special stickers [75], clothes, and bracelets [42] to blur the subject's face in a captured image. Still other prototypes support Stage 2 by restricting photos in controlled environments.

Most of the existing prototypes support multiple stage interventions with some customizations. However, our experience model identifies a gap in designs that consider workarounds through different photo stages for privacy management. These prototypes are too specific; there is a need for more general-level designs giving more control to the subjects and bystanders, not just the photographers. Designs also need to consider how participants often prefer collaborative and collective strategies to mitigate privacy risks.

Many of the design directions we introduce below involve the integration of different devices and software into the ecosystem of digital photography and SNS. We do not have easy answers on how this will be accomplished but surmise that we will need technical solutions coupled with new policies. Scholars will need to consider how standards and processes can be developed between disparate stakeholders (e.g., software and hardware camera companies). Alternatively, we will need designs that are capable of defending against adversarial systems; such advances may only be practical when the appropriate sensors or algorithms have been researched (e.g., sensors that can detect camera activity). In the next subsections, for each stage, we discuss design opportunities, speculative designs that address these opportunities, and the research challenges of implementing such designs.

6.1.1 First Stage: Designing for Potential Captures

Opportunity: Actors live daily with the ongoing potentiality of a photo harming their privacy being taken with or without their awareness.

Designs that work despite the absence of communication between visible and invisible actors taking photos. There is a distinct lack of communication between the photographer and subject; one does not realize that a photo will be taken in any given moment. New designs would allow photographer and subject-specific systems (i.e., smartphones or tablets) to interoperate with each other to notify subjects about covert attempts to take their photo.

Designs that maintain preferred practices to prevent the creation of photos vulnerable to context collapse. Participants act idiosyncratically based on their own WAs to protect their privacy (e.g., self-disciplining their physical behavior and appearance at all times) from unknown and unobservable capture. Based on self-selected behaviors, designs may remind users to maintain certain workarounds. Such designs are analogous to apps like the 'Drunk Mode' app used by our participants to prevent themselves from taking photos that might affect their self-presentation negatively.

6.1.2 Second Stage: Designing for In-the-Moment Maneuvers and Scene Alteration

Opportunity: Photographers felt it was unnecessary to obtain consent from parties that might be involved when a shot was imminent. It was up to subjects and bystanders to react immediately to evidence of such a capture.

Designs that support in-the-moment maneuvers. College students reverted to face-to-face, in-the-moment WAs because technological solutions to convey their privacy and personal preferences regarding imminent capture were not available. In-the-moment capture requires a time-sensitive solution that communicates preferences to the photographer, whereas the previous stage requires an omnipresent, overseer-type system. Researchers will be challenged to find solutions sensitive to the social nuances of negotiating capture when the photographer and subject are in proximity.

Designs that support socially unobtrusive rejection of capture. These designs would alert subjects that photography was imminent in their area, allowing them to move out of the camera's physical frame. Designs may help subjects and bystanders visualize an active photographer or the path of a camera's focus. Lastly, designs may alert photographers themselves of social, even formal rules of capture tied to a location and/or event (such as a sorority party). This solution, for instance, may require a form of crowdsourcing to label a current location as the site of an occasion with rules.

Designs that evaluate how 'safe' a photo is. Participants reported manually scanning the scene before captures to ensure subjects were safe from potential contextual collapse (e.g., no drinking, no embarrassing dancing, and no Solo cups). Designs should support participants' active, in-person evaluations and allow negotiation over with whom to share the photo – this suggests designs need to support segmentation (i.e., an awareness of multiple photo sharing platforms and their use-cases for particular audiences).

6.1.3 Third Stage: Designing for Photo Negotiation

Opportunity: In this stage, the photo exists, and participants enacted WAs to determine whether it should be shared.

Designs that support in-situ photo negotiation. When negotiating the sharing of a photo, participants asked (or were asked) for consent face-to-face, away from the online world in which the photos would be shared. Participants found it more expeditious to seek consent in-person immediately after the photo was taken and before sharing it. Previous research has highlighted the need to facilitate in-person collaboration over photo sharing in social media sites [72, 43, 85, 79, 16, 36, 50]. However, SNS still lack a negotiation tool to notify involved parties before sharing the photo. Parties

Current Design/ Prototypes	Stage 1	Stage 2	Stage 3	Stage 4
BlindSpot [58]		×		
World-driven access control [64]	×	×		
PrivateEye [62]; PrivacyApp, PrivacyFabric, Privacy Bracelet [42]	×		×	
PlaceAvoider [76]; TagMeNot [75]; ScreeenAvoider [41]	×		×	
Obscuring scene elements [31]; Cartooning [32]; Snapme [33]			×	
Collaborative Privacy Management [69, 36]			×	×
Restrict Others [6]; Facebook; Snapchat; Instagram				×

Table 1: Designs and the stages of the Experience Model they support

need a mechanism that works at the site of capture, in-situ, directly after the group photo to facilitate obtaining consent. Additionally, a remote version would allow sharing to be decided after-the-fact. Researchers will need to investigate solutions that incorporate information such as the identity of people in a photo, one's social network, and the physical locations of relevant parties.

6.1.4 Fourth Stage: Designing to Shield from Consequences

Opportunity: In this final stage, people's privacy has already been compromised. Students had to find WAs with photographers. Those who chose to avoid conflict with the photographer had to accept the risk of unwanted disclosure. A few participants resorted to technical means to protect their privacy, untagging themselves from posted photos.

Designs that mitigate consequences. Participants spoke powerfully of the effects on their reputation from photos that were vulnerable to context collapse. How can we both mitigate damage on reputation and help one recover their reputation from these already-taken photos? Future research should address this under-investigated area.

Designs that support socially acceptable workarounds of photo removal. Direct negotiation is a means by which many participants got photos to be deleted. Scholars will need to address the challenge of supporting negotiation that is tactful or even passive. Designs might allow users to convey that they want a photo to be removed through the use of humor about the content or subject of the photo – this avoids direct conflict between subject and photographer while still communicating that the photo is inappropriate for sharing.

Designs that ease concern about consequences of context collapse. Some participants, after some consternation, simply accepted that context collapse had happened. Uniquely, systems might help users come to a realization that a risky photo may not adversely effect their reputation. Such systems may pose scenarios with similar photos and demonstrate that the consequences were not as dire as they seem now, and that the users should take the photo as a learning moment. Researchers will need to create appropriate exemplars that can form the basis of these scenarios.

Designs that support vigilance (i.e., 'neighborhood watch'). Bystanders sometimes alerted subjects that someone had posted a photo of them on social media without their knowledge. A system that supports such neighborhood watch-type communities would leverage the power and motivation of particular organizations (e.g., Greek communities or college career services) or social groups (e.g., close friends). Such crowd-sourced watching might root out reputation-damaging photos before they propagate widely but would need to avoid inadvertently creating a system for cyberbullying.

6.2 Photo Trajectories: It Takes a Village...

Our experience model visualizes the trajectory of workarounds for managing privacy throughout a digital photo's lifespan. Thus, aside from gleaning what systems must do at each specific stage over a photo's life, a wider, more significant contribution of our model is in highlighting issues germane to the photo's trajectory. Our experience model highlights several concerns that are difficult to see in any individual stage but are eminently visible when we step back and look at the entire process. Our central message here is that for college students – akin to the African proverb, "it takes a village to raise a child" – it takes an entire social group to help manage each others' privacy, thanks to digital photography. This perspective, we believe, is more in-line with what our participants actually do to manage their privacy – they enact collaborative workarounds in individual stages of a photo in service to a long-term, curated representation of themselves. We identify design opportunities and challenges to support this perspective on privacy management to produce photos with a 'healthier', more privacy-sensitive life.

6.2.1 The Power of the Photographer: Empowering the Subject and Bystanders, too

In each stage of our model, we observe that the photographer remains a powerful actor. Photographers have overwhelming power in deciding to digitally capture, alter, and disseminate a photo. In Stage 1, the workaround space is large and untenable, out of the subject's control. The photographer's actions are not moderated, and instead, it behooves the subject to alter their own physical behavior or routines to protect their privacy. In Stage 4, the photographer has already released their photo to the online world where it can be endlessly modified and disseminated, after which the photographer bears no responsibility for their digital progeny. Even in a collective network like Facebook, the only two options for subjects to deal with unflattering photos they are tagged in are to untag themselves and ask the uploader to delete the photo [23], making the subjects feel helpless to enforce their privacy preferences. In Stage 2 and 3 – when the photo is imminent and taken, respectively – the photographer is bound by social conventions to negotiate with those physically around them. Yet, the camera device, perhaps a smartphone, remains under the photographer's control. The photographer may pass their smartphone around for their

friends' review, but it is understood that the owner of the smartphone will be the one pushing the 'delete' button.

The power of the photographer lies in the sites in which both capture and sharing happen – in the hardware and the software possessed by the photographer. We suggest that designs should *dilute* the concentrated power of photography that currently resides with the photographer by spreading it across all interested actors. For instance, we should investigate technical opportunities to give power to subjects and bystanders at all stages of the photo's existence. Such solutions will need to answer difficult questions on how to *work around* conflicts. For instance, we can imagine a user toggling a 'do not share photos of me' setting on their smartphone's OS. If the photographer takes a shot in a tourist spot, with many potential subjects having turned on this feature in their phone, this may ruin the photographer's experience – it will be impossible to frame a photo without an opt-out person in the background. Should we rob the photographer of their power to take a photo for their aesthetic goals? Such solutions could rely on the *propriety* of photographers [34] to honor 'requests' sent by the subjects – even if photographers may have the ultimate power of the veto, technical mechanisms are needed to enable a more seamless negotiation than the current status quo.

6.2.2 Making Past Workarounds Visible

Once a photo has been shared (Stage 4), the user has no idea what the photo has gone through. For example, did all co-owners (photographer and the subjects) of a particular photo approve it to be shared online? With whom? Could past workarounds of a photo be visualized? What if we could see to what degree different actors' decisions allowed the photo to reach its current state? If future designs are able to automate collaborative workarounds of various stages, a system may attach to the photo visualizations that indicate its negotiated nature – key decisions made, by whom, where (physically), and at what stage. Then, the photo would bear the mark of its history. This might look like the functionality for Facebook that allows anyone who can see a published post to see its 'edit history' [22].

By making past workarounds more visible, we can empower users, and even social platforms, to better determine the appropriateness (possibly including the factualness) of the photo in terms of privacy and context collapse. If a photo shows strong vetting, the social platform could prioritize the photo's appearance in contacts' 'newsfeeds', for example. Alternatively, a user may choose to alert a social platform of a poorly vetted photo and/or have an option, themselves, to prevent its further dissemination by refusing to share that photo. Novel technical mechanisms that reveal past workarounds of a photo may thus add assurances to the platform and its users as to how the photo should be displayed or further disseminated. This also interestingly suggests that the *solution* to human workarounds does not lie simply in eliminating them via technology (which can be technically intractable and perhaps unwanted) but rather in rendering them visible – via technology – to the user.

6.2.3 Making Future Trajectories Visible

Not only should systems support making past processes and practices visible in digital photography, they should also intelligibly highlight the possible privacy consequences of photos. Although researchers have suggested that social networking sites need to learn from the online, privacy-preserving behaviors of people [25, 30], our findings suggest a need to *combine data on behaviors in both the face-to-face and online worlds in order to address privacy.* For instance, mobile photo applications now support 'augmented reality' modes, where the camera feed is annotated in real-time. Social platforms can offer a comprehensive suite of tools that include a 'privacy-respecting camera' in Stages 1 and 2 in addition to affordances in other stages. These cameras could overlay the display with indicators of potential context collapse as well as subjects' privacy preferences. Facial analysis of expressions could attempt to predict vulnerability to context collapse. Inference of activities, such as parties and whether such activities constitute grounds for self-censorship, could provide additional data to algorithms designed to reduce context collapse. Importantly, while such analyses and predictions can be performed at later stages by the social media platform, creating a 'privacy sensitive camera' offers unique opportunities for social platforms to tackle privacy at the nascent stages of a photo's life.

7. CONCLUSION

Our study shows that college students are acutely aware of pervasive photography in their lives and how photos taken out of context can impact their privacy. They engage in various 'workarounds' (where technology fails) in an attempt to manage their privacy. Young adults engage in a combination of behaviors at various stages: they know a photo can be taken at any time and adjust their behaviors in case a photo is taken; when a photo is about to be taken, they employ explicit and implicit measures to prevent a photo from being taken; after a photo is taken the photographer and subjects deliberate whether the photo should be posted to social media; and finally, if a photo is shared, friends look out for each other and attempt to remove damaging photos. By reaching theoretical saturation with coding, we believe our findings accurately capture the workarounds undergraduates enact in a world of constant photographic capture and sharing. We, however, warn against generalizing since our participants were in higher education institutions in a particular cultural setting. Future work will further test the validity of our models, perhaps through the use of surveys that will reach a wider, more representative sample.

We organize our findings on workarounds using an experience model, an established framework to facilitate design insights from fieldwork, and present a design opportunity map based on the experience model. This design opportunity map surveys current privacy systems and identifies future design opportunities for privacy management. For instance, it highlights the need for designs to support interoperability between different ecosystems, rejection of imminent photo captures, in-situ negotiation before sharing photos, and easing the psychological anxiety of photo sharing. Importantly, our map provides a holistic framework for design that aligns with the temporal, long-term nature of privacy management. A remit to protect privacy impels us to reflect upon designs that challenge the concentrated power the photographer now wields and render visible the past and future work that make pervasive photography work for, not against, people captured in photos.

8. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under award CNS-1252697. Rashidi is funded by the College of Computers and Information Systems in Umm Al-Qura University, Saudi Arabia.

9. REFERENCES

[1] A. Acquisti and R. Gross. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. In *Privacy Enhancing Technologies*, pages 36–58. Springer, 2006.

[2] S. Alter. Theory of Workarounds. *Communications of the Association for Information Systems*, 34(1):1041–1066, 2014.

[3] I. Altman. The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding. 1975.

[4] L. Andrews. *I Know Who You are and I Saw What you Did: Social Networks and the Death of Privacy*. Simon and Schuster, 2012.

[5] R. Beers and P. Whitney. From Ethnographic Insight to User-Centered Design Tools. *Ethnographic Praxis in Industry Conference Proceedings*, 2006(1):144–154, 2006.

[6] A. Besmer and H. Richter Lipford. Moving Beyond Untagging: Photo Privacy in a Tagged World. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1563–1572, New York, NY, USA, 2010. ACM.

[7] d. boyd. Facebook's Privacy Trainwreck. *Convergence: The International Journal of Research into New Media Technologies*, 14(1):13–20, 2008.

[8] d. boyd. *Taken Out of Context: American Teen Sociality in Networked Publics*. PhD thesis, University of California, Berkeley, 2008.

[9] d. boyd. *It's Complicated: The Social Lives of Networked Teens*. Yale University Press, 2014.

[10] K. E. Caine. *Exploring Everyday Privacy Behaviors and Misclosures*. PhD thesis, Georgia Institute of Technology, 2009.

[11] K. E. Caine, C. Y. Zimmerman, Z. Schall-Zimmerman, W. R. Hazlewood, A. C. Sulgrove, L. J. Camp, K. H. Connelly, L. L. Huber, and K. Shankar. DigiSwitch: Design and Evaluation of a Device for Older Adults to Preserve Privacy While Monitoring Health at Home. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 153–162, 2010.

[12] R. Chalfen. *Snapshot Versions of Life*. University of Wisconsin Press, 1987.

[13] R. Chalfen. Interpreting Family Photography as Pictorial Communication. *Image-based Research: A Sourcebook for Qualitative Researchers*, pages 214–234, 1998.

[14] R. M. Chalfen. Film as Visual Communication: A Sociovidistic Study in Filmmaking. 1974.

[15] E. Chell. Critical Incident Technique. In *Essential Guide to Qualitative Methods in Organizational Research*, pages 45–60. SAGE Publications Ltd, 2004.

[16] H. Cho and A. Filippova. Networked Privacy Management in Facebook: A Mixed-Methods and Multinational Study. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 503–514, New York, NY, USA, 2016. ACM.

[17] E. K. Choe, S. Consolvo, J. Jung, B. Harrison, and J. A. Kientz. Living in a Glass House: A Survey of Private Moments in the Home. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 41–44. ACM, 2011.

[18] S. Das and A. D. Kramer. Self-Censorship on Facebook. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM '13, pages 120–172, 2013.

[19] R. De Wolf, K. Willaert, and J. Pierson. Managing Privacy Boundaries Together: Exploring Individual and Group Privacy Management Strategies in Facebook. *Computers in Human Behavior*, 35:444–454, 2014.

[20] Dedoose, 2016. `http://www.dedoose.com/` Accessed Mar. 1, 2016.

[21] T. Denning, Z. Dehlawi, and T. Kohno. In Situ with Bystanders of Augmented Reality Glasses: Perspectives on Recording and Privacy-mediating Technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2377–2386, New York, NY, USA, 2014. ACM.

[22] Facebook. How Do I Edit a Post That I've Shared From My Page?, 2018. `https://www.facebook.com/help/1376303972644600?helpref=uf_permalink`.

[23] Facebook. What If I Don't Like Something I'm Tagged In?, 2018. `https://www.facebook.com/help/196434507090362?helpref=faq_content`.

[24] L. Fasoli. Reading Photographs of Young Children: Looking at Practices. *Contemporary Issues in Early Childhood*, 4(1):32–47, 2003.

[25] P. W. Fong, M. Anwar, and Z. Zhao. A privacy Preservation Model for Facebook-style Social Network Systems. In *European Symposium on Research in Computer Security*, pages 303–320. Springer, 2009.

[26] M. Foucault. *Discipline and Punish: The Birth of the Prison*. Vintage, 1977.

[27] L. Gasser. The Integration of Computing and Routine Work. *ACM Transactions on Information Systems (TOIS)*, 4(3):205–225, 1986.

[28] E. Goffman. *The Presentation of Self in Everyday Life*. Garden City, 1959.

[29] R. Gross and A. Acquisti. Information Revelation and Privacy in Online Social Networks. In *Proceedings of the 2005 ACM Workshop on Privacy in the Misc society*, WPES '05, pages 71–80, New York, NY, USA, 2005. ACM.

[30] M. Hart, R. Johnson, and A. Stent. More Content-less Control: Access Control in the Web 2.0. *IEEE Web*, 2, 2007.

[31] R. Hasan, E. Hassan, Y. Li, K. Caine, D. J. Crandall, R. Hoyle, and A. Kapadia. Viewer Experience of Obscuring Scene Elements in Photos to Enhance Privacy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '18, pages 47:1–47:13, New York, NY, USA, 2018.

[32] E. T. Hassan, R. Hasan, P. Shaffer, D. Crandall, and A. Kapadia. Cartooning for Enhanced Privacy in Lifelogging and Streaming Videos. *CVPRW*, 1:4, 2017.

[33] B. Henne, C. Szongott, and M. Smith. SnapMe if You Can: Privacy Threats of Other Peoples' Geo-tagged Media and What We Can Do About It. In *Proceedings of the 6th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, WiSec '13, pages 95–106, New York, NY, USA, 2013. ACM.

[34] R. Hoyle, R. Templeman, D. Anthony, D. Crandall, and A. Kapadia. Sensitive Lifelogs: A Privacy Analysis of Photos from Wearable Cameras. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '15, pages 1645–1648, New York, NY, USA, 2015. ACM.

[35] R. Hoyle, R. Templeman, S. Armes, D. Anthony, D. Crandall, and A. Kapadia. Privacy Behaviors of Lifeloggers Using Wearable Cameras. In *Proceedings of the 16th International Conference on Ubiquitous Computing*, UbiComp '14, pages 571–582, New York, NY, USA, 2014. ACM.

[36] H. Jia and H. Xu. Autonomous and Interdependent: Collaborative Privacy Management on Social Networking Sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4286–4297, New York, NY, USA, 2016. ACM.

[37] R. Jones. Experience Models: Where Ethnography and Design Meet. *Ethnographic Praxis in Industry Conference Proceedings*, 2006(1):82–93, Sept. 2006.

[38] S. Jones. Grass Roots Campaigning As Elective Sociality (Or Maffesoli Meets 'Social Software'): Lessons From The Bbc Ican Project. *Ethnographic Praxis in Industry Conference Proceedings*, 2005(1):31–52, 2005.

[39] D. Kirk, A. Sellen, C. Rother, and K. Wood. Understanding Photowork. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 761–770, New York, NY, USA, 2006. ACM.

[40] R. Koppel, T. Wetterneck, J. L. Telles, and B.-T. Karsh. Workarounds to Barcode Medication Administration Systems: Their Occurrences, Causes, and Threats to Patient Safety. *Journal of the American Medical Informatics Association*, 15(4):408–423, 2008.

[41] M. Korayem, R. Templeman, D. Chen, D. Crandall, and A. Kapadia. Enhancing Lifelogging Privacy by Detecting Screens. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4309–4314, New York, NY, USA, 2016. ACM.

[42] K. Krombholz, A. Dabrowski, M. Smith, and E. Weippl. Exploring Design Directions for Wearable Privacy. In *Proceedings of the Workshop on Usable Security*, USEC '17, 2017.

[43] A. Lampinen, V. Lehtinen, A. Lehmuskallio, and S. Tamminen. We're in it Together: Interpersonal Management of Disclosure in Social Network Services. In *Proceedings of the SIGCHI Conference on human factors in computing systems*, CHI '11, pages 3217–3226, New York, NY, USA, 2011. ACM.

[44] A. Lampinen, S. Tamminen, and A. Oulasvirta. All my People Right Here, Right Now: Management of Group Co-presence on a Social Networking Site. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, GROUP '09, pages 281–290, New York, NY, USA, 2009. ACM.

[45] M. R. Leary. *Self-presentation: Impression Management and Interpersonal Behavior*. Brown & Benchmark Publishers, 1995.

[46] E. Litt, E. Spottswood, J. Birnholtz, J. T. Hancock, M. E. Smith, and L. Reynolds. Awkward Encounters of an Other Kind: Collective Self-presentation and Face Threat on Facebook. In *Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 449–460, New York, NY, USA, 2014. ACM.

[47] A. E. Marwick and boyd danah. Social Privacy in Networked Publics: Teens' Attitudes, Practices, and Strategies. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011.

[48] M. Massimi, K. Truong, D. Dearman, and G. Hayes. Understanding Recording Technologies in Everyday Life. *IEEE Pervasive Computing*, 9(3):64–71, 2010.

[49] J. M. Morath and J. E. Turnbull. *To Do No Harm: Ensuring Patient Safety in Health Care Organizations*. John Wiley & Sons, 2005.

[50] A. R. Murphy, M. C. Reddy, and H. Xu. Privacy Practices in Collaborative Environments: A Study of Emergency Department Staff. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 269–282, New York, NY, USA, 2014. ACM.

[51] D. H. Nguyen, G. Marcu, G. R. Hayes, K. N. Truong, J. Scott, M. Langheinrich, and C. Roduner. Encountering SenseCam: Personal Recording Technologies in Everyday Life. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, UbiComp '09, pages 165–174. ACM, 2009.

[52] C. Nippert-Eng. Privacy in the United States: Some Implications for Design. *International Journal of Design*, 1(2), 2007.

[53] C. E. Nippert-Eng. *Home and Work: Negotiating Boundaries Through Everyday Life*. University of Chicago Press, 1996.

[54] C. E. Nippert-Eng. *Islands of Privacy*. University of Chicago Press, 2010.

[55] H. Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

[56] G. Orwell. *Nineteen Eighty-Four*. New York: Harcourt Brace, 1977 [1949].

[57] L. Palen and P. Dourish. Unpacking Privacy for a Networked World. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 129–136. ACM, 2003.

[58] S. Patel, J. Summet, and K. Truong. *BlindSpot: Creating Capture-Resistant Spaces*, pages 185–201. Springer London, 2009.

[59] M. Q. Patton. *Qualitative Research*. Wiley Online Library, 2005.

[60] S. Petronio. *Boundaries of Privacy: Dialectics of Disclosure*. SUNY Press, 2012.

[61] Y. Rashidi, K. Vaniea, and L. J. Camp. Understanding Saudis' Privacy Concerns When Using WhatsApp. In *Proceedings of the Workshop on Usable*

Security, USEC '16, 2016.

[62] N. Raval, A. Srivastava, A. Razeen, K. Lebeck, A. Machanavajjhala, and L. P. Cox. What You Mark is What Apps See. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '16, pages 249–261, New York, NY, USA, 2016. ACM.

[63] N. M. Richards. The dangers of surveillance. *Harvard Law Review*, 126(7):1934–1965, 2013.

[64] F. Roesner, D. Molnar, A. Moshchuk, T. Kohno, and H. J. Wang. World-Driven Access Control for Continuous Sensing. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 1169–1181, New York, NY, USA, 2014. ACM.

[65] R. Sarvas, A. Oulasvirta, and G. Jacucci. Building Social Discourse Around Mobile Photos: A Systemic Perspective. In *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services*, pages 31–38. ACM, 2005.

[66] S. Singhal, C. Neustaedter, T. Schiphorst, A. Tang, A. Patra, and R. Pan. You are Being Watched: Bystanders' Perspective on the Use of Camera Devices in Public Spaces. In *Proceedings of the SIGCHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI '16, pages 3197–3203. ACM, 2016.

[67] D. J. Solove. *The Future of Reputation: Gossip, Rumor, and Privacy on the Internet*. Yale University Press, 2007.

[68] D. J. Solove. *Nothing to Hide: The False Tradeoff Between Privacy and Security*. Yale University Press, 2011.

[69] A. C. Squicciarini, H. Xu, and X. L. Zhang. CoPE: Enabling Collaborative Privacy Management in Online Social Networks. *Journal of the Association for Information Science and Technology*, 62(3):521–534, 2011.

[70] A. Strauss and J. Corbin. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, Thousand Oaks, CA, 1998.

[71] F. Stutzman and W. Hartzog. Boundary Regulation in Social Media. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 769–778, New York, NY, USA, 2012. ACM.

[72] F. Stutzman and J. Kramer-Duffield. Friends Only: Examining a Privacy-Enhancing Behavior in Facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1553–1562, New York, NY, USA, 2010. ACM.

[73] N. M. Su and L. Wang. From Third to Surveilled Place: The Mobile in Irish Pubs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '15, pages 1659–1668, New York, NY, USA, 2015. ACM.

[74] J. M. Such, J. Porter, S. Preibusch, and A. Joinson. Photo Privacy Conflicts in Social Media: A Large-scale Empirical Study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3821–3832, New York, NY, USA, 2017. ACM.

[75] TagMeNot. TagMeNot.info. `http://tagmenot.info/`, 2017.

[76] R. Templeman, M. Korayem, D. Crandall, and A. Kapadia. PlaceAvoider: Steering First-Person Cameras away from Sensitive Spaces. In *Proceedings of The 21st Annual Network and Distributed System Security Symposium*, NDSS '14, pages 23–26, 2014.

[77] M. Thibault and D. Walbert. Reading Photographs, 2006.

[78] J. Thom-Santelli and D. R. Millen. Learning by Seeing: Photo Viewing in the Workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 2081–2090. ACM, 2009.

[79] J. Vitak. Balancing Privacy Concerns and Impression Management Strategies on Facebook. In *Symposium on Usable Privacy and Security*, SOUPS '15, Ottawa, Ontario, Canada, 2015.

[80] J. Vitak and J. Kim. You Can't Block People Offline: Examining How Facebook's Affordances Shape the Disclosure Process. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 461–474, New York, NY, USA, 2014. ACM.

[81] J. Vitak, C. Lampe, R. Gray, and N. B. Ellison. Why Won't you Be my Facebook Friend?: Strategies for Managing Context Collapse in the Workplace. In *Proceedings of the 2012 iConference*, iConference '12, pages 555–557, New York, NY, USA, 2012. ACM.

[82] J. B. Walther, B. Van Der Heide, S.-Y. Kim, D. Westerman, and S. T. Tong. The Role of Friends' Appearance and Behavior on Evaluations of Individuals on Facebook: Are We Known by the Company We Keep? *Human Communication Research*, 34(1):28–49, 2008.

[83] S. D. Warren and L. D. Brandeis. The Right to Privacy. *Harvard Law Review*, pages 193–220, 1890.

[84] J. Winston. Photography in the Age of Facebook. *Intersect: The Stanford Journal of Science, Technology and Society*, 6(2), 2013.

[85] P. Wisniewski, H. Lipford, and D. Wilson. Fighting for my Space: Coping Mechanisms for SNS Boundary Regulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 609–618, New York, NY, USA, 2012. ACM.

Away From Prying Eyes: Analyzing Usage and Understanding of Private Browsing

Hana Habib, Jessica Colnago, Vidya Gopalakrishnan, Sarah Pearman,
Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, Lorrie Faith Cranor
Carnegie Mellon University
{htq, jcolnago, vidyag, spearman, thomasjm, acquisti, nicolasc,
lorrie}@andrew.cmu.edu

ABSTRACT

Previous research has suggested that people use the private browsing mode of their web browsers to conduct privacy-sensitive activities online, but have misconceptions about how it works and are likely to overestimate the protections it provides. To better understand how private browsing is used and whether users are at risk, we analyzed browsing data collected from over 450 participants of the Security Behavior Observatory (SBO), a panel of users consenting to researchers observing their daily computing behavior "in the wild" through software monitoring. We explored discrepancies between observed and self-reported private behaviors through a follow-up survey, distributed to both Mechanical Turk and SBO participants. The survey also allowed us to investigate why private browsing is used for certain activities. Our findings reveal that people use private browsing for practical and security reasons, beyond the expected privacy reasons. Additionally, the primary use cases for private browsing were consistent across the reported and empirical data, though there were discrepancies in how frequently private browsing is used for online activities. We conclude that private browsing does mitigate our participants' concerns about their browsing activities being revealed to other users of their computer, but participants overestimate the protection from online tracking and targeted advertising.

1. INTRODUCTION

Private browsing mode is a feature offered by most major web browsers. These modes promise users an increased level of privacy for their browsing activities. Typically, browsers clear data associated with a user's activities once they close a private browsing window. Though private browsing is an important tool for users, prior work has found that it does not address some major user privacy concerns, nor does it offer privacy protections that many users expect [10, 16, 41, 42]. Furthermore, though users may have privacy concerns regarding their online activities, they frequently fail to navigate privacy decisions to meaningfully address them [1].

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

Prior user studies have examined different aspects of private browsing, including contexts for using private browsing [4, 10, 16, 28, 41], general misconceptions of how private browsing technically functions and the protections it offers [10, 16], and usability issues with private browsing interfaces [41, 44]. A major limitation of much prior work is that it is based on self-reported survey data, which may not always be reliable. In answering surveys, participants may not remember all past activities, may be too embarrassed to report some of their private browsing behavior, or may misinterpret survey questions [23]. Moreover, it is unclear whether users' misconceptions reported in prior work are relevant to users' motivations for engaging private browsing mode, and thus, lead to privacy harms.

Our study builds on prior work to provide a better understanding of how people use private browsing, and identify the gaps that exist between users' perceptions of the privacy protections afforded by private browsing and the reasons they use it. To do so, we analyzed browsing data contributed by 451 participants over a three-year period to the Security Behavior Observatory (SBO), a longitudinal panel study actively collecting data related to privacy and security behaviors from participants' home Windows computers [7, 13, 14, 36]. We supplement this analysis with a survey which explored reasons for using private browsing, and common misconceptions about its actual protections. Our survey was distributed to both SBO and Amazon Mechanical Turk[1] participants so that we could compare our findings with the misconceptions explored in prior work [16], and determine whether our findings hold across two demographically different populations.

Our work contributes the following: 1) We leverage SBO browsing data to explore patterns in private browsing usage, such as how browsing activity differs between normal and private browsing modes. 2) We examine to what degree private browsing activities observed by the SBO differ from those reported in our survey, in order to investigate the impact of self-reporting bias on prior work. 3) We provide insights into why people use private browsing for specific use cases, and explore to what extent misconceptions about private browsing may be harming private browsing users.

Overall, private browsing occurred in only 4% of the 167,128 browsing sessions observed in the SBO, indicating that users likely only switch to private browsing to complete a specific

[1] Amazon Mechanical Turk: https://www.mturk.com/

task. The most common use cases for private browsing include using a service which required a login and performing a search engine query. We observed that websites categorized as adult content constituted a larger percentage of domains visited in private browsing than in normal browsing. Proportionally, participants also conducted searches about sensitive topics and watched age-restricted YouTube videos more frequently in private browsing mode than in normal browsing. We found discrepancies between the private browsing usage reported by SBO participants and that empirically observed, though overall, the most common activities observed were similar to those reported.

Similar to Gao et al., our survey found that although participants had misconceptions about the technical mechanisms behind private browsing, they did find utility in this tool. The most commonly reported use of private browsing was to prevent browsing or search activities from being stored to their device, and potentially being seen by other users. However, we found that some participants overestimated the protections offered by private browsing for the specific use cases they reported, which could lead them to use private browsing in potentially harmful ways. For example, some participants reported that their credit cards were better protected in private browsing mode during online shopping and that their social media activities were hidden from their employers when browsing at work. Identifying such misconceptions is necessary to educate users about the actual protections offered by private browsing, and help them navigate the privacy decisions they make online. We conclude with a discussion about the implications of our findings for browser design and usability.

2. BACKGROUND & RELATED WORK

In this section we present relevant literature and background information related to our study. We focus on prior literature examining privacy concerns of internet users, as well as that studying typical use cases for private browsing. Additionally, we provide a description of private browsing functionalities available in major web browsers currently offered on the market to better highlight user misconceptions observed in our study.

2.1 Privacy Sensitive Online Activities

Prior work has explored users' privacy concerns when they use the internet. Angulo studied users' concerns in "online privacy panic situations" such as account hijacking, leaking of data online, and identity theft. He found that financial harm, embarrassment, and reputation loss were users' primary concerns [5]. A 2013 Pew Research Center survey of 1,002 U.S. adults about online privacy and security concerns and behaviors found that 50% of participants reported being concerned about the amount of personal information collected about them online, and 59% did not believe it is possible to be anonymous on the internet. Most commonly, survey participants expressed a desire to hide their activities from hackers and advertisers, and more participants reported taking steps to avoid advertisers and uncomfortable social situations than to avoid employers or the government from knowing their activities [39].

Prior work has also found that users are willing to take measures to protect their privacy. In the same 2013 Pew survey, 86% of participants had taken steps to remove or hide their online activities, including clearing cookies or browser history and disabling cookies in their browser [39]. An interview study conducted by Kang et al. found that 77% of their non-technical participants reported taking some action to hide or delete their "digital footprints," including using private browsing mode [21].

Other research has highlighted that certain online activities, such as visiting adult content, performing search engine queries, and receiving targeted advertising, may be particularly sensitive. The Pew Research Center found that only 13–15% of their participants reported that they visited adult websites or shared adult content online [15]. Another Pew Research Center survey found that 73% of participants viewed the storing of searches by search engine providers, such as Google, as an invasion of privacy, and 68% opposed receiving targeted advertising [37]. Similarly, a study conducted by Panjwani and Shrivastava analyzing whether users are willing to trade off search personalization for privacy found that 84% of their participants considered at least one of their observed Google searches as sensitive and preferred personalized results for fewer than 20% of these types of searches [35]. In a vignette survey, Rader found that advertisements were a concern related to search engine queries, but participants viewed advertisements in Facebook as even more concerning [38]. However, the findings from an interview study conducted by Agarwal et al. suggest that though users are concerned about tracking on some types of websites, they generally may be more concerned with embarrassment stemming from particular types of advertisements, such as those promoting sexually explicit content, dating sites, or lingerie. The authors also observed that videos viewed by the participants were also often reported as sensitive [3].

Altogether, this prior work highlights reasons why users may choose to use private browsing mode when performing certain activities online. In our study we aim to explore these reasons in more detail to identify whether there are common misconceptions among online users about the protections provided by private browsing. Though users have concerns regarding their online privacy and try to take steps to protect it, they may often make mistakes in doing so [2].

2.2 Private Browsing Functionalities

Each major web browser has a private browsing mode. However, different browsers refer to it using different terms. Google Chrome call private browsing "Incognito Mode" [17], Internet Explorer and Microsoft Edge refer to it as "InPrivate Browsing" [25, 26], and Firefox, Opera, and Safari each refer to it as "Private Browsing" [6, 29, 34]. Generally, when users browse in private browsing mode, their browsing history, logins, form data, and cookies are not stored in their browser. Additionally, in some browsers, the files a user downloads during a private browsing session do not appear in their downloads list [6]. Table 1 summarizes the private browsing functionalities of each browser.

Primarily, private browsing prevents a user's browsing and search activities from being seen by other users of the device. It also provides some protection against online tracking and targeted advertising. Private browsing windows do not replay the cookies and other trackers previously placed by websites in normal browsing mode. Additionally, any

Browser	Browsing History Not Stored	Cookies Not Stored*	Login Info Not Stored	Form Data Not Stored	Tracking Protection Enabled	Downloaded Files Hidden
Safari 11.0.3	✓	✓	✓	✓	Do Not Track	✓
Internet Explorer 11	✓	✓	✓	✓	Do Not Track	✓
Firefox 58.0.2	✓	✓	✓	✓	Disconnect	✗
Edge 41.16299.15	✓	✓	✓	✓	✗	✓
Chrome 63.0.3239	✓	✓	✓	✓	✗	✗
Opera 51	✓	✓	✓	✓	✗	✗

Table 1: Summary of private browsing functions of six major web browsers. Safari, Internet Explorer, and Edge are the only browsers in which downloaded files do not appear in the user's downloads list during private browsing. *Cookies are still exchanged in private browsing, but are not stored beyond the session.

new cookies that were set during the browsing session are deleted once the user closes the window. Firefox and Safari also enable additional web tracking protection mechanisms. In Firefox, some web trackers identified by Disconnect[2] are automatically blocked when users enable private browsing mode [30], while Safari enables Do Not Track, a signal that requests websites not to track users [6]. However, private browsing does not prevent websites from seeing a user's IP address, nor does it hide a user's activities from their Internet Service Provider (ISP).

Prior work in the field of computer forensics has found that artifacts that can identify a user's browsing activities do still remain on the user's computer, even if they use private browsing mode [27, 40]. For example, Ohana and Shashidhar were able to recover usernames, cached images, and URL history from RAM for activities conducted in Internet Explorer's InPrivate mode [33]. A study by Aggarwal et al. highlighted that browser extensions could be particularly privacy violating if enabled in private browsing mode [4]. Soghoian argues that private browsing mode does not offer the level of privacy users expect, and may provide users a false confidence that their activities are truly private [42]. In our work, we aim to further explore whether or not users do have misconceptions about private browsing, particularly concerning the most common activities for which private browsing is used.

2.3 Private Browsing Usage

Prior work has explored how people use private browsing and the misconceptions users have about how it works. A recent survey of 5,710 U.S. participants about private browsing conducted by DuckDuckGo,[3] a privacy-protective search engine, found that 46% of participants had used private browsing at least once on their computer and 43% had used it on a mobile device. The survey also revealed that two-thirds of participants overestimated the privacy protections offered by private browsing, the most common misconceptions being that private browsing prevents tracking from websites and online advertisers, and that it hides searches from search engines [10]. This particular survey population may have been more privacy sensitive than the average online user. However, Gao et al. found similar misconceptions regarding online tracking in a survey study, and reported that many participants did not understand the technical mechanisms behind private browsing. Perceived benefits of private browsing mentioned by participants included that it protects against data collection from malicious sites, reduces page load times, and prevents viruses from being downloaded [16].

An interview study conducted by Shirazi and Volkamer highlighted several usability issues participants noted related to private browsing, including determining whether or not private mode was active in Firefox and Chrome, confusion with browser-provided descriptions of private browsing, and perceptions that it was hard to use or that websites would not be fully functional [41]. Similarly, Wu et al. found in an online study that nearly every disclosure of private browsing provided by major browsers failed to dispel common misconceptions about private browsing mode [44].

Common use cases for private browsing reported in these prior studies include performing "embarrassing" searches, visiting pornographic and dating sites, preventing targeted ads, avoiding cookies, accessing social media, browsing on unprotected Wi-Fi networks, and buying presents [4, 10, 16, 41, 44]. A report from Mozilla's Test Pilot study analyzed timing patterns related to private browsing usage and found that there are spikes in usage at lunch time, the end of the work hours, and after midnight. The report also revealed that most private browsing sessions have a duration of about 10 minutes [28].

Our work builds on these prior studies of private browsing. In our survey, we examine more nuanced use cases for private browsing determined from an analysis of actual user data collected through the Security Behavior Observatory. We also seek a deeper understanding of the threats users are seeking protection from specific to particular use cases. Furthermore, we aim to study users' understandings of the technical mechanisms behind private browsing and identify misconceptions that lead users to believe private browsing is protecting them in ways that it is not.

3. METHODOLOGY

In this section we describe our data collection and analysis methodology. Our study incorporates both empirical and survey data, collected from a longitudinal study, as well as Amazon's Mechanical Turk.

3.1 The Security Behavior Observatory

For our analyses of private browsing behaviors, we used browsing data collected from the Security Behavior Observatory (SBO), further described in Section 3.1.1. We provide additional details about the analyses we conducted for this study in Section 3.1.2.

[2]Disconnect: https://disconnect.me/
[3]DuckDuckGo: https://duckduckgo.com/

3.1.1 Data Collection

The Security Behavior Observatory (SBO) is a longitudinal panel capturing the usage and security behaviors of Windows computer users [7, 13, 14, 36]. The study has been continuously recruiting new participants and collecting data since late 2014 and, as of December 2017, has collected data from over 530 machines.

SBO participants' own home computers are instrumented with data collection software that is designed to collect data automatically with minimal effects on users' normal activities. The SBO data collection software includes system-level components, which allow collection of metadata related to system events, installed software, and other system events and user activities. The software suite also includes browser extensions for Google Chrome, Mozilla Firefox, and Internet Explorer that collect browsing history metadata including URLs and titles of pages visited by the user.

The study's protocol is approved by the Institutional Review Board (IRB) at all universities that work with data from the panel. Each participant completes an enrollment phone call with a member of the research team during which they are assisted with reviewing the study description and terms. During that phone call, participants sign a consent form that explicitly states that all browsing activity and network traffic may be subject to monitoring and that the full contents of web pages may be collected, with the exception of a few highly-sensitive data types.

After the participant has asked any questions they may have and has completed the consent form, a researcher assists each participant with installing the SBO system software, as well as the browser extensions for Google Chrome, Mozilla Firefox, and/or Internet Explorer. The researcher and the participant are connected via both phone and remote session during this entire process so that the researcher can explain each installation step and so that the participant may ask any additional questions that arise. In the case of Google Chrome, an explicit opt-in is required in order for the extension to be able to run and collect data in Incognito mode, so participants either observe the researcher enabling it (and have the opportunity to decline this or ask the researcher for more information) or undergo the step of enabling this functionality themselves.

Participants received $30 for enrolling, as well as $10 per month for continued participation, and are free to leave the study at any time. Given the breadth of the SBO's data collection, special considerations are made for the security and privacy of its participants. After collection, SBO data is encrypted in transmission and stored on hardened servers accessible only to research team members and maintenance personnel using a VPN and two-factor authentication.

We utilized data collected by the SBO's Chrome and Firefox extensions. These extensions collect data related to users' browsing histories, including page URLs, page titles, timestamps, and flags indicating the use of private browsing modes. They also collect a variety of metadata regarding browser configuration and preferences, including information about browser settings and extensions present in the browser. We excluded sessions comprised solely of activity from other browsers from our analysis, as only the Chrome and Firefox extensions report a private browsing flag.

3.1.2 Data Analysis

In this section we describe the analyses we conducted using browsing data collected through the SBO. The data was collected between October 15, 2014 and December 19, 2017 and was contributed by 451 distinct SBO participants. While the SBO has collected data from more participants, for our analysis we excluded participants who had technical issues in reporting browsing data. We also did not include those who solely used a browser other than Chrome or Firefox, as the SBO currently only collects private browsing activity from these two browsers. As the browsing data is stored in a MySQL database, much of our analysis was conducted using MySQL queries. To analyze browsing activity at a session level, or period of continuous browsing activity, the data used in our analyses were labeled with a session identifier. We identified browsing sessions as periods of browsing activity such that there was a gap of at least 30 minutes before the session started and ended. Wang et al. used a similar time-based definition to distinguish browsing sessions, with a threshold of 20 minutes [43].

To analyze the contexts in which private browsing was being used, we manually annotated all sessions containing private browsing data with the use cases listed in Table 3. These use cases were determined by annotating a subset of the private browsing data and finding commonly occurring activities. Definitions of what comprised sensitive browsing and sensitive searches were based off of the responses from Mechanical Turk participants to the survey question "What do you consider to be a sensitive search?" which were analyzed prior to manually coding the entire set of private browsing data. In their responses, participants most frequently mentioned the following categories: 1) pornography or adult content, 2) health or medical content, 3) financial activities, 4) terrorism or crime content, 5) illegal activity, 6) political content.

To ensure accuracy and consistency in coding, two researchers independently coded 25% of the private browsing data, achieving an agreement of $\kappa = 0.81$. All conflicting sessions were reviewed and resolved. The remaining data were coded by a single researcher. After coding the private browsing data, we identified the dominant use case for private browsing for each participant who used it. We determined this to be the use case that the participant did most frequently in private browsing and in at least half of their browsing sessions containing private browsing activity.

We also ran several analyses to compare activities in private browsing with those in normal browsing. The four primary attributes we analyzed were the set of domains visited, categories of the websites visited, search engine queries conducted, and types of YouTube videos viewed. We chose to focus on search engine queries and YouTube activity because conducting searches and streaming video or audio are among the most common use cases for private browsing, in both the observed SBO data and survey responses.

To make statistical comparisons between the two browsing modes, we used Pearson's chi-square tests, with $\alpha = 0.05$. We also report the effect size using the phi coefficient (ϕ), if the comparison was between two binary variables, or Cramer's V (V), if the variables compared had more than two levels, for the significant associations we observed. Both measures are reported on a scale from -1 to 1, where -1 indicates complete negative association and 1 indicates complete

positive association. Only results with at least a small effect (where the association is at least 0.1) are reported, which is an accepted threshold for reporting statistical results [8].

Domains Visited

In comparing the set of domains visited, we calculated the Jaccard similarity coefficient of the distinct domains (e.g., mail.google.com) visited in private browsing with those visited in normal browsing. Subdomains (e.g., chat.google.com and mail.google.com) were counted as distinct domains. Two sets are completely dissimilar (they have no members in common) if they have a coefficient of 0, and are completely similar (they have all members in common) if they have a coefficient of 1 [18].

Domain Categories

To compare the categories of visited websites, we used Amazon's Web Information Service (AWIS)[4] to classify the domains visited. We reduced the number of AWIS provided categories to those that directly mapped to the common use cases for private browsing identified in the manual analysis. These categories are listed in Appendix A. We chose not to use AWIS categories to identify specific use cases of private browsing as some common use cases, such as bypassing a paywall on a news website, can only be determined by looking at the browsing activity in context.

Search Engine Queries

We also used our manual analysis of private browsing activities to identify keywords that corresponded to search engine queries that people may consider sensitive, based on the definition determined from our Mechanical Turk survey responses. The lists of keywords are provided in Appendix B. We developed a script to compare the presence of these keywords in the queries made in both browsing modes. The results of the script were manually reviewed for searches that would not be considered sensitive, and to ensure searches were correctly categorized. Queries to Google, Bing, or Yahoo were identified using the domain and query parameters in the URL of the browsing activity.

YouTube Activity

To compare the types of videos visited in private browsing with those in normal browsing, we developed a script utilizing the PhantomJS WebKit[5] to parse the HTML of pages containing YouTube videos visited by SBO participants. For each video, we analyzed the element with the "unavailable-message" id, which we determined to be a sort of status indicator for the video. This element provided information about whether the video was blocked in restricted mode (indicating some sort of adult or sensitive content), removed for copyright reasons, or removed for violation of YouTube's site policy on sexual, violent, or deceptive content. A list of these codes are provided in Appendix C. Due to the computing resources required for running the script, we analyzed all unique 2,190 videos viewed in private browsing and 3,158 unique videos viewed in normal browsing (a random sample of 5% of all unique videos viewed in normal browsing).

[4]AWIS: https://aws.amazon.com/awis/

[5]PhantomJS: http://phantomjs.org/

3.2 Survey

We conducted a survey to better understand why people use private browsing for certain browsing activities and whether users understand how it works. We administered our survey through both the SBO and Mechanical Turk to collect data from a larger population, and to evaluate the generalizability of our findings by comparing the two populations.

3.2.1 Data Collection

The survey developed for this study contained a combination of open-ended response and multiple choice questions. In the survey, participants answered questions about their background and device configurations, such as devices and browsers they typically use, use of private browsing mode, if they shared their computer with others, demographics, their current cookie policy, any privacy-related browser extensions installed, and privacy consciousness (determined from the IUIPC scale for control, awareness, and collection [24].

Additionally, participants answered two open-ended questions asking what they expected to be protected from while using private browsing, and how they thought it functions. To investigate understanding of private browsing more deeply, the survey also presented 14 statements about technical details related to private browsing and participants selected one of the following options for each statement: "definitely correct," "probably correct," "probably incorrect," "definitely incorrect," or "I don't know." While some questions used more general terms, such as "anonymous," others included more specific wording (like "IP address") so that we could explore the consistency of potential misconceptions.

Participants who indicated ever having used private browsing on their browser were asked how frequently they had performed a list of 13 activities in private browsing mode, based on observed use cases from the SBO, during the past month. We chose to ask about activities in the past month to capture a more accurate representation of regular usage of private browsing, instead of activities that participants may have done only once or twice, a long time ago. Participants were asked a follow up open-ended question asking why they chose to use private browsing for each activity they indicated having done at least once in the past month. The list of all survey questions is included in Appendix D.

We first piloted the survey with 10 local participants who provided detailed feedback, and then conducted two rounds of pilot surveys on Mechanical Turk, with 20 participants each. After each round of piloting we improved the clarity of survey questions and developed additional questions. With approval from our IRB, we advertised this survey on Mechanical Turk as a survey about browsing habits, so as to potentially recruit participants who did not use private browsing. Mechanical Turk users who had a HIT approval rate of over 90% and were residents of the United States, over the age of 18, and not active military were eligible to take the survey. The survey was completed by 309 participants on Mechanical Turk who were compensated with $2.50.

Active SBO participants with Chrome or Firefox browsing data sent by a current version of the SBO browser extension were also invited to participate in the survey. This survey was optional for all SBO participants and did not affect their participation in the longitudinal panel. The survey contained the same questions that were distributed to

the Mechanical Turk sample. In keeping with the approved IRB protocol for optional surveys distributed to this panel, SBO participants received $7.50 for completing the survey. Survey invitations were sent to 344 participants, and 227 participants completed the survey.

3.2.2 Data Analysis

Prior to running our statistical analyses of the survey data, we reviewed for indicators of repeat Mechanical Turk respondents. We removed four responses submitted from IP addresses from which we had previously received survey responses.[6] Thus, we included 305 Mechanical Turk responses in our analyses. We did not have similar concerns about SBO participants completing the survey multiple times, because the SBO infrastructure prevents duplicate responses.

For statistical testing we used $\alpha = 0.05$. In comparisons in which both the independent and dependent variables were categorical, we ran Pearson's chi-squared tests, or Fisher's exact tests if any counts in the contingency table were below five. As in our categorical comparisons of SBO data, we also report the effect size of the association using the phi coefficient or Cramer's V. When testing whether a certain population used private browsing for a particular use case, responses to the question asking participants how frequently did they used private browsing for that use case in the past month were binned as a binary variable where the levels were "never" and "at least once." Responses to this question were confirmed with the participant's answer to the follow up question asking why they used private browsing for that use case. Participants who wrote that they did not use private browsing for that activity, or simply filled in "N/A" were excluded from the count of participants who used private browsing for that use case.

We used a binary logistic regression to test if demographics and privacy sensitivity influenced whether a participant had used private browsing. The independent variables for one regression were the categorical variables age, gender, education, and technical expertise. In another model, we tested for correlations with the IUIPC control, awareness, collection factors. The dependent variable of both regressions was whether or not the participant had used private browsing.

In measuring participants' understanding of private browsing, we used their responses to the 14 statements about the technical details of private browsing. Each participant was assigned a score based on the number of questions they answered correctly, with the "probably" and "definitely" options grouped together. To compare the average score of distinct populations, we ran two-sided t-tests or ANOVA tests, depending on the number of levels in the independent variable. We also used a linear regression to test the impact of demographics on participants' level of understanding, using the same independent variables as the logistic regression.

To analyze our qualitative data, we developed three separate codebooks; the first for the question about expected protections, the second for the question asking how private browsing works, and the third for responses to why private browsing was used for a specific use case. Codebooks were iteratively developed by reviewing a subset of responses to their respective questions for common themes. All responses were coded by two researchers independently, who then reviewed and resolved all conflicts. Our reporting of qualitative data is based on the resolved set of codes.

[6]Though it is possible that multiple people connected to the same network may have completed the survey, and thus had the same IP address, we thought it was more likely that the participants took our survey more than once under different Mechanical Turk accounts. In these cases, we analyzed only the first response submitted.

3.3 Limitations

While our study provides valuable insights into people's usage of private browsing, there are some limitations of our findings. The manual coding of private browsing data could have introduced some errors in our reporting, since there is a large degree of subjectivity in what is considered privacy sensitive. As what constitutes a sensitive activity varies from subject to subject, we cannot be certain that activities coded as sensitive were actually considered sensitive by the participant. Similarly, there may have been activities that participants considered sensitive that were not marked as such during the coding process. This limitation also impacted our analysis of search engine queries conducted in both browsing modes. We attempted to limit this subjectivity by identifying specific categories that survey participants indicated they considered sensitive.

Another limitation of the SBO is that it collects data only from Windows users. Additionally, our study analyzed browsing activities only conducted in Google Chrome and Mozilla Firefox. It is possible that the browsing habits of MacOS users, and users of other browsers, differ from the activities we observed in this population. However, we believe our findings still offer valuable insights into how people use private browsing in their daily lives.

Some of our findings are also impacted by the same limitations as prior work using self-reported methods. As discussed, some of these limitations include the misreporting of prior activities and misinterpretation of survey questions. We attempted to mitigate these potential issues by conducting multiple rounds of piloting and iterating our survey based on the feedback received after each round.

Our study also utilizes two convenience samples, neither of which are representative of the general population. However, Mechanical Turk has proven to be a valid source of high-quality human subjects data [22], and has been used successfully in prior privacy research (e.g., [12, 32]). Considering the consistency of reported behaviors across our two, demographically-different samples, we believe our study provides value in understanding how this important privacy enhancing technology is being used.

4. RESULTS

In this section we report findings from browsing activities observed in the SBO and our survey data. Participants were consistent in their usage of private browsing, and generally used it for practical reasons, such as logging into an account without leaving credentials on the computer, as well as for privacy-sensitive activities. Though there were some inconsistencies between observed and reported private browsing behaviors, overall the most common activities matched across the two data sources. We observed that participants were primarily concerned with their activities being revealed to other users of their device, but also desired protection from web tracking and targeted advertising.

4.1 Demographics

Demographics, displayed in Table 2, were significantly different between our two participant groups. The SBO population had a wider age distribution, with 10% of participants reporting to be age 65 or older. Additionally, the SBO group was significantly more educated, and a larger percentage were technical which was defined by ever holding a job or receiving a degree in computer science or any related technology field. The SBO also contained a larger proportion of females (all $p < 0.05$). Mechanical Turk participants were found to be somewhat privacy conscious, based on the IUIPC metrics for control, awareness, and collection factors, while those in the SBO were less privacy conscious.

We did observe some demographic differences in whether or not a participant had used private browsing within both participant groups. Male Mechanical Turk participants were more likely to use private browsing than females ($p = 0.01, \phi = 0.2$); 95% of males reported using private browsing but only 86% of females did. From the SBO survey, those age 45 and older reported using private browsing less than younger participants ($p < 0.001, \phi = 0.5$); 84% of those under 45 had used private browsing compared to only 39% of those older than age 45. Additionally, 93% of technical SBO participants had used private browsing, compared to 66% of non-technical participants, which was also a significant difference ($p < 0.001, \phi = 0.3$).

A significantly larger portion of participants from Mechanical Turk (91%) had used private browsing compared to participants in the SBO (73%), ($p < 0.001, \phi = 0.3$). We also found that Mechanical Turk participants reported using private browsing significantly more frequently than SBO participants ($p < 0.001, \phi = 0.2$). From Mechanical Turk, 28% reported that they had used private browsing at least half of the time in "the past week" (i.e., the week immediately prior to the survey being administered) on their computer and 23% had used it at least half of the time on their mobile device. In contrast, only 16% of SBO participants used it at least half of the time on their home computers and 15% used it at least half of the time on their mobile device.

Neither the participant's primary browsing platform nor operating system of their main home computer impacted whether and how much they used private browsing, in either the Mechanical Turk and SBO populations. Similarly, we found that having a shared computer did not correlate with more usage of private browsing.

4.2 Patterns in Private Browsing Usage

Of the 451 SBO participants whose browsing data was used for this analysis, 184 (41%) had used private browsing at least once. Overall, private browsing occurred in only 4% of browsing sessions captured by the SBO.

4.2.1 Use Cases for Private Browsing

Table 3 displays the results of our manual coding of 6,327 private browsing sessions. Though adult browsing and other sensitive activities were observed in a substantial proportion of private sessions, they were, surprisingly, not the most common use cases. The most common activities were using a service which required a login (38% of sessions) and performing a search query (33%). Activities that did not fall into a specific use case were categorized as "general browsing," which occurred in 37% of private sessions.

Looking at the dominant use case for which our participants used private browsing, 18% of participants most commonly used it for viewing adult content, 15% used it for general browsing, and 11% used it most commonly to log into an account. However, 22% had no discernible dominant use case. This indicates that the majority of private browsing users are generally consistent in their usage of private browsing.

4.2.2 Private vs Normal Browsing Activities

Next, we examined in more detail the differences in browsing activity between normal and private browsing modes. Among 167,128 total observed sessions, 96% contained only normal non-private browsing. Over 3% of sessions contained a mixture of private and normal browsing, and about 0.5% of sessions contained exclusively private browsing. Sessions containing private browsing comprised 6% of the total browsing sessions collected from observed private browsing users.

We found that, on average, "mixed" browsing sessions that contained a combination of private and non-private browsing sessions were longer than other sessions, with an average duration of approximately 1 hour and 44 minutes. Sessions containing only non-private browsing had an average duration of approximately 43 minutes, while sessions containing only private browsing had an average duration of approximately 23 minutes. On average, normal browsing sessions contained 73 page visits, while sessions conducted only in private browsing contained 40. The average mixed session contained 175 page visits, 34% of which were performed in private browsing windows. This suggests that typically, users switch to private browsing mode to accomplish a task and switch back to normal mode to resume their browsing.

We found the distribution of the browsers used in normal browsing to be significantly different than those used in private browsing ($p < 0.001, V = 0.2$). In normal browsing 65% of participants used only Chrome, 7% used only Firefox, and 31% had used both. However, in private browsing 83% used Chrome, 10% used Firefox, and only 7% used both browsers, indicating that some users of both browsers have decided to use one or the other for private browsing.

The set of domains visited in private mode was found to be dissimilar to those visited in normal browsing, with a Jaccard similarity coefficient of 0.02. The distribution of website categories between normal and private browsing was also found to be significantly different ($p < 0.001, V = 0.1$). Of the most common AWIS categories, email, news, portal, shopping, and social media domains comprised a larger proportion of domains visited in private browsing than in normal browsing. Financial, health, political, search, software, and streaming domains comprised a roughly equal proportion. We observed that 6% of all distinct domains visited in private browsing were categorized as an adult content website, while only 1% of domains were such in normal browsing.

The searches conducted in private browsing were significantly different than those conducted in normal browsing ($p < 0.001, V = 0.1$). Altogether, 16% of searches conducted in private browsing were categorized under a sensitive category, while only 2% of searches were such in normal browsing. The most prominent sensitive search categories in private browsing were searches for adult and health-related content. Searches for adult content comprised of 12% of all private browsing search queries, but only made up 0.5% of

Gender			Age			Education			Tech Expertise			IUIPC (average)		
	MTurk	*SBO*		*MTurk*	*SBO*		*MTurk*	*SBO*		*MTurk*	*SBO*		*MTurk*	*SBO*
Female	43%	61%	18-24	9%	32%	High School	16%	3%	Expert	16%	25%	Control	5.8	4.4
Male	55%	38%	25-34	58%	32%	Some college	20%	19%	Non-Expert	84%	75%	Awareness	6.2	4.9
Other	.3%	.4%	35-44	20%	11%	Trade School	2%	2%				Collection	5.6	5.8
No answer	1%	1%	45-54	9%	8%	Associates	13%	6%						
			55-64	4%	7%	Bachelors	40%	37%						
			65-74	1%	8%	Graduate	8%	34%						
			75-84	0%	2%	No answer	1%	.4%						
			No answer	0%	.4%									

Table 2: Demographic breakdown of our 305 Mechanical Turk participants and 227 survey participants from the SBO. A smaller proportion of SBO participants are male and have technical expertise, compared to the Mechanical Turk population. SBO participants are also more varied in age, more educated, and less privacy sensitive, as measured on the seven-point IUIPC scale.

Use Case	% of Private Sessions Activity was Observed	% of Private Browsing Users Who Did Use Case	% of Private Browsing Users - Dominant Use Case
Log into service	38%	57%	11%
General browsing	37%	66%	15%
General searches	33%	61%	6%
Access adult content	24%	49%	18%
Streaming video/audio	19%	41%	5%
Visit social media	15%	35%	3%
Shopping	12%	42%	5%
Adult-content searches	12%	42%	1%
Sensitive browsing	8%	33%	3%
Sensitive searches	5%	30%	0%
Look up someone's name/profile	3%	25%	1%
Pirate content	1%	7%	1%
Bypass news limits or ad-blocking detection	0.9%	5%	2%
Sensitive shopping	0.6%	10%	0%
Other	2%	11%	1%

Table 3: Summary of private browsing usage in the SBO, displaying the percentage of private browsing sessions in which participants used private browsing for that use case, the proportion of private browsing users in the SBO who used private browsing for each use case, as well as the percentage of private browsing users for which the use case was their dominant reason for using private browsing. About 22% of participants had no discernible dominant use case.

normal browsing queries. Health-related searches were 3% of private browsing searches but only 0.4% of normal searches. The distribution of the types of YouTube videos viewed in private browsing also was found to be significantly different from that viewed in normal browsing ($p < 0.001, V = 0.1$). Proportionally, three times as many videos viewed in private browsing were removed for violating the website's policy on nudity and sexual content and twice as many had content warnings indicating age restricted content. However, overall, these videos made up fewer than 5% of YouTube videos viewed in private browsing. Other videos tagged as infringing or graphic content occurred in roughly equal proportions.

4.3 SBO Observed vs Reported Activities

In this section we provide a comparison of the private browsing activities reported by SBO participants in their survey responses and those empirically observed by the SBO software, so that we can better understand the limitations of prior work utilizing only self-reported data. We find that there were discrepancies between the activities reported by participants and those observed by the SBO, which suggests that participants over-reported on the survey, or performed private browsing activities on other devices. However, the overall activities in the two data sources were similar, indicating that self-reported data is still a valuable means to study research problems in this area.

4.3.1 Usage of Private Browsing

As stated in Section 4.1, 166 (73%) of SBO survey participants reported that they had used private browsing mode in the past. However, only 101 (61%) of these participants had private browsing activities observed by the SBO. Some of these discrepancies are due to participants using private browsing on a non-SBO configured device. Of the participants for whom private browsing activity was reported but never observed, 58% also reported that they had used private browsing in the past month to browse or log into their account on a computer they did not own. 62% of these participants had reported using private browsing on their mobile device in the past week.

Thirteen (6%) of the SBO survey participants had reported that they had never used private browsing mode, even though the SBO software reported private browsing activity coming from their computer. Three of these 13 participants appeared to have opened a private browsing window once, perhaps accidentally, and did not actually perform any activities in private browsing. Six of the 13 participants had three or fewer private browsing sessions, most of which in-

cluded an account login. One explanation for these sessions could be that someone else may have briefly borrowed the SBO participant's computer. The last four participants had between nine and 44 private browsing sessions with various browsing activities, including visits to adult websites. This suggests that very few of our participants intentionally misrepresented their lack of private browsing usage.

4.3.2 Private Browsing Use Cases

Our survey participants were asked about the activities they did in private browsing during the past month. There were 21 participants from whom the SBO collected private browsing data from within the 30 days prior to their survey responses, and nine use cases for which we could make direct comparisons between the two data sources. Table 4 displays the discrepancies in the reported and observed private browsing usage of these 21 participants.

Overall, there were discrepancies in the specific activities participants reported doing in private browsing and those they were observed doing. Perhaps surprisingly, participants from the survey were *over*-reporting, rather than under-reporting, their private browsing usage compared to the measurements. Averaged over the nine use cases, only 40% of participants who reported using private browsing for a use case were also observed using it for that purpose within the 30 days prior to their response. Some activities, such as using private browsing to bypass a paywall or ad-blocking detection and pirate content, had particularly large discrepancies. For most private browsing activities compared, the overall total number of participants who reported using private browsing for that activity on the survey was similar to that observed in the SBO.

When considering the entire population of SBO participants, observed behaviors were similar to those reported among the top use cases for private browsing, as shown in Table 5. Conducting searches, accessing adult content, and logging into an account were the most prominent activities in both the observed and reported data.

4.4 Conceptions of Private Browsing

In this section we describe the reasons our participants use private browsing and their understanding of the privacy protections it offers. Participants were most concerned about their browsing and search activities being saved to their computer. Other reasons for using private browsing were to protect their account credentials and personal information. Overall, participants demonstrated a lack of understanding about the technical functions of private browsing, and had misconceptions consistent with those found in prior work.

4.4.1 Reasons for Using Private Browsing

In their responses to the open-ended question asking what they expected to be protected from during private browsing, participants were primarily concerned about their browsing history, cookies, and search activities being saved to their device. Specific threats participants frequently mentioned included other potential users of their computers, tracking by websites or search engines, or targeted advertising. Concerningly, 12% of SBO participants and 5% of Mechanical Turk participants expressed that they expected private browsing to protect them from malicious attacks, such as malware and being hacked, highlighting a serious misconception.

Participants also had various reasons for using private browsing in particular use cases, some of which included misconceptions. Of the 144 Mechanical Turk participants and 86 SBO participants who used private browsing for online shopping, 24% of these Mechanical Turk participants and 20% of these SBO participants expressed that they thought private browsing protected their credit card or other private information. 14% of both these populations stated they used private browsing to shop for gifts. Another perceived benefit was avoiding price discrimination while shopping for an item or booking airline travel, which was mentioned by 17% SBO participants and 4% of Mechanical Turk participants who shop online using private browsing. One participant explained, "[private browsing] lets me think I am seeing 'real' prices for tickets/items instead of prices generated for me based on my preferences or visits to competitors' websites."

The primary reason for using social media in private browsing was to access social media profiles or look up someone without it being associated to their account. Some participants also thought that private browsing hides their social media activity from their employers (e.g., "I just get on social media very quickly to access and to see was going on, but again I do this at work and we're not supposed to do that though"), which is not an actual protection it provides.

12% of SBO participants and 9% of Mechanical Turk participants who used private browsing for streaming video or audio stated that they did not want their video recommendations to be impacted, which was the most common reason cited after general privacy concerns. One participant explained, "I don't want my browsing history dictating what videos I might want to watch." Four participants from Mechanical Turk and one from the SBO also mentioned reduced load times when streaming content.

Participants who used private browsing on their computers to log into a service, such as their email, most frequently mentioned that they wanted to protect their passwords or private information. Additionally, 14% of these SBO participants and 9% of these Mechanical Turk participants reported that they used private browsing because they had multiple accounts for a service, and they did not want to log out of their account in their normal browser.

Across all use cases, feelings of privacy or security were mentioned in 11% of Mechanical Turk responses and 10% of SBO responses. A participant captured this sentiment stating that they use private browsing to conduct sensitive searches for "Privacy mostly, I don't know how much more secure it is but it makes me feel better." Some participants also mentioned usability benefits. We observed that 34% of SBO participants and 24% of Mechanical Turk participants who use private browsing to access content with ad-blocking detection specifically mentioned that they switched to private browsing to avoid turning off their ad-blocker.

4.4.2 Technical Understanding of Private Browsing

We also asked participants to describe how private browsing worked. Nearly half (47%) of SBO participants and 60% of Mechanical Turk participants correctly conveyed that browsing history was not stored after the session had ended. Many other responses indicated that private browsing did not permanently store other information types such as cookies, login information, and form data. However, 17% of SBO survey

Use Case	Total Reported	Total Observed	% Reported, Not Observed	% Both Observed & Reported	% Observed, Not Reported
General searches	15	10	40%	60%	10%
Access adult content	14	9	36%	64%	0%
Bypass paywall or ad-blocking detection	10	1	90%	10%	0%
Log into service	8	11	25%	75%	45%
Sensitive searches	12	9	42%	58%	22%
Shopping	5	4	60%	40%	40%
Visit social media	7	5	57%	43%	40%
Streaming video/audio	6	7	50%	50%	57%
Pirate content	5	0	100%	0%	NA
Any private browsing usage	21	21	0%	100%	0%

Table 4: Summary of the discrepancies between the observed and reported private browsing activities for 21 participants who sent browsing data to the SBO in the 30-day period prior to their survey response. Of the 14 participants who reported accessing adult content in private browsing, five (36%) were not observed using private browsing for this purpose, while nine (64%) had observed visited to adult websites. All nine participants observed using private browsing for visiting adult content reported their usage.

Use Case	% of PB Users - MTurk	% of PB Users - SBO
General searches	77%	76%
Sensitive searches	71%	64%
Access adult content	66%	52%
Log into service	60%	54%
Shopping	50%	52%
Streaming video or audio	44%	39%
Visit social media	42%	43%
Bypass ad-blocking detection	42%	35%
All browsing	41%	31%
Using a computer they don't own	40%	54%
Bypass news limits	34%	39%
Log into service from a device they don't own	33%	45%
Pirate content	25%	15%

Table 5: Summary of private browsing usage reported by Mechanical Turk and SBO survey participants, displaying the percentage of participants who reported using private browsing for that use case at least once in the past month.

participants and 6% of Mechanical Turk participants indicated they were not sure how private browsing worked.

Responses to this question also revealed a variety of misconceptions about the technical mechanisms behind private browsing. Some responses indicated that private browsing protected their computer's identity, such as their browser version and operating system. Others thought private mode enabled encryption of their browsing activities (e.g.,"history gets more encrypted so that it's not as accessible"). A couple of participants casted doubts that it offered any protection.

Participants in both survey groups, on average, correctly answered between eight and nine of the 14 technical questions about private browsing. These questions also revealed participant misconceptions. One of the most glaring misconceptions indicated as correct by 22% of both Mechanical Turk and SBO participants was that private browsing prevents the browser from sending any cookies to websites. In reality, websites can still place cookies in the browser during a private browsing session but they are deleted after the session has ended. However, an even more alarming misconception is that private browsing allows for browsing the web anonymously, which was answered incorrectly by 39% of both Mechanical Turk and SBO participants. Additionally, 39% of SBO participants and 26% of Mechanical Turk participants thought that private browsing clears all browsing history from their computer after they close the browser window. This is also not correct, as only history from the private browsing session is cleared.

In both survey populations, those who had used private browsing mode answered one or two more questions correctly, on average. As seen in Table 6, the largest gaps in understanding between users and non-users of private browsing were related to information exchange between the user's computer and another entity, such as the ability of the Internet Service Provider to see browsing activity and the computer's IP address being shared with websites. Demographics were not correlated with understanding in the Mechanical Turk survey population, but females and those older than 65 were observed to have answered fewer questions correctly in the SBO survey population. In both survey populations, higher privacy awareness, measured by reported cookie policy, presence of a privacy-related browser extension, and the IUPIC control, awareness, and collection factors, did not correlate with a better understanding of private browsing.

Our results are in line with those observed by Gao et al [16]. Participants in their study showed a similar awareness that browsing history and cookies are deleted in private browsing mode, and desired to keep their activities private from other users of their computer. They also demonstrated similar misconceptions as participants in our study, such as private browsing can block all tracking from websites and will prevent viruses and advertisements.

5. DISCUSSION

Our study accomplishes three goals: investigate how people use private browsing, learn if there are discrepancies between reported and empirically-measured private browsing behaviors, and determine whether private browsing offers users the security and privacy protections they expect to receive. We analyzed a combination of empirical data from the SBO, and survey data from the SBO and Mechanical Turk.

Technical Understanding Question	% of Users Who Answered Correctly		% of Non-Users Who Answered Correctly	
	MTurk	*SBO*	*MTurk*	*SBO*
Private browsing clears my browsing history for that session from my computer after I close the browser window	89%	85%	81%	69%
Private browsing does not save my login information after I end that session.	87%	84%	77%	64%
Private browsing clears most cookies for that browsing session from my computer after I close the browser window.	84%	81%	88%	69%
Private browsing clears all the information that I fill into forms in that session from my computer.	83%	74%	62%	54%
Private browsing blocks ads on the websites I visit.*	73%	71%	54%	54%
Private browsing does not allow my Internet Service Provider (e.g., Comcast, Verizon) to see which websites I visited during that session.*	66%	69%	38%	33%
Private browsing blocks some tracking by advertisement and social media companies.	62%	60%	85%	67%
Private browsing prevents companies from targeting ads to me based on my browsing history from previous private browsing sessions.	61%	62%	77%	64%
Private browsing does not allow websites to get my computer's IP address or any information about my web browser or computer.*	61%	58%	31%	28%
Private browsing prevents companies from targeting ads to me based on any of my previous browsing history.	60%	49%	77%	62%
Private browsing causes the information I send to websites to be encrypted.*	55%	50%	35%	20%
Private browsing allows me to browse the web anonymously.*	51%	52%	31%	40%
Private browsing clears all my browsing history from my computer after I close the browser window.*	27%	42%	12%	31%
Private browsing prevents my browser from sending any cookies to websites.*	24%	26%	0%	13%

Table 6: Percentage of correct responses by users and non-users of private browsing to the 14 technical understanding questions. Statements marked with a "*" are a false statement about private browsing, while all others are true.

Distributing the survey to two populations, especially ones with different demographics, allows us to consider the generalizability of our results. The Mechanical Turk population was younger and likely more technically savvy than the SBO group. Additionally, Mechanical Turkers, on average, reported higher privacy concern on the IUIPC scale compared to the SBO population, and have been found to be more privacy conscious than the U.S. population as a whole [20]. These two factors likely contributed to why Mechanical Turk participants reported using private browsing more frequently than the SBO participants. Despite the differences in the amount of private browsing usage, the top activities performed in private browsing were the same across both populations. This suggests that the most common activities for which private browsing is used may be universal.

Overall, we observed a variety of activities for which people use private browsing, including log-ins to Internet services and search engine queries. Though there were disparities in the usage reported by SBO participants and that which was observed, the most common private browsing activities were the same across both data sources. Lastly, we found that some participants use private browsing for purposes that do not match with the actual protections it provides.

5.1 Usability and Design Implications

We observed that the typical pattern for private browsing usage is that users start a private browsing session for a specific task, and then switch back to normal browsing mode. This could be due to usability reasons, as users might enjoy the convenience of different functions of their browser, such as password auto-fill or browser extensions. Another explanation is that users realize that the protections offered by private browsing, such as hiding activity from other users or avoiding targeted ads, are diminished if they leave their private browsing window open. Perhaps ironically, some users, especially those of shared computers, may intentionally use normal browsing mode for some of their activities to throw off suspicion about their browsing habits. To better support this usage pattern, browsers could implement functions that automatically close private browsing windows after a certain amount of time, similar to how online banking sites automatically log off users after several minutes of inactivity.

Another usability reason for which people use private browsing is to log into a secondary account on their computer without having to log out of their first. However, it is unclear why this behavior is as prominent as it is, since major online services, such as Google, allow users to link their accounts and be logged into multiple accounts at once. It could be that participants may be unaware of this functionality, or that it is not implemented on many websites they use. Another possibility is that our participants prefer to keep their multiple accounts unlinked.

Participants also cited other reasons for using private browsing related to convenience. For example, many participants choose to use private browsing as an alternative to turning off an ad-blocker browser extension on websites that use ad-blocking detection. This indicates that users might find these interfaces too confusing to be able to efficiently disable it to access content.

Some participants also reported using private browsing because they experienced reduced page load times. Certain browsers, such as Firefox, may run faster in private browsing, compared to normal browsing, as browser extensions are disabled by default. Firefox also blocks certain trackers in private browsing, which could also allow pages to load faster. While this aspect of private browsing is not currently advertised by major desktop browsers, it may become more prominent in the future, as some mobile apps such as Firefox Focus already mention this benefit in their description [31].

Recent work has found landing pages for private browsing to be ineffective for dispelling certain misconceptions [44]. Our findings support the changes to private browsing disclosures recommended by the authors, such as directly stating that IP addresses can still be collected by websites. Additionally, we suggest that browsers clarify that cookies are still used in private browsing, but those placed in the browser during private browsing will not be saved beyond that session.

Our study did not comprehensively examine whether users prefer other privacy enhancing strategies over private browsing mode. While we did not find a correlation between private browsing usage and the usage of privacy and security related browser extensions, it is possible that some tools, such as Tor, lead people to use private browsing less frequently. To explore this further, future work could analyze the use of privacy enhancing strategies at an eco-system level.

5.2 Reliability of Self-Reported Methods

In comparing the observed and reported data for the SBO population, a larger proportion of participants reported using private browsing than were observed using it. Many of these participants could have used private browsing on devices not monitored by the SBO, such as their mobile device, as 62% reported doing in the past week, or on someone else's computer, which 58% reported doing in the past month. Some may have used it prior to joining the SBO. Alternatively, we might be observing a form of the Hawthorne effect, such that participants may have unintentionally reported behaviors that align with their interpretation of the study's goals – in that case, affirming more security- and privacy-concerned behavior than they actually evidence.

On the other hand, very few participants whose computer sent private browsing activity to the SBO reported on the survey that they had never used private browsing mode. Additionally, all of the participants who were observed accessing adult content in private browsing reported that activity on the survey. This seems to indicate that people are willing to report some behaviors truthfully on a survey, even if it requires the revelation of activities some may find private or embarrassing to disclose.

Our findings highlight that there are limitations to both empirical and self-reported methods for studying behaviors such as private browsing. Though empirical data collection, like that implemented by the SBO, can provide ground truth for users' activities, it is very difficult to capture everything they do online, as people tend to use multiple devices. While self-reported methods can capture information about all the activities a user does, they suffer from the biases discussed earlier. Studies should utilize both types of methods to maximize coverage and minimize bias.

5.3 Is Private Browsing Enough?

For many users, private browsing functionality matches the privacy protections they expect. Participants most commonly reported using private browsing to hide their activities from other users of their computer. Interestingly, usage of private browsing was found to be independent of whether or not a participant had a shared computer. In their qualitative responses, those who did not typically share their computers frequently referred to rare occasions in which someone might use their computer. Despite having some protections, users should be aware that there is still privacy risk to their private browsing activities. Though private browsing does not permanently store browsing data that is easily accessible to other users, the browsing activity of a prior user could still be potentially seen if their private browsing window was left open, or if they had logged into an account, such as Google, which synced their browsing activity to their browser.

Another common threat participants seek protection from is tracking by websites or search engines. Private browsing does provide a degree of protection against web tracking, as some tracking information, such as cookies are not persistent once the user closes the browsing window. Additionally, many participants used private browsing so that certain activities were not linked to their Google account, which by default they are logged out from in private browsing mode. However, we found that some participants performed certain tasks to prevent Google and other websites from recording the activity, and not just to prevent it from being linked to their computer or account. Users may not be aware that their search and YouTube activities are still being sent to Google even if they are not logged in, which some might still consider as a privacy invasion. Similarly, websites still record the activities of visitors to their website using various trackers that do not require an account login.

Many participants also expressed concerns about receiving targeted advertising. Though private browsing will prevent access to tracking cookies set in normal browsing mode, it does not prevent new ones from being set. Furthermore, advertisement agencies can still use other practices such as browser fingerprinting [11] and IP targeting [9] to serve targeted advertisements to a user or household. Safari and Firefox do enable some additional tracking protections in private browsing, but they still do not offer full protection against such techniques.

Our results indicate that people also use private browsing for security reasons, beyond generally maintaining their privacy. Some thought that private browsing would prevent attackers from hacking into their accounts or stealing their identity, for which private browsing does provide some protection. For example, private browsing does mitigate session hijacking attacks which use active logins [19]. However, it is likely that users are more concerned about vulnerabilities introduced by forgetting to log out of an account. In some cases, participants overestimated the protection against the security threats. For example, private browsing mode does not prevent users from downloading viruses or malware, nor does it provide additional protections than those offered from normal browsing in the transmission of their credit card and other personal information.

In about 10% of responses, participants were not sure exactly what private browsing protected them from, but expressed that they used private browsing because it provided some feeling of privacy or security. These misconceptions can be especially dangerous if users naively choose to use private browsing to conduct online activities which put them at risk, thinking they are being protected.

6. ACKNOWLEDGMENTS

This work was funded by the National Security Agency (NSA) Science of Security Lablet at Carnegie Mellon University (contract #H9823014C0140). Additional support was provided by the National Science Foundation (NSF) and the

Hewlett Foundation, through the Center for Long-Term Cybersecurity (CLTC) at the University of California, Berkeley. We also would like to thank our reviewers for their feedback.

7. REFERENCES

[1] A. Acquisti and J. Grossklags. Privacy and rationality in individual decision making. *IEEE Security & Privacy*, 3(1):26–33, 2005.

[2] A. Acquisti, S. Komanduri, P. G. Leon, S. Wilson, L. F. Cranor, N. Sadeh, Y. Wang, I. Adjerid, R. Balebako, L. Brandimarte, L. F. Cranor, N. Sadeh, F. Schaub, M. Sleeper, and Y. Wang. 44 nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys*, 50(44), 2017.

[3] L. Agarwal, N. Shrivastava, S. Jaiswal, and S. Panjwani. Do not embarrass: Re-examining user concerns for online tracking and advertising. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 8:1–8:13, 2013.

[4] G. Aggarwal, E. Bursztein, C. Jackson, and D. Boneh. An analysis of private browsing modes in modern browsers. In *Proceedings of the USENIX Security Symposium*, 2010.

[5] J. Angulo. "WTH..!?!" experiences, reactions, and expectations related to online privacy panic situations. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 19–38, 2015.

[6] Apple Support. Browse in private. `https://support.apple.com/guide/safari/browse-privately-ibrw1069`, November 2017.

[7] C. Canfield, A. Davis, B. Fischhoff, A. Forget, S. Pearman, and J. Thomas. Replication: Challenges in using data logs to validate phishing detection ability metrics. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[8] J. Cohen. Statistical power analysis for the behavioral sciences. *NJ: Lawrence Earlbaum Associates*, 2, 1988.

[9] DBS Interactive. IP targeting 101: Smart display advertising. `https://www.dbswebsite.com/blog/2016/03/16/ip-targeting-101-smart-display-advertising/`, November 2017.

[10] DuckDuckGo. A study on private browsing: Consumer usage, knowledge, and thoughts. Technical report, 2017. `https://duckduckgo.com/download/Private_Browsing.pdf`.

[11] P. Eckersley. How unique is your web browser? In *Proceedings of the International Symposium on Privacy Enhancing Technologies (PETS)*, pages 1–18, 2010.

[12] M. Fagan and M. M. H. Khan. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 59–75, 2016.

[13] A. Forget, S. Komanduri, A. Acquisti, N. Christin, L. F. Cranor, and R. Telang. Security Behavior Observatory: Infrastructure for long-term monitoring of client machines. Technical Report 14-009, Carnegie Mellon University CyLab, 2014.

[14] A. Forget, S. Pearman, J. Thomas, A. Acquisti, N. Christin, L. F. Cranor, S. Egelman, M. Harbach, and R. Telang. Do or do not, there is no try: User engagement may not improve security outcomes. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 97–111, 2016.

[15] S. Fox. Adult content online. *Pew Research Center*, 2005.

[16] X. Gao, Y. Yang, H. Fu, J. Lindqvist, and Y. Wang. Private browsing: An inquiry on usability and privacy protection. In *Proceedings of the Workshop on Privacy in the Electronic Society (WPES)*, pages 97–106, 2014.

[17] Google Chrome Help. Browse in private. `https://support.google.com/chrome/answer/95464?co=GENIE.Platform{%}3DAndroid{&}hl=en`, November 2017.

[18] P. Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.

[19] M. Johns. SessionSafe: Implementing XSS immune session handling. In *Proceedings of the European Symposium on Research in Computer Security (ESORICS)*, pages 444–460, 2006.

[20] R. Kang, S. Brown, L. Dabbish, and S. Kiesler. Privacy attitudes of Mechanical Turk workers and the US public. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 37–49, 2014.

[21] R. Kang, L. Dabbish, N. Fruchter, and S. 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)*, pages 39–52, 2015.

[22] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 453–456, 2008.

[23] J. Lazar, J. Feng, and H. Hochheiser. *Research Methods in Human-Computer Interaction*. Morgan Kaufmann, 2017.

[24] 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.

[25] Microsoft Support. Browse inprivate in microsoft edge. `https://support.microsoft.com/en-us/help/4026200/windows-browse-inprivate-in-microsoft-edge`, November 2017.

[26] Microsoft Support. Change security and privacy settings for internet explorer 11 - windows help. `https://support.microsoft.com/en-us/help/17479/windows-internet-explorer-11-change-security-privacy-settings`, November 2017.

[27] R. Montasari and P. Peltola. Computer forensic analysis of private browsing modes. In *Proceedings of the Communications in Computer and Information Science (CCIS)*, volume 534, pages 96–109, 2015.

[28] Mozilla Blog of Metrics. Understanding private browsing. `https://blog.mozilla.org/metrics/2010/08/23/understanding-private-browsing/`, October 2017.

[29] Mozilla Support. Private browsing - use Firefox without saving history | Firefox help. `https://support.mozilla.org/en-US/kb/private-`

browsing-use-firefox-without-history, November 2017.

[30] Mozilla Support. Tracking protection | Firefox help. https://support.mozilla.org/en-US/kb/tracking-protection, November 2017.

[31] Mozilla Support. Firefox focus. https://support.mozilla.org/en-US/kb/focus#w_performance, February 2018.

[32] P. E. Naeini, S. Bhagavatula, H. Habib, M. Degeling, L. Bauer, L. Cranor, and N. Sadeh. Privacy expectations and preferences in an IoT world. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[33] D. J. Ohana and N. Shashidhar. Do private and portable web browsers leave incriminating evidence?: A forensic analysis of residual artifacts from private and portable web browsing sessions. *EURASIP Journal on Information Security*, 2013(1):6, 2013.

[34] Opera Help. Private browsing. http://help.opera.com/Mac/12.00/en/private.html, November 2017.

[35] S. Panjwani and N. Shrivastava. Understanding the privacy-personalization dilemma for web search: A user perspective. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 3427–3430, 2013.

[36] S. Pearman, J. Thomas, P. E. Naeini, H. Habib, L. Bauer, N. Christin, L. F. Cranor, S. Egelman, and A. Forget. Let's go in for a closer look: Observing passwords in their natural habitat. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, pages 295–310, 2017.

[37] K. Purcell, J. Brenner, and L. Rainie. Search engine use 2012. *Pew Research Center*, 2012.

[38] E. Rader. Awareness of behavioral tracking and information privacy concern in Facebook and Google. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 51–67, 2014.

[39] L. Rainie, S. Kiesler, R. Kang, M. Madden, M. Duggan, S. Brown, and L. Dabbish. Anonymity, privacy, and security online. *Pew Research Center*, 2013.

[40] K. Satvat, M. Forshaw, F. Hao, and E. Toreini. On the privacy of private browsing - a forensic approach. *Journal of Information Security and Applications*, 19(1):88–100, 2014.

[41] F. Shirazi and M. Volkamer. What deters jane from preventing identification and tracking on the web? In *Proceedings of the Workshop on Privacy in the Electronic Society (WPES)*, pages 107–116, 2014.

[42] C. Soghoian. Why private browsing modes do not deliver real privacy. *Center for Applied Cybersecurity Research*, pages 79–94, 2011.

[43] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao. You are how you click: Clickstream analysis for sybil detection. In *Proceedings of the USENIX Security Symposium*, volume 9, 2013.

[44] Y. Wu, P. Gupta, M. Wei, Y. Acar, S. Fahl, and B. Ur. Your secrets are safe: How browsers' explanations impact misconceptions about private browsing mode. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 217–226, 2018.

APPENDIX

WARNING: Appendix B contains explicit content relating to search terms used to identify sensitive search engine queries.

A. DOMAIN CATEGORIES

The domain categories returned by AWIS were categorized into the following categories:

- adult
- audio
- education
- email
- financial
- health
- news
- political
- portal
- search
- shopping
- social_network
- software
- video
- other

B. SEARCH ENGINE QUERIES

The following keyword lists were used to identify sensitive searches conducted by SBO participants.

Adult: 2 girls, 4 girls, adult, ageplay, anal, aphrodisiac, asshole, august, bdsm, bikini, blow job, blowback, boob, brenner bolton, chaturbate, cheating, christian mingle, cock, dating site, derpibooru, dick, digital playground, ennio gaurdi, erotic, fetish, fleshlight, foursome, fuck, gay, gaydar, gianna michaels, hentai, horny, jackinworld, lesbiantube, literotica, madison scott, masturbat*, mfc, naked, naughty, nip slip, nipslip, nsfw, nude, nudography, orgasm, osiris blade, pigtails in paint, porn, pussy, reality kings, redtube, riley reid, sex, slut, squirt, strip club, strip poker, strip tease, sucking, threesome, tit, topless, tub girl, upskirt, vagina, virgin, xhamster, xkeezmovies, xtube, xvideo

Health: alcohol tolerance, aloe, anorexia, anxiety, asperger, bedsore, blister, body fat, burn, cabergoline, calories, careprostcanada, colonoscopy, concussion, condom, counseling, creatine, creatinine, cyproheptadine, dht, dim, doctor, dopamaine, dopamien, dry scalp, ephedra, ephedrine, feeling weak, fingering, fingured, glycemic, hair grow, heart beat, heartburn, hepatitis, hernia, hydrocodone, hypogonadism, hypogonadism, hysterectomy, infection, ingrown, insurance, itchy, leprosy, lice, lower back, malaria, medicaid, medical, menstrual, metamucil, minoxidil, mylanta, nose, nurofen, organ, pain, pediatric, penis, physical, pregnancy, proctolgist, prolactin, provider lookup, rohypnol, scar, serotonin, sickness, sneez*, ssri, stomach, sudafed, swollen, tattoo care, testosterone, thalidomide, therapy, tibulus, upset stomach, urine, valtrex, vicodin, yellow fever, zyrtec

Financial: american express, bank, bitcoin, bond, capital one, credit card, fcu, financial, income, interest rate, loan, pnc, salar*, stipend, stock, tax, wells fargo

Copyright: 1337x, dvdrip, ebook, piracy, pirate, piratebay, torrent

Political: bannon, bush email, Donald Trump, election, flag burning, free speech, heavens gate, jared kushner, jeff sessions, kim jung un, march for life, potus, president, protest, scaramucci, science march, spicer, trans murders, trump, vote

Other Sensitive: a joint, abuse, attack, cannabis, darknet, dies, eaze, fire, genocide, parramore, pcp, personal injury, pot, pulse, rape, weapons, weed

C. YOUTUBE ACTIVITY

The text of the element "unavailable-message" from the HTML of YouTube videos returned the following codes which indicated infringing, sensitive, or adult content related videos:

- Content Warning
- Copyright Violation
- Nudity/Sexual Content Violation
- Scam/Deceptive Practices Violation
- Terms of Service Violation
- Violent/Graphic Content Violation
- Community Guidelines Violation

D. SURVEY QUESTIONS

Description **For the duration of this survey we ask that you answer questions based on your behaviors and expectations associated with browsing the internet on your main home computer (desktop or laptop), unless stated otherwise.**

1. Which browsers do you regularly use? Check all that apply.

   ☐ Chrome ☐ Opera
   ☐ Edge ☐ Safari
   ☐ Firefox
   ☐ Internet Explorer ☐ Other

Every browser listed above has a built-in feature that allows users to engage in private browsing. However, they each refer to it slightly differently.

- Chrome refers to this feature as **Incognito mode**
- Edge and Internet Explorer call it **InPrivate Browsing**
- Firefox and Safari use **private browsing**
- Opera calls it **private tab**

Throughout this survey we will refer to this feature simply as "private browsing."

2. Have you ever used private browsing mode on your web browser?
   - ○ Yes
   - ○ No
3. Do you share the computer you regularly use for private browsing with other people (e.g. siblings, parents, partners, etc.)?
   - ○ Yes, but it is mainly mine
   - ○ Yes, and it is mainly someone else's computer
   - ○ Yes, and it is a shared/family computer
   - ○ No, I am the only user
4. When you use private browsing, which of the following browsers do you use? (If you use more than one browser for private browsing, select the one you use most often.)

   ○ Chrome ○ Internet Explorer
   ○ Edge ○ Opera
   ○ Firefox ○ Safari

Broad Understanding

5. What would you expect to be protected from when using private browsing in the [Q3 response] browser?
6. To the best of your knowledge, what do you think actually happens when you use private browsing in the [Q3 response] browser?

Specific Understanding. *Participants were shown the following Likert-style options for the set of statements below:*

Definitely correct	*Probably correct*	*I don't know*	*Probably correct*	*Definitely correct*

Please select if the following statements are correct.

7. Private browsing in the [Q3 response] browser causes the information I send to websites to be encrypted.
8. Private browsing in the [Q3 response] browser clears all my browsing history from my computer after I close the browser window.
9. Private browsing in the [Q3 response] browser clears most cookies for that browsing session from my computer after I close the browser window.
10. Private browsing in the [Q3 response] browser blocks some tracking by advertisement and social media companies.
11. Private browsing in the [Q3 response] browser clears my browsing history for that session from my computer after I close the browser window.
12. Private browsing in the [Q3 response] browser does not allow my Internet Service Provider (e.g. Comcast, Verizon) to see which websites I visited during that session.
13. Private browsing in the [Q3 response] browser prevents companies from targeting ads to me based on any of my previous browsing history.
14. Private browsing in the [Q3 response] browser prevents companies from targeting ads to me based on my browsing history from previous private browsing sessions.
15. Private browsing in the [Q3 response] browser blocks all ads on the websites I visit.

16. Private browsing in the [Q3 response] browser clears all the information that I fill into forms in that session from my computer.
17. Private browsing in the [Q3 response] browser does not save my login information after I end that session.
18. Private browsing in the [Q3 response] browser allows me to browse the web anonymously.
19. Private browsing in the [Q3 response] browser prevents my browser from sending any cookies to websites.
20. Private browsing in the [Q3 response] browser does not allow websites to get my computer's IP address or any information about my web browser or computer.

Private Browsing Usage. *Participants were shown the following options for each of the use cases in Q21 below:*

- ○ *Never*
- ○ *Once or a few times*
- ○ *A few times each week*
- ○ *Almost every day*
- ○ *Multiple times per day*
- ○ *Prefer not to answer*

21. How often did you perform each of the following activities in private browsing in [Q3 response] during the **past month**?
    (a) Shopping online
    (b) Performing any type of searches
    (c) Accessing social media
    (d) Logging into accounts on someone else's computer
    (e) Using a computer that isn't mine (e.g. public, friend's, or work computer)
    (f) Logging into accounts on my computer
    (g) Performing sensitive searches
    (h) Viewing adult content
    (i) Streaming content (music/video)
    (j) Accessing news websites that have a viewing limit
    (k) Accessing websites that have ad blocking detection (i.e., won't let me me access the content if my adblocker is on)
    (l) Pirating content (software, videos, music, etc)
    (m) Using it for all of of my browsing
22. Are there any other activities for which you use private browsing in [Q3 response]?
    - ○ No
    - ○ Yes, I use it for... _____
23. What do you consider to be a sensitive search?

Specific Scenarios. *The question below was repeated for each activity the respondent indicated using private browsing in Q21.*

24. What are the reasons you use private browsing in [Q3 response] when [Q21 response]?

Cookie Policy

25. What is your current cookie policy for [Q3 response]? Select all that apply.
    - ☐ Whatever is the default option
    - ☐ Block all cookies
    - ☐ Allow cookies from the current website only
    - ☐ Allow cookies from websites I visit
    - ☐ Allow all cookies (third-party included)
    - ☐ Allow session cookies
    - ☐ Keep cookies only until I close my browser window
    - ☐ I don't know

Privacy Plugins and Other Steps

26. Please select which of the following types of browser plugins and extensions you use. Select all that apply.
    - ☐ Protect you from malware or phishing websites
    - ☐ Browse anonymously
    - ☐ Block ads
    - ☐ Encrypt your communications
    - ☐ Protect children
    - ☐ Prevent websites from tracking your browsing activity
    - ☐ Manage passwords
    - ☐ Other privacy or security functions _____
    - ☐ None of the above
27. Do you take any other steps to protect your privacy while browsing (other than private browsing, if you use it)?
    - ○ Yes
    - ○ No
    - ○ I don't know
28. Which steps do you normally take?

IUIPC. *Participants were shown the following Likert-style options for the set of statements below:*

- ○ *Strongly agree*
- ○ *Agree*
- ○ *Somewhat agree*
- ○ *Neither agree nor disagree*
- ○ *Somewhat disagree*
- ○ *Disagree*
- ○ *Strongly disagree*

Please select how much you agree with the following statements.

29. Consumer online privacy is really a matter of consumers' right to exercise control and autonomy over decisions about how their information is collected, used, and shared.
30. Consumer control of personal information lies at the heart of consumer privacy.
31. I believe that online privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction.

32. Companies seeking information online should disclose the way the data are collected, processed, and used.
33. A good consumer online privacy policy should have a clear and conspicuous disclosure.
34. It is very important to me that I am aware and knowledgeable about how my personal information will be used.
35. It usually bothers me when online companies ask me for personal information.
36. When online companies ask me for personal information, I sometimes think twice before providing it.
37. It bothers me to give personal information to so many online companies.
38. I'm concerned that online companies are collecting too much personal information about me.

Demographics

39. How often did you use private browsing in [Q3 response] in the **past week** on your **main home computer**?
    - ○ Every time
    - ○ Most of the time
    - ○ About half the time
    - ○ Sometimes
    - ○ Rarely
    - ○ Never
40. How often did you use private browsing in [Q3 response] in the **past week** on your **main mobile device**?
    - ○ Every time
    - ○ Most of the time
    - ○ About half the time
    - ○ Sometimes
    - ○ Rarely
    - ○ Never
41. How similar were the activities you did in private browsing on your mobile device to the activities you did in private browsing on your main home computer?
    - ○ Completely the same
    - ○ Sometimes the same and sometimes different
    - ○ Completely different
42. What was different about the activities you did in private browsing on your mobile device?
43. How old are you?
    - ○ 18-24
    - ○ 25-34
    - ○ 35-44
    - ○ 45-54
    - ○ 55-64
    - ○ 65-74
    - ○ 75-84
    - ○ 85 or older
    - ○ I prefer not to answer
44. How do you self identify?
    - ○ Male
    - ○ Female
    - ○ _____ Other
    - ○ I prefer not to answer
45. What is the highest level of education you have achieved?
    - ○ Less than high school
    - ○ High school graduate
    - ○ Some college
    - ○ Trade/Technical school
    - ○ Associate degree
    - ○ Bachelor's degree
    - ○ Advanced degree (Master's, Ph.D., M.D.)
    - ○ I prefer not to answer
46. Which of the following best describes your primary occupation?
    - ○ Administrative Support (e.g., secretary, assistant)
    - ○ Art, Writing, or Journalism (e.g., author, reporter, sculptor)
    - ○ Business, Management, or Financial (e.g., manager, accountant, banker)
    - ○ Education or Science (e.g., teacher, professor, scientist)
    - ○ Legal (e.g., lawyer, paralegal)
    - ○ Medical (e.g., doctor, nurse, dentist)
    - ○ Computer Engineering or IT Professional (e.g., programmer, IT consultant)
    - ○ Engineer in other field (e.g., civil or bio engineer)
    - ○ Other _____
    - ○ Service (e.g., retail clerk, server)
    - ○ Skilled Labor (e.g., electrician, plumber, carpenter)
    - ○ Unemployed
    - ○ Retired
    - ○ College student
    - ○ Graduate student
    - ○ Mechanical Turk worker
    - ○ I prefer not to answer
47. Have you ever held a job or received a degree in computer science or any related technology field?
    - ○ Yes
    - ○ No
48. Are you either a computer security professional or a student studying computer security?
    - ○ Yes
    - ○ No
49. Which platform do you use most frequently for web browsing?
    - ○ Laptop/Desktop
    - ○ Phone/Tablet
    - ○ I use both equally
50. Which operating system do you use on your main home computer?
    - ○ Windows
    - ○ MacOS
    - ○ Linux distribution
    - ○ Other _____
51. If you have any other comments or feedback, please use the space below.

Online Privacy and Aging of Digital Artifacts

Reham Ebada Mohamed
School of Computer Science
Carleton University
riham.mohamed@carleton.ca

Sonia Chiasson
School of Computer Science
Carleton University
chiasson@scs.carleton.ca

ABSTRACT

This paper explores how the user interface can help users invoke the *right to be forgotten* in social media by decaying content. The decaying of digital artifacts gradually degrades content, thereby becoming less accessible to audiences. Through a lab study with 30 participants, we probe the concept of aging/decaying of digital artifacts. We compared three visualization techniques (pixelating, fading, and shrinking) used to decay social media content on three platforms (Facebook, Instagram, and Twitter). We report results from qualitative and quantitative analysis. Visualizations that most closely reflect how memories fade over time were most effective. We also report on participants' attitudes and concerns about how content decay relates to protection of their online privacy. We discuss the implications of our results and provide preliminary recommendations based on our findings.

1. INTRODUCTION

Online sharing contributes to individuals' well-being and social interactions [1, 9, 13, 39]. For example, directed communication on Online Social Networks (OSN) can promote social bonding and positive feelings [13]. It can also facilitate the process of finding and interacting with classmates [1] or maintaining relationships with family and acquaintances [28]. In addition, the use of social media provides individuals with needed social support in case they experience negative feelings such as grief [72] or loneliness [39]. Online communication and social media can also positively contribute to adolescent development through increasing self-esteem and providing an outlet for identity experimentation [9, 68].

However, many incidents and research have also demonstrated the potential negative consequences of online sharing. For example, OSN data may be considered during important selection processes (such as in hiring or school admission decisions) [38, 73], resulting in individuals' professional [3, 23, 45, 46], or academic future [65] being compromised by their digital footprints. Moreover, the Internet exploits the fact that a privacy paradox [1, 2, 18, 33, 49, 58, 66] exists among users by making salient the desire to divulge while downplaying the desire for privacy [33]. In addition, Coopamootoo and Groß suggest that it may be challenging for users to follow both a *privacy attitude* and a *sharing attitude* simultaneously because the two attitudes stem from two opposing forces or emotions: fear and happiness, respectively [18].

OSNs and other online repositories have contributed to making ephemeral information permanent. In the European Union (EU), the *Right to be forgotten* entitles individuals, after a certain time has passed and under other specific conditions, to ask search engine companies to de-index and delete potentially damaging personal digital material. As a result, forgetting digital memories [45] has become an important principle to diminish the potential negative repercussions resulting from the persistent reproduction of our digital footprints. While there exists a general emphasis on reminiscing [21, 56], forgetting digital memories introduces a converse emphasis on dissociating from obsolete and irrelevant digital artifacts. In this regard, there has been an emphasis on representing the passage of time in the field of Human-Computer Interaction (HCI) [42] to preserve the temporal contextual integrity of previously published information [5, 10, 48, 50].

One approach to dissociate from obsolete content is to visualize the passage of time within the user interface (UI) by having older content gradually age or decay. Aging of content has two conceivable purposes. It provides temporal context to viewers and it provides some privacy advantages as posts become less accessible to viewers. Different temporal cues for indicating the age of Facebook online content were proposed by Novotny [50] who implemented and partially evaluated one such prototype.

The literature suggests opportunities to design forgetting mechanisms that support users' online identity management needs. However, it is unclear how the UI can provide temporal context that is non-obtrusive and natural to users, while also protecting their privacy. Our study examines the concept of aging of social media digital artifacts from the user's perspective. It aims to identify representations that match users' metaphor of aging and explores the representation of temporal cues [50] on OSN profiles for supporting user privacy. In particular, we were interested in these two research questions: (RQ1) Which of the three studied visualizations best represents digital aging on social media from a user perspective? and (RQ2) What are users' attitudes and concerns relating to digital aging on social media?

Through a lab study with 30 participants, we compare three different visualization techniques that decay OSN content visible to other users on three different social media platforms. Using both qualitative and quantitative analyses, we identify which visualization best represents aging of digital artifacts. We report participants' attitudes and concerns, and discuss their preferences regarding content decay. We further offer some preliminary recommendations for using decay to enhance online privacy.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

2. BACKGROUND AND RELATED WORK

Since the phenomenon of sharing data online is broad and includes various dimensions, some aspects are beyond the scope of our review. Among these dimensions are issues of practical implementation and enforcement of privacy laws. Other issues relate to sensor data privacy [57] and data collection and behavioural tracking by institutions or apps [15, 40, 59, 77]. While these are important concerns, they are tangential to our current research questions. We concentrate our literature review on privacy issues relating to interpersonal sharing on OSNs.

2.1 The Privacy Paradox

A dichotomy exists between online users' reported attitudes and their actual behaviour towards privacy, coined as a privacy paradox [1, 2, 4, 5, 18, 33, 49, 58, 66]. Online users report willingness to protect their own privacy [2], but studies show that few actions are performed for that purpose [2, 4, 5, 66]. Moreover, even privacy concerned individuals unknowingly disclose information that might be sensitive when they are in specific web contexts, such as online shopping [66], or when expecting a payoff or a reward [2].

2.1.1 Regretting

Some information users disclose online might be regrettable in the future [33, 74]. Digital footprints may bring unintended negative consequences in important selection processes [3, 23, 38, 45, 46, 65, 73]. Furthermore, although users are keen to reveal details about themselves through social media posts [13,33,58], their willingness to re-share the same content significantly decreases with time [4].

2.1.2 Preventing regret

Researchers have investigated ways to embed privacy management tools in OSNs to handle past or regrettable disclosed information. For example, Wang et al. [75, 76] introduced nudging to Facebook users to prevent potential regret when sharing status updates. Three nudges were introduced [76]: reminding users about the audience of the post, delaying publishing the post, and giving feedback regarding content containing strong sentiments. Although perceived as beneficial, users started to ignore the nudges within days. Moreover, while users liked the first nudge, they found the second and third nudges intrusive. Another way to prevent potential future regret about disclosed information is to set an expiry date for the published information [4], as discussed in Section 2.2.2.

2.2 Forgetting in the digital age

The idea that individuals should be able to move beyond their past artifacts and actions has been most prominently discussed by the EU. Principles of the *Right to be forgotten (RTBF)* were upheld in May 2014 by the European Court of Justice in González's versus Google [67]. The court ruled that search engines must remove links to pages that "appear to be inadequate, irrelevant or no longer relevant or excessive in the light of the time that had elapsed" when requested by individuals. In May 2018, the EU's new General Data Protection Regulation [55] comes into effect. Concepts of data minimization from the RTBF, however, have been included in earlier data protection laws and in the EU Data Protection Directive of 1995 [54]. During the same year, a joint study [70] by the Dutch Data Protection Authority and the Information and Privacy Commissioner of Ontario in Canada also explored a new approach to privacy and identity protection, that served as basis for seven *Privacy by Design (PbD)* principles [14], namely:

1. Proactive not reactive; preventative not remedial
2. Privacy as the default setting
3. Privacy embedded into design
4. Full functionality — positive-sum, not zero-sum
5. End-to-end security — full lifecycle protection
6. Visibility and transparency — keep it open
7. Respect for user privacy — keep it user-centric

The PbD principles serve as a framework for proactively embedding privacy during the system engineering process and more broadly within organizational practices. The framework's main goal is to make central the concern for individual privacy by promoting user trust and accountability when handling personal data.

Two decades have passed since these initial efforts, but many issues remain unresolved or only partially addressed. More recently, the right to be forgotten in the context of digital artifacts was described as a fundamental need in Mayer-Schönberger's [45] book, *Delete: The Virtue of Forgetting in the Digital Age*. The book illustrated several examples of individuals who have had their professional lives compromised because of their digital footprints, and emphasized the importance of "forgetting" in the digital age.

2.2.1 Deletion

Ochrat and Toch [4] and Ayalon and Toch [5] found that users' willingness to re-share information decreases with time, as it becomes less relevant. In the meantime, the probability that they delete such irrelevant information was low [4, 5] and there was no obvious tendency of users to permanently change their old content. Thus, users' reported approaches towards sharing do not align with their actual behaviour, which could be explained by the privacy paradox [5]. However, other reasons also include the desire to keep past posts for reminiscing [5, 7]. Similarly, participants in Zhao et al.'s [80] study appreciated reflection over their past and revisited their older content, expressing regret over their deletion decisions. Thus mechanisms that permanently delete content do not appear appropriate for most users as a solution for long-term retrospective privacy [5] or when curating their online-self [80]. These mechanisms include solutions such as the Web 2.0 Suicide Machine [17], or deleting content after a certain amount of time [7].

2.2.2 Expiry, Archival, and Decay

Based on the identified gap between users' sharing preferences and their willingness to delete, Ochrat and Toch [4] proposed having an information expiry feature on Facebook. They [4, 5] also suggested other mechanisms for ongoing privacy management instead of deletion: archiving, compaction, and blocking.

When considering an information expiry feature, it might be challenging to set expiration defaults to accommodate preferences for sharing information for short periods and long periods [4]. In addition, Bauer et al. [8] cast doubt on the usefulness of content expiration and suggested that extensive archival features would not be appropriate for users. Through two studies about privacy settings using the temporal dimension, Bauer et al. [8] found a gap between users' prediction about how their own privacy preferences would change over time and the actual change in their preferences. They instead suggested designing interfaces that promote reflection on older content [8]. Gulotta et al. [25] suggested that a more subtle mechanism to handle irrelevant content, such as selective archiving rather than extensive archiving, would be more helpful to users because archiving moves irrelevant content to a secondary storage that remains accessible only by the content publisher [5].

A more concrete and elaborate theoretical proposal of forgetting mechanisms and interfaces was discussed by Barua et al. [7]. They set forth theoretical foundations for the design of user-controlled

forgetting mechanisms in HCI that parallel forms of human forgetting. They discuss the benefits and consequences of implementing five forgetting mechanisms: decay, deletion, compaction, blocking, and archival. For example, they demonstrate that a decay mechanism that provides gradual removal of obsolete content would simulate the decay process in human memory [11, 62].

2.2.3 Information Obfuscation

Another approach is to fully or partially obfuscate sensitive photo elements [32, 41] or user attributes [16, 61]. Obfuscating attributes, however, may not be effective against inference attacks [16]. Li et al. [41] further showed that some commonly used face and body obfuscation are neither [41] effective for privacy nor preferred by users. One limitation to face and body obfuscation in OSNs is that it does not provide integrated protection of all contextual cues [41] or personally identifiable information [71]. For example, other parts of the photo or post (e.g., background, comments, time, and location check-in) can be recognized by other users [41].

2.3 Remembering and Reflection

While arguments for enabling forgetting aim to allow people to move beyond their past, there are benefits to remembering and allowing individuals to reflect on their histories. People tend to keep physical artifacts with certain tangible or intangible value [35], and online users also tend to keep and archive their digital artifacts [35, 43]. It is thought that the capabilities of digital technology should be used to eliminate limitations of human memory and to provide a valuable lifelong remembering experience [21, 56]. Therefore, some HCI practices seek to support everyday reminiscing [21, 56], use the web as a personal archive and for information management [43], consider digital inheritance [53], and enable reflection on social relationships [64] or personal past [69]. As discussed in Section 2.2.1, reflection over old content is important to users, especially when maintaining their online identity. Ayalon and Toch [4] suggested that the format of the Facebook timeline offers a reasonable starting point for enabling users to review and reflect on old content, and to manage their privacy. The following section reviews existing work on authoring of history and self-reflection. However, we focus on contextual privacy and its potential in providing better control and space for users when curating and reflecting on their online self.

2.4 Contextual privacy

Barth et al. [6] proposed a formal model of privacy and contextual integrity that links protection of personal data to norms in specific contexts. *Contexts* refer to how individuals act in certain roles within distinctive social domains [6]. The model serves as a conceptual framework endorsing the concept that privacy is not about secrecy, and individuals willingly share personal information if they are assured that specific social norms have not been violated. Online users, as individuals in the society, transact in different capacities by managing their online identity. They present themselves in a way that matches current social circumstances [5, 25, 26, 29].

2.4.1 Online Self

The literature has shown that maintaining online identity is not an ephemeral act, rather, it is an enduring one [29, 79]. Harper et al. [27] and Hogan [29] explored the concept of identity articulation through time on Facebook. They reflected on how outdated content can resurface, highlighting that social media focuses on "now" even though the associated events may have occurred in the past [27,29]. OSN content associated with online identity can become an exhibit that is encountered by different audiences, in different times, and in different contexts [29]. In this regard, some researchers proposed ways in which users can edit their past histories or choose how these histories should be handled in the future. For example, Thiry et al. [69] used the timeline metaphor to introduce a framework that allows authoring of personal histories to build meaning between the present and the past. In addition, Gulotta et al. [25] proposed three systems that prompt users to choose how their digital artifacts should be handled in the future. Based on their findings, Gulotta et al. [25] encouraged the development of tools that provide users with selective archiving and safekeeping of digital data denoting experiences outside of daily activities.

2.4.2 Contextual Privacy Settings

The literature also recognizes a need for contextual privacy settings [5, 6, 44]. Users curate online self-representational data to meet current circumstances [5, 25, 26, 29]. Madejski et al. [44] showed that Facebook privacy settings did not match users' sharing intentions, and identified a need for contextual privacy settings. Zhao et al. [80] noted that Facebook did not support an obvious personal space for private reflection when users curate their data and exert control over how they will be exhibited. In addition, Novotny and Spiekermann [51] showed that users desire control over their disclosed personal information in OSNs and need to dissociate from obsolete information that represents their past identity.

2.4.3 Visualizations for temporal integrity

One approach to dissociate users from obsolete information is to visualize time within the UI and have older content gradually decay [50]. As suggested by Novotny [50], this approach can preserve information's temporal contextual integrity, which is one of the key building blocks of user privacy [10, 48]. Based on a focus group, Novotny [50] proposed a catalogue of temporal interface cues to indicate the age of Facebook posts. He [50] classified these cues into *temporal indices* that incorporates time as a property of the posted information and *temporal symbols* that can be used as additional visual cues. A table summarizing Novotny's catalogue is available in Appendix A. The temporal indices manipulate the display properties of the information (e.g., through size, motion, decay), while temporal symbols include objects that indicate the time of the post (e.g., adding pictograms) and methods to manipulate the layout (e.g., horizontal or vertical) or typography [50].

Although an interesting proposal, few of Novotny's [50] temporal interface cues have been evaluated. A Facebook prototype visualized the passage of time by gradually decreasing the size of posts, and posts were arranged horizontally on the user's timeline. Although properties of the photo in the post and the caption were manipulated, other contextual cues that might be revealing, such as date of the post [41] were not manipulated. It was also suggested [50] that shrunk posts should still be clickable to ensure readability but it was not clear whether the prototype implemented this feature. However, we think that making the original information available defeats the purpose of degrading them. The prototype was partially evaluated in a study with 14 participants by having them complete one task. The horizontal arrangement of posts did not appeal to participants because it did not match other familiar interfaces which display posts vertically in chronological order.

Another experimental lab study [52] adopted two other temporal cues (temporal order and graphical timelines) in a hiring process simulation where reputation profiles of job-seekers were shown to participants acting as employers [52]. Results showed that the graphical timeline helped users more easily disregard obsolete information compared to the temporal order cue. However, the other temporal cues suggested by Novotny [50] remain untested.

2.5 Existing Gap

Although the literature has explored different forgetting mechanisms and provided insights on how to better match users' goals, such as providing contextual privacy settings and allowing reflection over older content, it is unclear how these mechanisms apply within OSNs. For example, how can an OSN timeline support forgetting and reflection simultaneously? How can an interface provide an immediate contextual visual cue that can promote privacy whilst presenting a natural non-obtrusive metaphor to users? There also remains other open design and research questions about visualizing the passage of time in HCI [42]. How should designs handle the disconnect between representations of time and time as remembered? Which metaphors represent a clearer analogue to human experience? And, how should the passing of time be depicted? [42]. Furthermore, how do users prefer to depict the passage of time to others, to represent their current personalities, and to show progression in life? And would users actually want time to be depicted; what benefits or concerns exist with such mechanisms? And finally, what are the privacy implications relating to these issues?

3. OUR STUDY

In our present study, we further probe the concept of having older content gradually decay and become less accessible to audiences. We believe this approach simulates the idea of archiving as a subtle mechanism to handle digital artifacts. It also provides an immediate contextual cue to the viewer about the age of posted content. We extend Novotny's [50] study by comparing three different visualizations on three different OSN platforms. We choose three different OSN platforms instead of one to see if our findings are applicable across platforms. We also chose three distinct visualizations that degrade content differently and fall under two of Novotny's [50] suggested temporal indices: *display salience* and *degrading display quality*. Our study partially answers some of the remaining open research questions regarding visualizing time in OSNs. Our two research questions are:

RQ1: Which of the three studied visualizations best represents digital aging on social media from a user perspective?

RQ2: What are users' attitudes and concerns relating to digital aging on social media?

4. METHODOLOGY

Our study explores visualizations of social media posts to simulate the decay or fading of memories over time. The visualizations are intended to illustrate that posts are getting older or aging to the viewer. The visualization is applied to content viewed by "others" as opposed to content that is self-accessed. For example, it is applied to Jane's Facebook profile as viewed by her friends, not content solely viewable by Jane. Aging of posts has two possible interrelated purposes. It provides temporal context to viewers and it provides some privacy advantages as posts become less accessible by viewers. There are, however, several dimensions when considering aging of posts, such as information sensitivity, access control options, and determining the appropriate decay function given a specific sharing scenario. For this first study, we focus on identifying the best decay representation out of three studied visualizations from a user perspective, recognizing that further work focusing on the other dimensions will be needed in other studies. Our study also captures users' attitudes and concerns regarding the concept and its potential purposes, including privacy. During the study, we introduced the concept of "aging" as posts getting older over time, but we carefully avoided mentioning "privacy" as a reason why this might be desirable until the very end of the study to avoid unduly influencing participants' perspectives.

To answer RQ1, we gauged users' preferences as determined by responses to Likert-scale questions and interview questions about the preferred visualization for use on their own data. Likert-scale questions considered aspects such as meaning, intuitiveness, most natural metaphor, and visual appeal.

To address RQ2, we collected more in-depth answers from users through interview questions and open-ended questions in a wrap-up questionnaire. For example, some questions explored their interpretations and impressions of the visualizations, if they think the concept of aging digital artifacts is necessary, and if they would like their own artifacts to age. We also asked about how aging should take place and if they could think of cases in which aging is more useful than deletion or content expiration. Other questions were relevant to the process itself, e.g., what are the thresholds for the aging process, what should the settings look like, and how does this concept relate to their privacy.

The study methods and questionnaires were pilot tested prior to data collection. Descriptions of the study tasks, interview guide, and questionnaires are available in Appendices B through D.

4.1 Recruitment

The study was cleared by our Research Ethics Board. Recruitment was done through posters posted across campus and a social media page for advertising the university's HCI studies. Participants were compensated $15 for their time. Before beginning the session, participants read and signed a consent form that explained the purpose and the procedure of the study, and it reminded them that the session will be audio-recorded. Personally identifiable information collected from participants was limited to their voice; responses were pseudo-anonymized and non-attributable.

We had 30 participants; 12 were male and 18 were female, with a mean age of 26 (Std. Dev = 9 years). They reported having an average of three social media accounts each and spending an average of three hours (Std. Dev = 2 hours) online daily. The majority were university students; 16 participants were undergraduate students, 9 were graduate students, and 5 were university staff.

Participants were assigned a username that is not linked to their identity and these usernames were used during data compilation and to report results in the paper. Usernames were generated according to participants' assigned platform (e.g., Facebook: FB1-FB10, Twitter: TW1-TW10, Instagram: IG1-IG10).

4.2 Prototype

We created a fictitious social media profile on three different social networks: Facebook, Twitter, and Instagram. We choose several platforms to explore whether our results applied across a range of interfaces. Facebook, Twitter, and Instagram are among the top 5 most popular OSN sites [20, 34] and each has a distinct purpose.

In our prototypes, the profile layout and arrangement imitated the existing look and feel of July 2017 UI versions of each of the three platforms. The content on both Facebook and Twitter was identical; it included miscellaneous photo posts with captions and status updates. To conform with Instagram's layout, its fictitious content included only photos with captions. We intentionally included content that is personal in nature [30], such as family photos, photos of a car with the licence plate number visible, and photos of a house with a visible street address. Status updates included personal sentiments and opinions about potentially sensitive subjects [37] (e.g., political views, support for LGBT).

We implemented decay techniques on the three OSN platforms (Facebook, Twitter, Instragram), using three different approaches: (1) *content fading*, (2) *content pixelation*, (3) *content shrinking*, resulting in $3 \times 3 = 9$ prototypes. We chose techniques from Novotny's taxonomy [50] that seemed likely to convey *privacy* and *aging* based on our literature review; others could also be considered.

The dates of the fictitious posts were separated by a month and each prototype showed posts spanning one year. The decay was applied linearly across posts; for example in the fading prototype, transparency levels were reduced by equal increments between any two sequential posts. Figures 1, 2, and 3 show the nine prototypes on Facebook, Twitter, and Instagram respectively. Each prototype was displayed to the user as a scrollable webpage.

Unlike face, body, or object obfuscation, the decay techniques in our prototypes degraded the entire post. To ensure that limitations posed by such obfuscation techniques [41] were avoided, we manipulated all the contextual cues related to a post that might be recognizable [41] along with the image itself. These manipulated cues included the image's caption, its intended audience, publishing date and time, comments, date and time of the comments, and tagged friends. Manipulated content was also unclickable to prevent retrieving or accessing the original unmodified post. Moreover, although pixelation is ineffective for privacy as an obfuscation technique [41], we choose it as one of the decaying visualizations as the obscuring effect was linearly increased across multiple time-related posts, which is a different application of pixelation than its application in obfuscation of time-unrelated data.

4.3 Procedure

Thirty participants took part in our 3×3 mixed design lab study featuring one between-subject variable (social media type) and one within-subject variable (decay technique); ten participants were assigned to each of three social media types, and each participant saw all three visualizations. Assignment of social media types and presentation order of the visualizations was controlled using a full latin square to ensure that all combinations were cycled and to avoid possible ordering effects. For example Participant X saw {Facebook-Fading, Facebook-Pixelating, Facebook-Shrinking} and Participant Y saw {Twitter-Pixelating, Twitter-Shrinking, Twitter-Fading}.

We collected participants' feedback verbally and through online questionnaires in a 60-minutes session. A session unfolded as follows:

1. View and explore Prototype A
2. Complete visualization questionnaire A
3. View and explore Prototype B
4. Complete visualization questionnaire B
5. View and explore Prototype C
6. Complete visualization questionnaire C
7. Interview/conversation about the concepts and prototypes
8. Complete wrap-up questionnaire

In Steps 1, 3, and 5, participants viewed the social media content as if they were previewing another user's social media profile, not their own. We asked some probing questions while participants viewed each prototype, e.g., what was their interpretation of the visualization, what was most appealing/confusing, and whether they would change anything in the design. Other questions explored if the visualization was meaningful in terms of conveying the idea that posts were getting old.

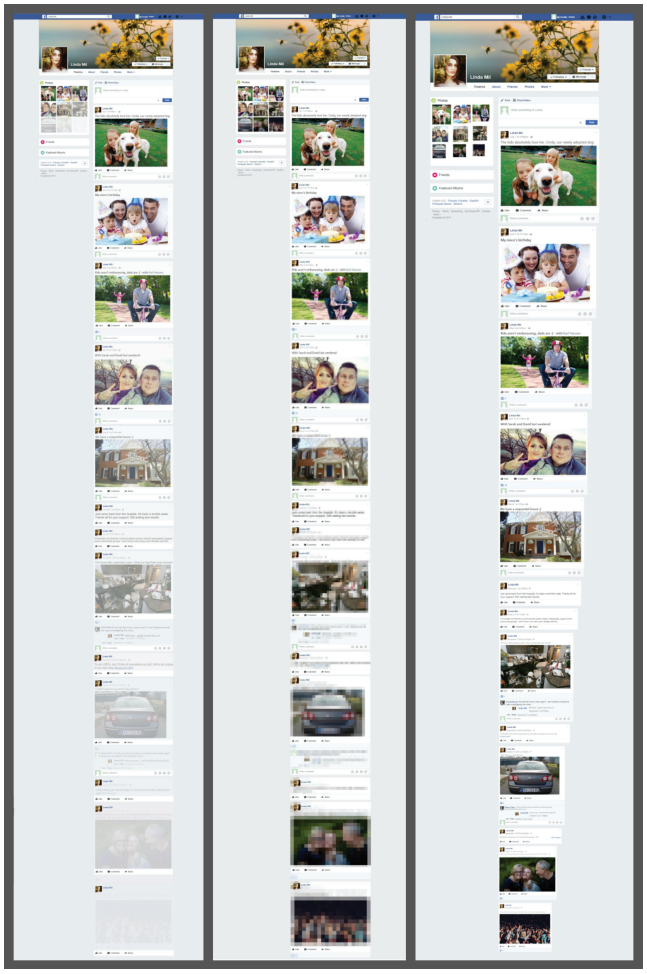


Figure 1: Facebook fading (L), pixelating (M), shrinking (R).

In Steps 2, 4, and 6, the visualization questionnaires consisted of 10 Likert-scale questions covering agreement with the visual representation: (Q1) easily showed posts were getting old, (Q2) was meaningful, (Q3) was confusing, (Q4) was complete, (Q5) changed their perspective, (Q6) was appropriate to the content, (Q7) was obtrusive, (Q8) of photo posts was intuitive, (Q9) of text posts was intuitive. And finally, (Q10) whether they would use the visual representation on their social media account.

In Step 7, the wrap-up interview questions sought to learn about users' attitudes and concerns as both *a user browsing another user's profile* and as *an owner of the profile* concerned about other users. For example, we asked for participants' reaction if they came across a profile that uses one of the content decay visualizations. Other questions examined participants' perception of aging of digital artifacts, how necessary it is, and by which means it should be implemented in OSNs (e.g., by deletion, expiration, or decay). More questions probed whether participants would use one of the visualizations to display their own digital artifacts when accessed by other user groups, whether the study changed how they would use social media in the future, and whether content decay would promote their online privacy.

In Step 8, the wrap-up questionnaire consisted of one Likert-scale question and three open-ended questions. In total, each participant

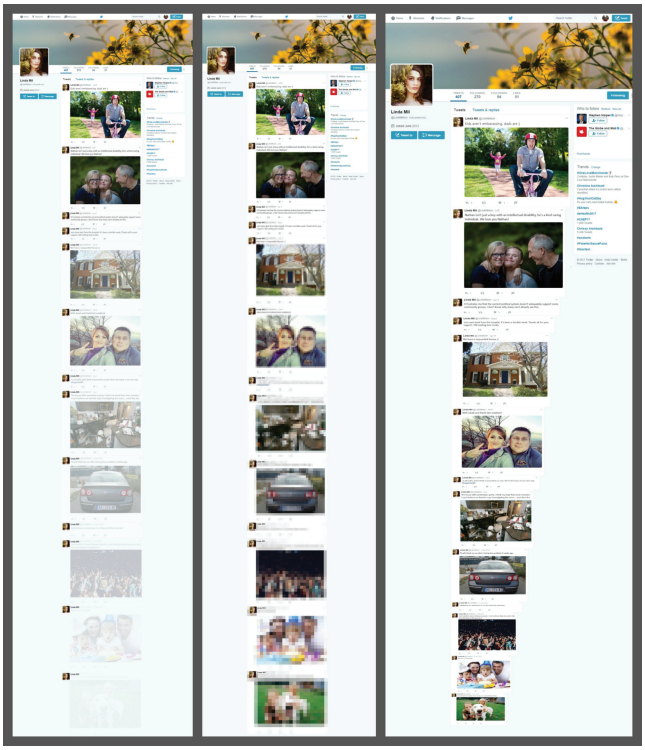


Figure 2: Twitter fading (L), pixelating (M), shrinking (R).

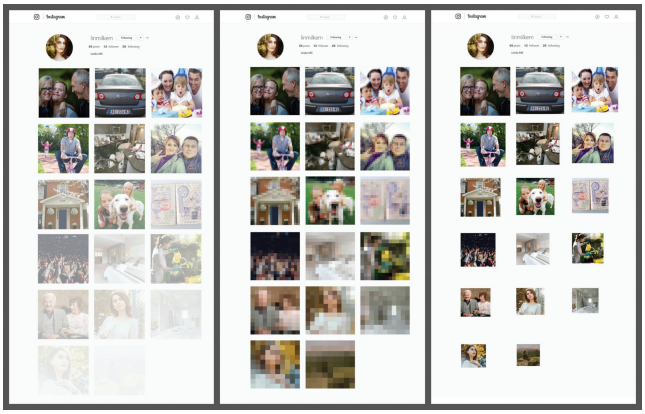


Figure 3: Instagram fading (L), pixelating (M), shrinking (R).

gave feedback on three different prototypes, filled out four online questionnaires, and shared opinions pertaining to the concept of content decay.

4.4 Analysis Process

To answer our two research questions, we performed both quantitative and qualitative analyses.

For the statistical analysis, we were primarily concerned with our within-subject variable, *visualization type*, with three levels (fading, shrinking, and pixelating). We used Friedman tests (significance level of $p < 0.05$) to test for main effects of visualization type. In cases of significant omnibus test results, we followed up with pairwise Wilcoxon signed-rank test with Bonferroni correction applied (significance level of $p < 0.017$).

Table 1: Mean values out of 5 per question for each approach. Highest values are highlighted in gray. *Q3 and Q7 were negatively worded; responses were reversed for analysis so that a higher score signifies a more positive response.

Question	*Fading*	*Pixelating*	*Shrinking*
Q1: Vis. easily shows aging	3.87	2.87	3.53
Q2: Meaningful vis.	3.57	2.17	3.67
Q3*: Confusing vis.	3.30	2.03	3.33
Q4: Complete vis.	2.80	1.87	2.90
Q5: Changed my perspective	2.50	2.53	2.70
Q6: Appropriate to the content	3.10	2.40	3.30
Q7*: Obtrusive vis.	2.80	1.93	3.00
Q8: Vis. of photos was intuitive	3.63	2.60	3.67
Q9: Vis. of text was intuitive	3.27	1.97	2.87
Q10: Would use this vis.	2.17	1.63	2.63

The qualitative data consisted of 23 hours of audio-recordings from the interviews and open-ended questions from the questionnaires. We transcribed the relevant parts of the interviews. We used content analysis methodologies [31] following an inductive process which included multiple iterations across the transcripts. We categorized data primarily according to 1) expressed preferences/dislikes/reasons for each, 2) attitudes towards digital aging, 3) interpretation of the purpose of the visualizations, and 4) interest in incorporating aging into their social media, and the requirements/settings for such functionality. The main researcher compiled the data and extracted the main themes looking for key patterns and particularly insightful feedback through several rounds. A second researcher was involved refining the patterns, interpreting the data, and handling any complicated cases, but did not independently code the data.

5. RQ1 ANALYSIS AND RESULTS

We summarize results of our statistical and qualitative analysis pertaining to our first research question: *Which of the three studied visualizations best represents digital aging on social media?*

5.1 Statistical Analysis of Questionnaire

Participants completed ten 5-point Likert scale questions per visualization technique. Mean responses are available in Table 1; higher means indicate more positive scores.

Using the within-subjects variable, visualization technique, we compared questionnaire responses to see if participants favoured any technique. We found a significant difference in nine out of the ten questions. Friedman's test results are presented in Table 2, with significant differences highlighted in gray. Table 3 shows the pairwise comparison between the three approaches and the associated p values (Bonferroni corrected).

Mean responses to the questionnaire ranged from negative to neutral, suggesting that participants were generally unenthusiastic about the visualization techniques. Reasons for this are discussed in Section 5.2; participants were mainly concerned that the visualizations might obstruct browsing within social media.

The statistical analysis showed that pixelation was least favourable to participants. Shrinking was the most favourable, but participants did not significantly favour it over fading. Nevertheless, shrinking and fading were significantly more preferable than pixelation.

For completeness, we also verified whether there was a main effect of *social media type* (Facebook, Instagram, Twitter). This was a

Table 2: Friedman test statistic and significance values. Degrees of freedom = 2, n = 30.

Question	$\chi^2(2)$	*p*
Q1	10.308	0.006
Q2	17.883	0.000
Q3	16.673	0.000
Q4	15.721	0.000
Q5	0.886	0.642
Q6	12.869	0.002
Q7	16.071	0.000
Q8	10.659	0.005
Q9	12.060	0.002
Q10	12.976	0.002

Table 3: Asymp. Sig. values as reported from the pairwise comparison using Wilcoxon signed-rank test; Values with Bonferroni-corrected significant differences are highlighted in gray.

Question	*Pixel-Fade*	*Shrink-Fade*	*Pixel-Shrink*
Q1	0.002	0.350	0.067
Q2	0.000	0.543	0.001
Q3	0.001	0.853	0.000
Q4	0.004	0.792	0.004
Q5	*NA*	*NA*	*NA*
Q6	0.017	0.471	0.017
Q7	0.002	0.388	0.001
Q8	0.002	0.814	0.010
Q9	0.001	0.349	0.015
Q10	0.059	0.169	0.003

between-subjects variable and we performed Kruskal-Wallis tests on the 10 questions. We found no significant effect of media type; with one exception: Q9 showed a significant difference, with Instagram having a lower mean. We believe this single difference occurred because Q9 asked about "text posts", which Instagram does not support.

5.2 Feedback on the prototypes

The written and verbal feedback from participants aligned with the Likert scale results: shrinking was the most favourable visualization, followed closely by fading; pixelating was least favourable.

As suggested by the feedback for each prototype, detailed next, participants found the shrinking technique most visually pleasing as it looked more "natural". Moreover, it was best associated with memory and the passage of time; putting less significance on older posts by making them tinier. Participants also liked the fading visualization because the idea of graying out posts resembled how artifacts fade in real life. In both cases, the visualizations were reasonable metaphors that provided a logical parallel with their impression of how human memories work. They recognized and brought up their understanding of the metaphors without prompting.

Next, we discuss specific feedback relating to each visualization.

Prototype 1: Pixelating: The initial reaction to pixelation for fourteen participants was that there might be a glitch in the system/website or that the Internet connection was slow and pictures were not loading correctly. Mostly, participants had no idea what was going on. They reported various negative emotions, including thinking of something bad/criminal (FB6), feeling irritated (FB1), angry (IG8), and scared/lost (FB10). In addition, ten participants felt confused or annoyed. Moreover, they thought that someone using the technique on social media must be hiding something (n = 7) from specific people (e.g., non-friends), blocking someone (n = 4), or that the content had been censored (n = 2).

Participants thought that it was pointless to keep posts in such a representation, and felt that it would be better if the post was simply deleted. Overall, participants neither associated such representation with the passage of time nor found it visually appealing. Clearly, the pixelating visualization failed to convey the appropriate metaphor, and instead invoked other negative connotations.

Prototype 2: Fading: Fifteen participants found the fading effect intuitive and indicative of its purpose. In addition, it was visually appealing since the gradual fade-out inherently showed a smoother transition between posts. Eleven participants liked the idea that they could see details about the post, text in particular, compared to the pixelating and the shrinking techniques. Furthermore, the idea of fading the posts resonated for some participants (n = 12) with the metaphor of memories or physical photos fading over time.

As IG3 explained: *"[Fading is] really intuitive, and it's a nice metaphor of fading memories [...] and that's what happens to photos often, when they're older, they get faded [...] but making the pictures smaller? I didn't think of it that way [...] even the pixelated, it was effective, it's visually hard to ignore [...] I just assumed something is wrong with the image [...] so the fading is really nice."*

Nevertheless, some participants (n = 4) thought that faded out content would raise suspicion about the user, for example, suggesting that the user had something to hide. Others were unsure whether they would have guessed its purpose if they suddenly saw this visualization on their OSNs.

Prototype 3: Shrinking: Overall, the majority (n = 17) thought the shrinking approach was most intuitive and visually appealing. TW10 explains: *"It's more clever; like fuzzy memories; recent memories occupy more space in your head."*. Participants could see the appeal to instantly realizing what content is most recent without having to look at the dates. As explained by FB3: *"It's like a visual way of seeing that it's a later post [...] the way the time grows the way the grid grows, it kinda correlates that way [...] it would take time to be used to it, but if Facebook had come like this, I'd be more accustomed to it, I wouldn't really have a problem."*

However, some participants (n = 8) initially thought that bigger posts were of higher importance and relevance to the user publishing the content. They believed that the user had somehow chosen to make some posts larger, rather than realizing that size was an automatic characteristic that varied over time. The most common complaint from participants (n = 27) was being unable to have clear legibility of posts as they shrunk. However, between fading and shrinking, they thought shrinking offered better visibility.

Users' preferences: when asked to choose one visualization to be applied to their own artifacts, 14 participants favoured the shrinking prototype, 11 participants preferred fading, three participants were undecided between both prototypes, one participant wanted both combined, and one participant preferred pixelation.

5.3 RQ1 Summary

Participants expressed clear preferences for the Shrinking and Fading visualizations, and these successfully conveyed the metaphor of memories fading over time. The Pixelating visualization was disliked and held negative connotations; it did not meet the goal of representing aging of digital content. We recommend either Shrinking

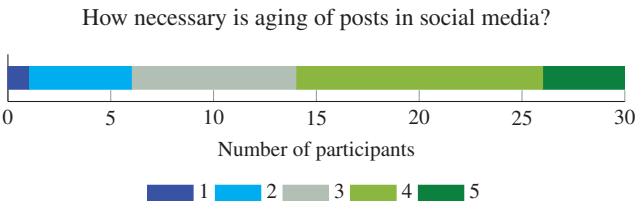


Figure 4: Responses to one of the wrap-up questions (1 = Not at all necessary, 5 = Very necessary)

or Fading as appropriate visualizations for conveying digital aging.

Interestingly, participants (who were not initially told that this study was about privacy) expressed annoyance resulting from not being able to clearly read the posts as they decayed. Some mentioned a preference for the fading technique because it enabled them to decipher the details of the posts for the longest time. So while they understood the metaphor, they still favoured the visualization which showed the least decay. We note that levels of decay could be adjusted for any of these visualizations and that in an actual implementation, the effect would appear much more gradual since there would likely be more posts in the span of a year.

6. RQ2 ANALYSIS AND RESULTS

We next summarize results of our analysis pertaining to answering our second research question: *What are users' attitudes and concerns relating to digital aging on social media?*

We concluded the session with an interview and a wrap-up questionnaire to capture participants' opinions regarding the concept of aging digital artifacts and to discuss if it would increase their online privacy. This part of the session took place after participants had seen all three visualizations and had provided their feedback about each one. The next subsections summarize the responses from the interview and the open-ended questions of the questionnaire.

6.1 Necessity of digital aging

The first question of the wrap-up questionnaire asked *"How necessary is aging of posts in social media?"*. Sixteen participants thought that digital content aging is necessary, while a third were neutral. Figure 4 shows participants' Likert-scale responses.

In our interviews, we asked participants if they would opt-in to the content decay feature for their own content, if available. Two-third of participants (n = 20) thought they would, believing that digital content should age, whether to reflect the person they are today, to depict different time periods, or to protect their online privacy. The remaining participants disagreed, or were concerned about how aging of digital artifacts would impact their access to content on their own profiles. While they thought that aging might be appropriate for others viewing their profile, they wanted to retain access to the unedited versions of their own content.

6.2 Deletion, expiration, or decay

We further discussed with participants what it means to have their social media content age, and how this should happen.

Eighteen participants recognized that social media content lost relevance as time passed. Two-thirds of participants (n = 20) wanted to either *delete* or *archive* content themselves or potentially have content *decay*. Their choice of method depended on the social media platform and the content itself. Some explicitly mentioned that they wanted to delete content when it no longer reflected their current personalities [25, 26, 29] or the impression that they wanted to convey to the world.

Secondly, two-third of participants (n = 20) saw a need for a *decay* feature on OSNs: one-third would unconditionally opt in and one-third would opt for it conditionally, i.e., if they retained some control over the operation of the decay feature. For example, if it was programmed to allow an undo of the decay, and if the decay did not apply to their own self-view. Other preferences included being able to select which decay visualization technique should be applied. Moreover, the majority (n = 17) wanted to select which content should decay rather than have it automatically executed. Their choice would depend on specific time thresholds, or the context of the content itself. Fifteen participants thought decay should depend on characteristics of the content more than on how much time has passed. In addition, eight participants wanted to choose which audience views the decayed content.

Participants (n = 25) thought *decay* would be particularly beneficial in several situations. For instance, they thought it might reduce information overload when browsing other users' profiles. They also thought it would be beneficial if they might regret deletion of specific content. As one participant explained: *"Sometimes you delete something in the spur of the moment then you think I shouldn't have deleted that [...] and there's no point in putting it back cause everybody already saw it [...] with all the comments [...] people regret deleting things." -FB3.* Others thought it would be useful for fact checking data, for keeping track of their online activity, or for archiving or compressing content. As FB4 explained: *"Maybe fact checking data, for politics and election season, sometimes it's important to check news and when they happened, which is something that's easily overlooked in social media."* Others thought it might help to keep only relevant memories and forget irrelevant ones, which might be helpful in the healing process after a breakup. As FB10 illustrated: *"Delete is [...] computation oriented, faded feels more like personal, more human, more like in my memory [...] more natural, I have that association. When you become older, you forget many things [...] right now, social media does not make any differentiation in all our memories, they are all equally relevant, and it happens that along our lives, not all our memories are equally relevant."*

Lastly, two participants thought that the only cases in which content should automatically *expire* is when the person is deceased or the profile is no longer in use. Alternatively, they suggested the family of deceased person could choose to *decay* the content instead.

6.3 Privacy

We also wanted to explore participants' perspectives on online privacy. We asked a group of participants[†] if content decay would protect their online privacy. Participants' opinions were polarized. A minority (n = 3) thought the idea does not contribute to online privacy at all. Their main concern was that concealing content would raise questions about the content or the user, hence, they found no contribution to privacy. However, most (n = 11 out of 17) thought it was the only purpose for using decay. For instance, they would decay obsolete content when seeking employment, or when beginning new chapters in their lives. As explained by FB5: *"It would be beneficial to me if I was applying for a new job, or even entering a new relationship, I would not want the company or person to be able to scroll and see my old posts and judge me by them."*

[†]We explicitly asked 17 participants at the very end of the interview. However, it was implicitly discussed with other participants.

When the eleven participants who thought decay is beneficial for privacy were asked which visualization technique is most preferable for privacy, six participants favoured the pixelating technique. They thought pixels hid the content appropriately since pixelation obscures content more quickly by nature. Four participants thought either the fading or shrinking techniques might be helpful to privacy as well, depending on how fast/gradual they decay the content. One participant did not specify a preference.

6.4 RQ2 Summary

Participants thought that digital artifacts should age to accommodate changes in their real lives. Decaying digital content was appreciated, and if available on social media, participants would opt-in to the feature. They generally found it useful for online privacy, but responses varied for which visualization they would adopt for their own accounts. Specifically for privacy, pixelation was most popular but is also held negative connotations for several participants.

7. OTHER COMMENTS

Changing of perspective: Eleven participants said that introducing the concept of aging of digital artifacts changed their perspective on how they use social media today. For instance, they intended to go through their own content, re-examine their privacy settings, and re-think which posts remain appropriate for their current lives. This aligns with previous research suggesting that conversations about privacy lead users to reflect on their own practices [2,33,58].

We observed a shift during some sessions. Participants initially were concerned about how aging of digital artifacts would affect the visibility of their content to themselves and to others. As the session progressed, they accepted the concept and realized its value for online privacy when displaying content to others.

Other participants (n = 6) expressed no major change in perspective. They were already careful with what they post, or they were accustomed to the look and feel of social media today and saw no reason to change. As one participant noted: *"If I have choice between changing and not changing, I'm not gonna change [...] if they have it changed and I'm forced, I'm not gonna change it either." -FB1*

Downsides of decay: While participants realized that the feature has merit, three participants expressed concerns. Examples in which the feature would be problematic include translating decayed content for people with accessibility issues, or when the content is needed as an evidence to verify information (i.e., in a police investigation of a criminal activity).

8. DISCUSSION

Our motivation exploring how to represent the aging of digital artifacts within the UI. We further investigated what aging of digital artifacts means for users and to what extent incorporating this concept within the UI would conform to their sharing and privacy needs. We elaborate on the privacy and design implications of our findings in the following subsections. We then translate those implications into a tentative set of system design recommendations.

8.1 Aging vs. Privacy Paradox

We found that participants' mental models of how their content should appear online depended largely on whether they were considering aging or privacy at the time. In our study, we intentionally avoided mentioning privacy until late in the session so that we could determine if privacy concerns arose unprompted.

When participants considered the management of their data in terms of aging, they favoured a gradual fading/shrinking of artifacts over time because it matched their idea that memories lose prominence over time, as suggested by human memory decay theory [11, 62]. As with real memories, they also expected the UI to differentiate between important memories or life events that are clearly remembered despite the passage of time and everyday happenings that are gradually forgotten.

They expressed that the visualization should represent the natural forgetting process and should not seem like the artifacts were being manipulated. For example, several participants specifically disliked the pixelation visualization because it suggested that something was being intentionally obscured and this raised suspicion.

When prompted to consider privacy implications of digital artifacts, we observed a shift in priorities and requirements. This aligns with previous research regarding the privacy paradox [1,2,4,5,18,33,49, 58, 66]; people do not intuitively consider privacy risks and sometimes accept them until prompted to consider privacy. Some participants felt that the pixelating visualization best reflected the idea of privacy by making it clear that something was intentionally being kept private. Pixelation fit with these participants' mental model of privacy: content was being censored or obscured. They also noted a more discrete dimension to privacy: something should be either kept private or made public. It was not necessarily viewed as a gradual process whereas "aging" was clearly gradual.

We are left with this interesting paradox: users want gradual, natural decaying of digital artifacts (with exceptions for important events) to more accurately reflect human memory, but at the same time want discrete, intentional private/public representation of artifacts to reflect their concept of privacy. For participants, these were two distinct requirements, whereas the literature generally views them as closely related [5, 45, 50, 51].

In both cases, however, participants recognized the benefits of removing irrelevant content and recognized that their preference for the visibility of specific digital artifacts would likely change over time. The question remains: how do we best reconcile these two distinct intentions while displaying digital artifacts in OSNs?

8.2 {Self <—> Public} Spectrum

Participants require distinct rules when representing aging on their *self* profiles versus their *public* profiles. Aligned with previous research [35, 43], participants wanted their own content to always be visible to themselves. However, they then had complex rules for how their content should be displayed to different user groups. Those rules differ significantly depending on the category of the content published on their profiles and the intended audience.

Although this was not our intention, participants re-iterated that they expected that the representation of profiles should not be automatically altered to represent aging when *self* accessed. Normally, participants use the web and OSNs as backup repositories to retain their digital possessions [43]. Our participants were concerned that their view of their own data would be altered or the data would become inaccessible without their consent, losing access to the artifacts representing these milestones. Therefore, when the *self* UI visualizes aging, the default representation should not decay content. While not the intended purpose of decay, the discussion does serve as an anchor for participants' explanation of how things should work for content viewable by others.

When being accessed by the *public/others*, participants desired different rules. Because they are concerned about their online presence and their availability to other online users, it is important that their content is visible to their audience. However, they wish to

manage the visibility and aging settings of their online content for both availability and privacy purposes. In this case, the audience comprises a spectrum of *closest friends*, *specific circles of friends*, and moving outwards to the *public*. Participants wished to consider two main factors when visualizing aging on the UI for other audiences: (1) the context/category of the published content and (2) where its intended audience falls on the spectrum. Other practices [41] in the online photo sharing domain similarly adopt a privacy framework by controlling two elements: content and recipient. Indeed the two factors are significant determinants of privacy [41] since some online artifacts are more personal in nature than others [30] (e.g., a self-portrait versus a photo of a landscape). However, the rules are individualized to each user and can be complex as they encompass all possible scenarios and exceptions. Moreover, rules changed dynamically based on specific contexts or based on exceptions for a specific audience. For example, Joe might enjoy sharing his life memories with others, but Jim prefers having personal photos or embarrassing photos decay when viewed by work colleagues and unmodified when accessed by family members or close friends. Complexity might further increase if Jim also wanted the same artifacts decayed when viewed by a cousin and unaltered for a specific work colleague. Accurately reflecting users' real intentions could quickly become untenable.

This suggests that incorporating controls into the UI that maintain such rules becomes an added effort for users. Firstly, it is impractical that each user can internalize all their desired rules and adjust the rules whithin the UI whenever they publish new content. Secondly, because the desired rules change as time passes and circumstances change, it is unlikely that a system could generalize these rules to match the preference of every user. This leaves us with another question: should we integrate such complex functionality for controlling the display of digital artifacts in OSNs and can we do so without adding undue effort to users?

8.3 Privacy as an intangible subject

The literature show that although users rationally accept privacy risks as a trade-off for the benefits of online sharing, they also express an intuitive concern when prompted [1, 2, 4, 5, 18, 33, 49, 58, 66]. Very few of our participants initially realized the privacy merit of content decay, but opinions evolved throughout the sessions, as presented in Sections 6.3 and 7 (Change in perspective). Initially, participants who favourably viewed content decay said they would opt-in for different purposes. For example, they wanted it as a way of compressing, keeping track of their activity, or forgetting specific memories. Privacy is an intangible subject [33] to users; our participants did not intentionally ignore it, but rather it did not immediately occur to them. However, when prompted about privacy [33] and the ways in which aging of digital artifacts contributes to privacy, they started to realize its potential added value.

In some instances, privacy could be viewed as a positive by-product of decaying content. Some users liked the idea of decaying digital artifacts for reasons other than privacy (e.g., it makes it easier to quickly tell how recently information was posted). These users might be persuaded to adopt the visualization due to its perceived usability benefits, but subsequently also gain privacy benefits with no additional effort.

The literature has shown that some Facebook users manage their privacy by trusting their abilities in manually controlling information being shared [1]; few changed Facebook's default privacy settings [24]. Our participants thought they would simply *delete* what they no longer wanted available online. Although they expressed interest in retaining detailed control, practically speaking and, as shown in the literature [2,24,66], this is unlikely. Moreover, even if participants had the time and initiative to delete old content, this is actually very difficult to do in OSNs; for example, Facebook only loads a bit of data at a time, in reverse chronological order. And even though the "activity log" allows a user to review older content by year, there is no way to easily access and manage that content. Our participants thought that after the study, they would revise their own OSN content and delete what is no longer relevant. However, this intention only arose because they were specifically primed to consider the privacy of their OSN data [2]. This suggests that normally users remain indifferent to the need to perform retrospective privacy management.

8.4 Preliminary Recommendations

Based on the literature and our findings, we provide the following recommendations. Given that this study has raised additional questions and other aspects should be explored, these recommendations are preliminary in nature and intended to fuel further discussion.

R1: Have digital decay features enabled by default as a fail-safe mechanism: A principle of usable security and privacy is to include the safest outcome in the path-of-least-resistance since it is likely what users will choose [63,78]. Given that the ultimate path-of-least-resistance for users is to do nothing [78], system settings should be secure by default [63, 78]. The privacy paradox [1, 2, 4, 5, 18, 33, 49, 58, 66] also suggests that users' actions rarely match their privacy intentions. Thus, fail-safe decay mechanisms could at least partially protect users from their unintended self-disclosure on public profiles. This further aligns with the Privacy-by-Design principles [14] of having preventative and default measures.

As a result, users would be mostly relieved of the burden associated with retrospectively managing their digital artifacts. Digital aging gives temporal context to the viewer and emphasizes content that is currently most timely, indirectly supporting their online privacy by gradually removing content from the public sphere as it ages.

In practice, the aggressiveness of the decay algorithm could be increased for additional privacy protection and to avoid possible reversal of deteriorated posts at the early stages of decay. As shown in the literature on using redaction for visual privacy [60], increasing the strength of a privacy filter [19,36] and the masked area [19] increases privacy. Although digital decay does not address all aspects of online privacy, such a fail-safe mechanism could be a key component to minimize the negative consequences associated with long-term availability of OSN digital artifacts.

R2: Match the aging metaphor: Metaphors are a helpful tool that serve humans' cognitive functions [22] and metacognitive strategies [12]. Metaphors link an abstract concept to a concrete concept [22], allowing extraction of common properties from both concepts to better understand the abstract concept [22]. Metaphors have had a radical impact on interface design practices [47]. The use of metaphors in the UI can reduce the mismatch between the designer's intention and the user's mental model of the system [47].

As discussed in Section 8.1, participants felt that visualizations for aging of digital artifacts should reflect the natural forgetting process. Based on our early findings, the shrinking and fading visualizations were found to best depict the metaphor of decaying memories [11,62] and could be used either individually or potentially in combination. However, if a system designer is faced with selecting only one approach, shrinking would be recommended since it was most preferred by participants and was thought to be most intuitive

and natural. Other research suggests that visualizations such as pixelation or blurring are actually ineffective at preserving privacy of social media photos [41]. Our study found that the pixelation visualizations were interpreted as "concealing"; they invoked negative connotations and aroused suspicion. Taken together, these results suggest that pixelation should be avoided as a method for increasing privacy. Online sharing and privacy are guided by complex social norms and expectations [48]; any visualization used should be carefully implemented to ensure that it does not inadvertently make the user appear as if they are breaking such social norms.

R3: Allow overrides: Users should be allowed to override decay defaults, if they wish to. As suggested in R1, the default settings should be secure, but allowing users to have control over their content is also important. By allowing overrides, users can disable the feature or adjust settings to perform more selective decaying [25] and to control the decay rate [7] based on the context and specific online content. Whereas automating such privacy decisions may be desirable, the complex, personal, and dynamic nature of these decisions makes it unlikely that they can be performed algorithmically in a fully automated way. In particular, the risks of miscategorization could lead to privacy violations if the user expects something to automatically decay and it does not.

Given these constraints, users should remain involved in decisions to make some digital artifacts visible beyond the normal decaying period, or to avoid the decaying process altogether. It is possible that they could be assisted by the system, but the ultimate choice should rest with the user and involve a distinct, conscious decision by the user that enables them to reflect on their intended privacy and sharing needs. This could also support existing recommendations [4,8,80] suggesting that the UI should promote user reflection of aged content.

We believe our early recommendations align with Principles 1, 2, 3, and 7 of the Privacy-by-Design framework [14]. Our recommendations place privacy as a core function of the user interaction (*Principle 3; privacy embedded into the design*) by reducing the long-term exposure of digital artifacts and reducing risks of privacy violations (*Principle 1; proactive not reactive*). They seek to insert privacy into the design of OSNs by default as a fail-safe feature (*Principle 2; privacy as the default setting*). The recommendations aim to maximize privacy defaults, while giving users granular privacy options to customize their privacy preferences based on their privacy and sharing requirements (*Principle 7; keep it user-centric*). By supporting the aging metaphor, the recommendations also focus on matching users' mental models as closely as possible (*Principle 7; keep it user-centric*).

8.5 Feasibility

Within OSNs, several implementation issues would need to be addressed when implementing decay visualizations. First, digital content shared on OSNs may not be exclusively controlled by its publisher/owner [71]. For example, other users may be tagged in a post, or content may be re-shared by other users. In these cases, and cases where multiparty access control is required [71], it is unclear what should happen to decaying content. Do all instances decay at the pace set by the original owner? Should other users be able to override decay settings? What happens if content is reposted/shared after significant time has elapsed? Does it reset to full visibility or get posted partially decayed?

Another significant concern is that traces of the digital content might still be available elsewhere outside the original OSN. For example, content may be copied or downloaded by others before the decaying process begins, leaving unaltered instances of the digital artifacts. The owner of the content may also have shared copies of content on other mediums. Thus, the feasibility of decaying social media digital artifacts might be limited when considering other aspects of online sharing and availability of online data.

9. LIMITATIONS AND FUTURE WORK

The study had the usual limitation common to lab studies; asking participants to share feedback about a partially unfamiliar concept in a limited amount of time in an artificial environment. A future field study could be designed to complement our findings. Furthermore, the sample size of thirty participants might be small when considering that they were divided across three social media platforms (although every participant saw all three visualizations), and the university sample of users is not necessarily representative of the whole population. Additionally, when designing the study prototypes, we distributed fewer than 20 posts across a year to more easily and clearly show the effect of decay. Had we added more posts to the prototypes, the change in visualizations would have appeared more gradual, which could have impacted participants' opinions. We chose to use artificial data in the prototypes rather than applying the visualizations to the users' own content. This may have made the content seem more abstract to participants since it was disconnected from any particular context or personal connection. However, this design decision was taken because protecting the privacy of participants was viewed as more important than the slight methodological advantage to be gained in these early stages of the research. The study offers a starting point in empirically testing visualizations for aging of digital artifacts. Further comparison with other visualizations should also be considered.

This research has led to several possible future research directions. Further research could explore design or technical aspects of implementing decay features or examine how feasibility limitations of the feature can be addressed in specific online sharing scenarios. It could also subjectively [41,60] or objectively [36] evaluate technical privacy protection offered by the decay visualizations, and explore whether decay visualizations might lead users to become less proactive in managing their online content (e.g., by leaving content online rather than deleting it).

The design of a future study could consider other relevant aspects such as information sensitivity, access control options, different types of artifacts, and parameters of the decay function. It could include scenarios to help users with specific contexts, and to provide insights into identifying the primary factor for choosing to decay artifacts: characteristics of the artifacts or their age. It could also consider using real social media data that is connected to participants. Another future study could empirically examine how aging digital artifacts on an OSN profile affects viewers' impression of its owner. This could be explored in several different social contexts: political, employment, or relationships/dating contexts.

10. CONCLUSIONS

We conducted a lab user study exploring the concept of aging or decaying of digital artifacts and reported results from both qualitative and quantitative analyses. Results showed an inclination towards visualizations that closely represent fading memories over time. Because of the nature of human memory, and users' mental model of privacy, we identified distinct user requirements when addressing either aging or privacy in the UI. These two distinct purposes should be further explored to determine how they can be best reconciled in interaction design.

A balanced approach to addressing users' requirements would seek to promote privacy while minimizing user effort and simultaneously enabling user reflection. Towards this goal, we provided three preliminary design recommendations. Although decay features do not address every aspect of online privacy and long-term data availability dimensions, it can help minimize the potential unintended consequences associated with data availability on OSNs.

To summarize, this work compares three OSN content decay visualizations, investigates users' attitudes and concerns about the aging of digital artifacts, and provides early recommendations that would contribute to users' privacy and sharing needs. We also believe the study is a step towards answering currently open research questions pertaining to visualizing passage of time in OSNs.

11. ACKNOWLEDGMENTS

We thank our study participants and paper reviewers for their time and valuable feedback. R. Mohamed acknowledges funding from the Ontario Trillium Scholarship (OTS) program. S. Chiasson acknowledges funding from NSERC for her Canada Research Chair and Discovery Grant.

12. REFERENCES

[1] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *Privacy Enhancing Technologies*, pages 36–58. Springer Berlin, 2006.

[2] A. Acquisti and J. Grossklags. Losses, gains, and hyperbolic discounting: An experimental approach to information security attitudes and behavior. *2nd Annual Workshop on Economics and Information Security*, pages 1–27, 2003.

[3] E. Andrew-Gee. Gaffes: When candidates accidentally tell the truth. `http://www.theglobeandmail.com/news/politics/elections/gaffes-when-a-candidate-accidentally-tells-the-truth/article26713595/`, 2015.

[4] O. Ayalon and E. Toch. Retrospective privacy: Managing longitudinal privacy in online social networks. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 1–13. ACM, 2013.

[5] O. Ayalon and E. Toch. Not even past: Information aging and temporal privacy in online social networks. *Human-Computer Interaction*, 32(2):73–102, 2017.

[6] A. Barth, A. Datta, J. C. Mitchell, and H. Nissenbaum. Privacy and contextual integrity: Framework and applications. In *Proceedings of the 2006 IEEE Symposium on Security and Privacy*, SP '06, pages 184–198. IEEE Computer Society, 2006.

[7] D. Barua, J. Kay, B. Kummerfeld, and C. Paris. Theoretical foundations for user-controlled forgetting in scrutable long term user models. In *Proceedings of the 23rd Australian Computer-Human Interaction Conference*, OzCHI '11, pages 40–49. ACM, 2011.

[8] L. Bauer, L. F. Cranor, S. Komanduri, M. L. Mazurek, M. K. Reiter, M. Sleeper, and B. Ur. The post anachronism: The temporal dimension of facebook privacy. In *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society*, WPES '13, pages 1–12. ACM, 2013.

[9] P. Best, R. Manktelow, and B. Taylor. Online communication, social media and adolescent wellbeing: A systematic narrative review. *Children and Youth Services Review*, 41:27–36, 2014.

[10] K. Borcea-Pfitzmann, A. Pfitzmann, and M. Berg. Privacy 3.0 := data minimization + user control + contextual integrity. *Information Technology*, 53:34–40, 2011.

[11] M. E. Bouton, J. B. Nelson, and J. M. Rosas. Stimulus generalization, context change and forgetting. *Psychological Bulletin*, 125:171–186, 1999.

[12] L. Bowler and E. Mattern. Visual metaphors to model metacognitive strategies that support memory during the process of refinding information. In *Proceedings of the 4th Information Interaction in Context Symposium*, IIIX '12, pages 250–253. ACM, 2012.

[13] M. Burke, C. Marlow, and T. Lento. Social network activity and social well-being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1909–1912. ACM, 2010.

[14] A. Cavoukian. Privacy by design, the 7 foundational principles: Implementation and mapping of fair information practices. *Information and Privacy Commissioner of Ontario, Canada*, 2010.

[15] F. Chanchary and S. Chiasson. User perceptions of sharing, advertising, and tracking. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 53–67. USENIX Association, 2015.

[16] T. Chen, R. Boreli, M.-A. Kaafar, and A. Friedman. On the effectiveness of obfuscation techniques in online social networks. In E. De Cristofaro and S. J. Murdoch, editors, *Privacy Enhancing Technologies*, pages 42–62. Springer International Publishing, 2014.

[17] C. Conley. The right to delete. In *Spring Symposium Series*, pages 53–58. AAAI, 2010.

[18] K. P. Coopamootoo and T. Groß. Why privacy is all but forgotten: An empirical study of privacy & sharing attitude. *Privacy Enhancing Technologies*, 2017 (4):97–118, 2017.

[19] J. Demanet, K. Dhont, L. Notebaert, S. Pattyn, and A. Vandierendonck. Pixelating familiar people in the media: Should masking be taken at face value? *Psychologica Belgica*, 47:261–276, 2007.

[20] eBizMBA Guide. Top 15 most popular social networking sites. `http://www.ebizmba.com/articles/social-networking-websites/`, 2017.

[21] C. Elsden, D. S. Kirk, and A. C. Durrant. A quantified past: Toward design for remembering with personal informatics. *Human Computer Interaction*, 31(6):518–557, 2016.

[22] D. Gentner and B. Bowdle. *Metaphor Processing, Psychology of.* (John Wiley & Sons, Hoboken, 2006.

[23] S. R. Greysen, T. Kind, and K. C. Chretien. Online professionalism and the mirror of social media. *Journal of General Internal Medicine*, 25(11):1227–1229, 2010.

[24] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, WPES '05, pages 71–80. ACM, 2005.

[25] R. Gulotta, W. Odom, J. Forlizzi, and H. Faste. Digital artifacts as legacy: Exploring the lifespan and value of digital data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1813–1822. ACM, 2013.

[26] O. L. Haimson, J. R. Brubaker, L. Dombrowski, and G. R. Hayes. Digital footprints and changing networks during online identity transitions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2895–2907. ACM, 2016.

[27] R. Harper, E. Whitworth, and R. Page. Fixity: Identity, time

and durée on facebook. *Selected Papers of Internet Research (SPIR)*, IR:1–21, 2012.

[28] A. Heravi, D. Mani, K.-K. R. Choo, and S. Mubarak. Making decisions about self-disclosure in online social networks. In *Proceedings of Hawaii International Conference on System Sciences (HICSS)*, pages 1922–1932. ScholarSpace, 2017.

[29] B. Hogan. The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society*, 30(6):377–386, 2010.

[30] R. Hoyle, R. Templeman, S. Armes, D. Anthony, D. Crandall, and A. Kapadia. Privacy behaviors of lifeloggers using wearable cameras. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, pages 571–582. ACM, 2014.

[31] H.-F. Hsieh and S. E. Shannon. Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9):1277–1288, 2005.

[32] P. Ilia, I. Polakis, E. Athanasopoulos, F. Maggi, and S. Ioannidis. Face/off: Preventing privacy leakage from photos in social networks. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 781–792. ACM, 2015.

[33] L. K. John. *The Consumer Psychology of Online Privacy: Insights and Opportunities from Behavioral Decision Theory*. Cambridge University Press, 2015.

[34] P. Kallas. Top 15 most popular social networking sites and apps. `https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/`, 2017.

[35] D. S. Kirk and A. Sellen. On human remains: Values and practice in the home archiving of cherished objects. *ACM Transactions on Computer-Human Interaction*, 17(3):10:1–10:43, 2010.

[36] P. Korshunov, A. Melle, J.-L. Dugelay, and T. Ebrahimi. Framework for objective evaluation of privacy filters. *Proc.SPIE*, 8856:8856–12, 2013.

[37] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[38] N. Kotamraju, S. Allouch, and K. van Wingerden. *Employers' Use of Online Reputation and Social Network Sites in Job Applicant Screening and Hiring*, pages 247–269. Greyden Press, 2014.

[39] K.-T. Lee, M.-J. Noh, and D.-M. Koo. Lonely people are no longer lonely on social networking sites: The mediating role of self-disclosure and social support. *Cyberpsychology, behavior and social networking*, 16:413–8, 2013.

[40] Y. Li, A. Kobsa, B. P. Knijnenburg, and M.-H. Carolyn Nguyen. Cross-cultural privacy prediction. *Proceedings on Privacy Enhancing Technologies*, 2017(2):113–132, 2017.

[41] Y. Li, N. Vishwamitra, H. Hu, B. P. Knijnenburg, and K. Caine. Effectiveness and users' experience of face blurring as a privacy protection for sharing photos via online social networks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61(1):803–807, 2017.

[42] S. Lindley, R. Corish, E. Kosmack Vaara, P. Ferreira, and V. Simbelis. Changing perspectives of time in hci. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 3211–3214. ACM, 2013.

[43] S. Lindley, C. C. Marshall, R. Banks, A. Sellen, and T. Regan. Rethinking the web as a personal archive. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 749–760. ACM, 2013.

[44] M. Madejski, M. L. Johnson, and S. M. Bellovin. The failure of online social network privacy settings. Technical report, Columbia University Academic Commons, 2011.

[45] V. Mayer-Schönberger. *Delete: The Virtue of Forgetting in the Digital Age*. Princeton University Press, 2011.

[46] R. E. Mohamed, T. B. Idalino, and S. Chiasson. When private and professional lives meet: The impact of digital footprints on employees and political candidates. In *Proceedings of the 8th International Conference on Social Media & Society*, #SMSociety17, pages 48:1–48:5. ACM, 2017.

[47] D. Neale and J. Carroll. *The Role of Metaphors in User Interface Design*. Elsevier, Amsterdam, 1997.

[48] H. Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

[49] P. Norberg, D. R. Horne, and D. Horne. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, 41:100–126, 2007.

[50] A. Novotny. Signs of time: Designing social networking site profile interfaces with temporal contextual integrity. In *Human Aspects of Information Security, Privacy, and Trust*, pages 547–558. Springer International Publishing, 2015.

[51] A. Novotny and S. Spiekermann. Oblivion on the web: an inquiry of user needs and technologies. In *European Conference on Information Systems*, pages 1–15, 2014.

[52] A. Novotny and S. Spiekermann. Oblivion of online reputation: how time cues improve online recruitment. *International Journal of Electronic Business*, 13(2-3):183–204, 2017.

[53] W. Odom, R. Banks, D. Kirk, R. Harper, S. Lindley, and A. Sellen. Technology heirlooms?: Considerations for passing down and inheriting digital materials. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 337–346. ACM, 2012.

[54] E. Parliament and Council. Directive 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Union L281*, 38:31–50, 1995.

[55] E. Parliament and Council. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *Official Journal of the European Union L119*, pages 1–88, 2016.

[56] S. T. Peesapati, V. Schwanda, J. Schultz, M. Lepage, S.-y. Jeong, and D. Cosley. Pensieve: Supporting everyday reminiscence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2027–2036. ACM, 2010.

[57] A. J. Perez and S. Zeadally. Design and evaluation of a privacy architecture for crowdsensing applications. *SIGAPP Appl. Comput. Rev.*, 18:7–18, 2018.

[58] 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*, CHI '16, pages 5240–5251. ACM, 2016.

[59] Y. Pu and J. 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.

[60] J. Ramón Padilla-López, A. Chaaraoui, and F. Flórez-Revuelta. Visual privacy protection methods: A survey". *Expert Systems with Applications*, 42:4177–4195, 2015.

[61] S. Reddy and K. Knight. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26. Association for Computational Linguistics, 2016.

[62] T. J. Ricker, E. Vergauwe, and N. Cowan. Decay theory of immediate memory: From brown (1958) to today (2014). *Quarterly Journal of Experimental Psychology*, 69(10):1969–1995, 2016.

[63] J. H. Saltzer and M. D. Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975.

[64] V. Schwanda Sosik, X. Zhao, and D. Cosley. See friendship, sort of: How conversation and digital traces might support reflection on friendships. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1145–1154. ACM, 2012.

[65] E. Shugerman. 10 students have harvard acceptances withdrawn over facebook memes. `http://www.independent.co.uk/news/world/americas/harvard-facebook-memes-student-acceptance-taken-away-withdrawn-university-a7775991.html`, 2017.

[66] S. Spiekermann, J. Grossklags, and B. Berendt. E-privacy in 2nd generation e-commerce: Privacy preferences versus actual behavior. In *Proceedings of the 3rd ACM Conference on Electronic Commerce*, EC '01, pages 38–47. ACM, 2001.

[67] D. J. Stute. Privacy almighty? the cjeu's judgment in google spain sl v. aepd. *Michigan Journal of International Law*, 36(4):649–680, 2015.

[68] K. Subrahmanyam and D. Smahel. *Digital Youth: The Role of Media in Development*. Springer-Verlag New York, 2011.

[69] E. Thiry, S. Lindley, R. Banks, and T. Regan. Authoring personal histories: Exploring the timeline as a framework for meaning making. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1619–1628. ACM, 2013.

[70] H. van Rossum, H. Gardeniers, J. Borking, A. Cavoukian, J. Brans, N. Muttupulle, and N. Magistrale. *Privacy-Enhancing Technologies: The Path to Anonymity*. Information and Privacy Commissioner / Ontario, Canada & Registratiekamer, The Netherlands, 1995.

[71] N. Vishwamitra, Y. Li, K. Wang, H. Hu, K. Caine, and G.-J. Ahn. Towards pii-based multiparty access control for photo sharing in online social networks. In *Proceedings of the 22nd ACM on Symposium on Access Control Models and Technologies*, SACMAT '17 Abstracts, pages 155–166. ACM, 2017.

[72] J. Vitak, P. Wisniewski, Z. Ashktorab, and K. Badillo-Urquiola. Benefits and drawbacks of using social media to grieve following the loss of pet. In *Proceedings of the 8th International Conference on Social Media & Society*, #SMSociety17, pages 23:1–23:10. ACM, 2017.

[73] M. Vroman and K. Stulz. Employer liability for using social media in hiring decisions. *Journal of Social Media for Organizations*, 3:1–13, 2016.

[74] Y. Wang, S. Komanduri, and P. Giovanni. "i regretted the minute i pressed share: A qualitative study of regrets on facebook". In *Symposium on Usable Privacy and Security (SOUPS)*, pages 10:1–10:16. ACM, 2011.

[75] Y. Wang, P. G. Leon, A. Acquisti, L. F. Cranor, A. Forget, and N. Sadeh. A field trial of privacy nudges for facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2367–2376. ACM, 2014.

[76] Y. Wang, P. G. Leon, K. Scott, X. Chen, A. Acquisti, and L. F. Cranor. Privacy nudges for social media: An exploratory facebook study. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 763–770. ACM, 2013.

[77] Y. Yao, D. Lo Re, and Y. Wang. Folk models of online behavioral advertising. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1957–1969. ACM, 2017.

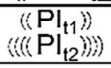
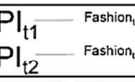
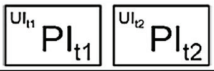
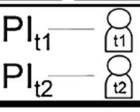
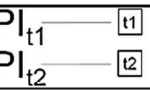
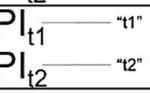
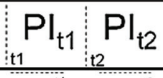
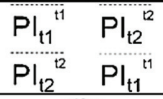
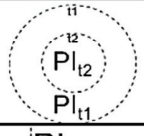
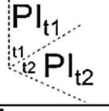
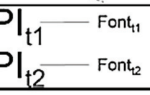
[78] K.-P. Yee. User interaction design for secure systems. In *Proceedings of the 4th International Conference on Information and Communications Security*, ICICS, pages 278–290. Springer-Verlag, 2002.

[79] X. Zhao and S. Lindley. Curation through use: Understanding the personal value of social media. In *SIGCHI conference on Human Factors in computing systems (CHI 2014)*, pages 2431–2440. ACM, 2014.

[80] X. Zhao, N. Salehi, S. Naranjit, S. Alwaalan, S. Voida, and D. Cosley. The many faces of facebook: Experiencing social media as performance, exhibition, and personal archive. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1–10, 2013.

APPENDIX

APPENDIX A: NOVOTNY'S SUGGESTED TEMPORAL SIGNS FOR SOCIAL NETWORKING SITES (SNS) - TABLE REPRINTED FROM [47].

Temporal sign	Explanation	High time granularity	High amount of PI	Dynamic generation	Generic representation of UI design
Temporal indices					
Sedimentation	Older PI_{t1} is covered by newer layers of PI_{t2}	x	x	x	
Display salience					
Size	Older PI_{t1} is displayed in smaller size	x	x	x	
Motion	Newer PI_{t2} moves faster on screen			x	
Degrading display quality					
Decay	Older PI_{t1} is displayed in decayed state		x	x	
Greying	Older PI_{t1} is displayed in lighter gray color		x	x	
Outdated display technology	Older PI_{t1} is displayed using earlier technology		x	x	
Fashion					
Fashion of content	Fashion of old PI_{t1}'s content is adapted to earlier time		x		
Fashion of UI design	Older PI_{t1} is displayed using old-fashioned UI design		x	x	
Historic snapshot of the person	Older PI_{t1} is displayed together with old profile picture		x	x	
Temporal symbols					
Symbolic objects					
Time pictograms	Temporal context is annotated using a graphical symbol	x	x	x	
Textual symbols	Temporal context is annotated using text or dates	x	x	x	
Screen space	Screen areas containing PI are assigned to time segments				
Horizontal	Older PI_{t1} is displayed left of newer PI_{t2}	x	x	x	
Vertical	Older PI_{t1} is displayed on top or bottom	x	x	x	
Concentric	Newer PI_{t2} is displayed closer to the screen's center	x		x	
Radial	Radial sections of the screen are assigned clock-wise to newer PI			x	
Typography	Older PI_{t1} is displayed in typefaces perceived as classic		x	x	

PI … SNS profile information, t_i … temporal context at point in time *i*

APPENDIX

APPENDIX B: STUDY TASKS AND INTERVIEW QUESTIONS

Part 1: Basic tasks and questions per each prototype

Example Study Tasks:

1. Browse post history by year.
2. Click on any year you wish to drill down through.
3. Click on any month of the year for its posts to be displayed.
4. Scroll through the displayed posts.

Example probing questions asked during or after task completion:

- Can you explain your interpretation of this visual representation of posts?
- Is such arrangement/representation of posts appealing to you?
- What do you like about such interface? / What worked well for you with this design?
- What don't you like? / What was most annoying or confusing to you?
- What would you change?
- Are any features missing?

To conclude this part of the study:

Which interface do you think is most:

- Helpful or useful
- Appealing or making sense to you?

APPENDIX

Part 2: Interview questions (after they've used the 3 prototypes)

A. As a user browsing another friend's page:

- What was your interpretation when you see posts fading away?
- What was your reaction when you see posts fading away?
- Did you care about seeing the original post? – when posts fade away, did that make you more curious/doubtful?
- Which technique/visual representation was more helpful in showing the decay/aging of posts?

B. As an owner of the profile:

- Would you opt for decaying/fading posts as they're getting older?
- How would you like your posts to decay, which technique was most likable to you?
- At what point, if any, would you stop caring about such artifacts/posts – when they're 1 year old? 3? 5? 10?
- In which cases do you think digital artifacts should expire/disappear? Should they expire? How? By decaying? Or by deleting forever?
- Would you prefer having the option to keep old posts the same without decaying as a way to reminiscing or highlighting a blast from the past?
- Would you want the process of decaying to be automated? Or manual? What kinds of settings would you want?
    - Select specific posts to decay based on: time of publishing, specific keywords in the caption/status, pictures taken with specific friends, posts/pictures with specific location?
- Did our study change the way you browse social media today?
- Do you think decaying can protect your online privacy? If so, which visualization from the ones you saw today would you use for privacy?

APPENDIX

APPENDIX C: STUDY QUESTIONNAIRE

Prototype A questions

Each are 5-point scales

1. The visual representation of posts easily shows that they are getting old.
( ) Strongly agree to ( ) Strongly disagree ( ) Prefer not to answer

2. The visual representation of posts was.
( ) Very meaningful to ( ) Not at all meaningful ( ) Prefer not to answer

3. The visual representation of posts was.
( ) Very confusing to ( ) Very understandable ( ) Prefer not to answer

4. The visual representation of posts was.
( ) Very complete to ( ) Missing many features that I expected
( ) Prefer not to answer

5. The visual representation of posts made me change my perspective on how I use social media today.
( ) Major change in perspective to ( ) No change in perspective
( ) Prefer not to answer

6. The aging technique used in the posts was.
( ) very appropriate for the content to ( ) did not apply to the content at all
( ) Prefer not to answer

7. The visual representation of posts was.
( ) Very obtrusive to ( ) Not at all obtrusive ( ) Prefer not to answer

8. The visual representation of photo posts was intuitive to me.

( ) Very intuitive to ( ) Not at all intuitive ( ) Prefer not to answer

9. The visual representation of text posts was intuitive to me.

( ) Very intuitive to ( ) Not at all intuitive ( ) Prefer not to answer

10. If available, I would choose to use this visual representation for my social media account.
( ) Strongly agree to ( ) Strongly disagree ( ) Prefer not to answer

Prototype B questions
[Same questions above to be copied]

Prototype C questions
[Same questions above to be copied]

APPENDIX

APPENDIX D: WRAP-UP QUESTIONNAIRE

How necessary is aging of posts in social media?
( ) Very necessary to ( ) Very unnecessary ( ) Prefer not to answer

If available, would you choose to have your posts age? Why or why not?

Can you describe a situation where aging of posts would have been particularly beneficial to you?

Can you describe a situation where aging of posts would have been particularly problematic for you?

"I've Got Nothing to Lose": Consumers' Risk Perceptions and Protective Actions after the Equifax Data Breach

Yixin Zou, Abraham H. Mhaidli, Austin McCall, Florian Schaub
School of Information
University of Michigan
{yixinz, mhaidli, mccallau, fschaub}@umich.edu

ABSTRACT

Equifax, one of the three major U.S. credit bureaus, experienced a large-scale data breach in 2017. We investigated consumers' mental models of credit bureaus, how they perceive risks from this data breach, whether they took protective measures, and their reasons for inaction through 24 semi-structured interviews. We find that participants' mental models of credit bureaus are incomplete and partially inaccurate. Although many participants were aware of and concerned about the Equifax breach, few knew whether they were affected, and even fewer took protective measures after the breach. We find that this behavior is not primarily influenced by accuracy of mental models or risk awareness, but rather by costs associated with protective measures, optimism bias in estimating one's likelihood of victimization, sources of advice, and a general tendency towards delaying action until harm has occurred. We discuss legal, technical and educational implications and directions towards better protecting consumers in the credit reporting system.

1. INTRODUCTION

In the United States, credit bureaus (also called credit reporting agencies) are private, for-profit organizations that create aggregated reports of individual consumers' credit information. They offer this information as a service to businesses that need to assess their customers' creditworthiness. For instance, lenders use credit reports and credit scores to determine whether they approve a loan and at what interest rate; landlords may check credit scores before offering a lease for an apartment; employers may consider credit reports in hiring decisions [27]. As such, credit bureaus play a significant role in the lives of U.S. residents by impacting their access to many necessities. In the United States, there are hundreds of credit bureaus serving specialized credit reporting needs. The biggest among them are the three National Consumer Reporting Agencies (NCRAs) [15]: Equifax, Experian and TransUnion.

In 2017, Equifax suffered a large-scale data breach that resulted in hackers stealing sensitive data of over 146.6 million consumers [45]. The data stolen included names, social security numbers, birth dates, addresses, and driver's license numbers, along with credit card numbers for about 209,000 consumers and dispute documents for another 182,000 consumers [38].

The size, scale and potential consequences of this data breach are unprecedented: the 2017 Equifax breach put almost half of the U.S. population at risk of identity theft. Defined as "the unlawful use of another's identifying information for gain" [89], identity theft often manifests itself through fraudulent use of existing accounts (e.g., credit card, telephone, online and insurance) [40], opening of new accounts or credit lines in the victim's name, as well as non-financial crimes [62]. In 2014, about two-thirds of identity theft victims experienced an average financial loss of $1,343, and about 40% of identity theft victims reported emotional distress resulting from the incident [40].

Despite the identity theft risks posed by the Equifax breach, evidence suggests that consumers took little to no protective action after it became public. Surveys following the breach conducted by Credit Sesame, a credit monitoring site aggregating consumer data from TransUnion, showed that 10 days after the breach was announced in September 2017, only 0.44% of credit reports at TransUnion had a credit freeze on file — a slight 0.8% increase from June 2017 [18]. The percentage of consumers who placed effective credit freezes, i.e., freezing their credit reports at all three major bureaus, would only be smaller. While a credit freeze restricts access to one's credit report and is associated with fees in many states, fraud alerts, which are free, had not been used by most consumers either. The Credit Sesame report found that only 7% of its members had a fraud alert on their credit report at TransUnion as of September 2017 [18].

To investigate the seeming contradiction between the severity of the Equifax data breach and the apparent lack of action by consumers, we conducted semi-structured interviews with 24 participants to gain insights on people's mental models of credit bureaus (how credit bureaus work, how credit bureaus collect/use data, etc.), risk perceptions of identity theft, the protective actions they took in response to the Equifax data breach, and reasons for inaction.

Our key findings show that (1) participants' mental models of credit bureaus varied regarding perceived purpose and information flows. (2) The majority of participants were generally aware of the Equifax data breach and the resulting risks, but most did not take protective action after the

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

breach. (3) We find that this inaction was not primarily influenced by accuracy of mental models or risk awareness, but rather by costs associated with protective measures, optimism bias in estimating one's likelihood of victimization, and a general tendency towards delaying action until harm has occurred. (4) Sources of advice appeared to be an influential factor in initiating actions; many participants who took action acted on advice from people they trust. Yet, taken actions also created a false sense of security for some, leading them to overlook other measures.

Based on our findings, we conclude that current protective measures offered by credit bureaus are insufficient to protect consumers. We discuss our findings in the context of prior research on privacy and security behavior, and suggest technical, legal and educational approaches to better protect consumer credit data and empower consumers with usable protection measures.

2. BACKGROUND

As context for our study, we first give an overview of how the U.S. credit reporting system operates; relevant regulation, protective measures; and the current state of data breaches and identity theft in the United States.

2.1 The U.S. Credit Reporting System

The U.S. credit reporting system relies on complex information flows between National Credit Reporting Agencies (NCRAs), smaller credit bureaus, data furnishers, public record repositories, users of credit reports, and consumers [15]. As the core entity of this ecosystem (see Figure 1), credit bureaus gather information about consumers' credit-related activity (referred to as trade lines) from data furnishers, including banks, credit unions, credit card issuers, auto and mortgage lenders, and many other entities who can provide information related to their transactions or experiences with consumers. NCRAs also purchase public record data on individuals' bankruptcy filings, tax liens, and court judgments. Some NCRAs also keep track of debts collected by third parties on behalf of the original creditors [15]. When such data is reported to credit bureaus, it is associated with Personally Identifiable Information (PII) of consumers, such as name, current and former addresses, birth date, and social security number (SSN). Each NCRA has their own channels to collect data, which they typically do not share with other credit bureaus. The amount of data processed by credit bureaus is vast: each of the NCRAs receive information on over 1.3 billion trade lines from data furnishers and updates on over 200 million credit files on a monthly basis [47].

The key function of credit bureaus is to provide credit reports on individual consumers. These reports typically include the consumer's name, current and former addresses, SSN, birth date, phone numbers, trade lines, public record information, and inquiries for the credit report by other entities [15]. Credit bureaus also calculate a credit score for the consumers, which may differ across NCRAs. Credit bureaus then sell these reports and scores to businesses who use them to assess the creditworthiness of their customers; primarily creditors and lenders, but also landlords, insurance companies, employers, debt collectors, utility services, and government agencies [44].

In the United States, the Fair Credit Reporting Act (FCRA) regulates the activities of credit bureaus. It details obligations for NCRAs, data furnishers and credit report users, and grants consumers the right to obtain a free copy from each NCRA annually and dispute errors on their credit files. In practice, however, errors on credit files are common [30], credit bureaus and data furnishers do not conduct thorough investigations into consumers' dispute requests [55], and the way NCRAs use consumer information and advertise their services has raised controversy [17]. Further legislation aims at combating identity theft. For instance, the Fair and Accurate Credit Transactions Act (FACTA) requires debit/credit card issuers to validate customer's address changes [8] and enables consumers to place fraud alerts at NCRAs [29].

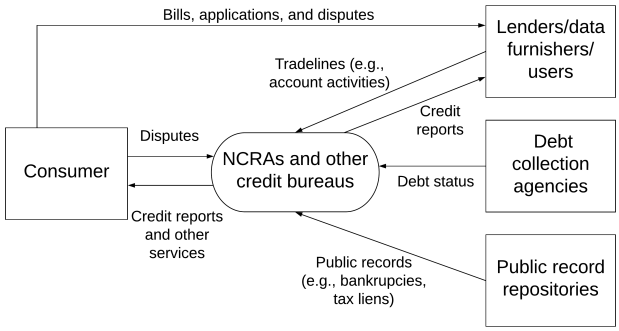


Figure 1: A simplified diagram of information flows around credit bureaus.

2.2 Available Protective Actions

Consumers have options for protecting and limiting access to their credit data, such as credit freezes, fraud alerts, checking credit reports, and using a credit monitoring service.

Credit freezes block inquiries for one's credit report, thus preventing new accounts or loans that require credit checks to be opened under the consumer's name. Unfreezing credit requires contacting the respective NCRA with a PIN to lift the freeze. However, a credit freeze is specific to a credit bureau [9], so effective protection requires placing credit freezes with each of the three NCRAs. Freezing or unfreezing one's credit typically costs $5-10 for each action with each NCRA, although some state laws prohibit those fees. A credit freeze only limits access to the credit report and thus does not protect against other types of identity theft that do not require credit checks (e.g., tax fraud).

Compared to credit freezes, fraud alerts are free but less effective. Creditors can still conduct credit checks on consumers' credit reports, but reports including fraud alerts signal that the consumer is at risk of credit fraud. Under this circumstance, creditors are expected to perform expanded identity verifications [26], but sometimes they may ignore such alerts and take no actions [61].

Consumers can further request their credit reports from the three NCRAs for inspection for free on a yearly basis, with additional costs for more frequent requests. Credit monitoring services and identity protection services are offered by NCRAs and other companies (e.g., LifeLock) as paid subscriptions. Credit monitoring alerts consumers about suspicious activity on credit reports only [82]. Identity protection services monitor more extensive information, but do

not capture tax or government benefits fraud [82]. Identity theft victims may consider identity recovery services and insurance to remediate or compensate harm, although the quality of these services seems to vary [11]. The the Federal Trade Commission (FTC) offers free online resources for victims of identity theft to help with recovery [84].

After the Equifax breach, in addition to the suggestions above, the FTC further recommended monitoring current accounts for fraudulent activity, using Equifax's dedicated website[1] to check whether one's information has been exposed, making use of a free year of credit monitoring offered by Equifax [83], as well as filing taxes early to prevent identity thieves from claiming a tax refund under one's name.

Given the range of protective measures and their respective caveats, it is not clear to what extent consumers are aware of these offerings, as well as their strengths and limitations. Furthermore, the complexity of these offerings is exacerbated by usability and reliability issues. For example, Equifax's official site for the data breach provided inconsistent results when consumers checked if they were affected [37]. Equifax even promoted the wrong web address in a tweet, sending consumers to a fake website instead [12]. In addition, Equifax tried to bundle the free credit monitoring they offered with a forced arbitration clause, so that consumers would have waived their right to sue Equifax in class-action lawsuits [57]. Other consumers were not able to place credit freezes at any of the three NCRAs, possibly due to a large volume of requests after the Equifax breach became public [22].

2.3 Data Breaches and Identity Theft

Data breaches have become increasingly prominent: 64% of the U.S. public having been affected by a major data breach [60]. Before Equifax, companies like Yahoo, Friend Finder, and eBay have suffered larger-scale data breaches [6]. Unlike these previous cases however, consumers have no choice to opt out of data collection by NCRAs.

Starting with California in 2003, most states in the U.S. have passed data breach notification laws, requiring companies to disclose data breaches immediately in a timely manner if the breach compromised consumer information. Romanosky et al. [71] found that the adoption of these laws reduced identity theft caused by data breaches about 6% on average. More recent amendments make further requirements about the compensation that affected companies should offer to consumers, such as providing free credit monitoring services if the breach involves SSN [77]. These compensations have shown to be effective in restoring customer sentiments [49].

However, there is a troubling gap between consumers' concerns and protective behaviors after data breaches. A report in 2014 [63] revealed that, following a data breach, consumers had a 21% increase of concern about being identity theft victims, but 32% of them reported that they ignored the notification and did nothing. This issue also surfaced after the Equifax data breach, when there was a less than 1% increase of newly initiated credit freezes at TransUnion [18]. These statistics imply the possibility of a pattern similar to the "privacy paradox" [48]: people claim they are concerned about their exposed data, but may not take protective actions. In our study, we investigate underlying reasons for this paradoxical behavior in the context of the Equifax data breach.

[1] `equifaxsecurity2017.com`

3. RELATED WORK

We discuss related work on security and privacy mental models, prior literature on risk perception, and user behaviors in reaction to security advice.

3.1 Security and Privacy Mental Models

Mental models are the representations of how objects or systems function in people's minds [19]. They have been studied to understand human cognition [85], reasoning [46], and decision-making [52]. Because mental models can be incomplete, imprecisely stated, with obscure or impugnable facts [35], sometimes they are also referred to as "folk models" [23, 91, 88]. In the context of human-computer interaction, studying mental models can provide insights on people's knowledge and understanding of a specific domain [41, 36], as well as help explain and predict people's interactions with complex interfaces or systems [59].

Mental models have been studied in usable privacy and security research to provide insights into users' understanding and behaviors [88, 10, 42, 65, 51, 10, 92, 91, 43, 14]. For example, Wash [88] investigated folk models of security threats and found that gaps in these models prevented users from taking actions against botnets. Bravo-Lillo et al. [10] examined mental models of security warnings and suggested that different interpretations of cues led users to diagnose underlying risks and respond differently. Zeng et al. [92] studied mental models of smart homes, revealing that end users had very limited technical understanding and concerns of potential security issues, which helped explain a lack of sophisticated mitigation strategies. Yao et al. [91] found that users had incomplete or inaccurate mental models of how online behavioral advertising (OBA) works, highlighting the importance of user education. Among these mental models studied, there are substantial discrepancies between experts and non-experts [43]. Understanding mental models of end-users hence can provide rich insights for effective communication regarding privacy and security risks [14].

Yet, little is known about consumers' mental models of credit bureaus. Studying and understanding mental models in this context can provide insights on consumers' reasoning, decision-making and behavior related to the Equifax data breach in particular, and credit bureaus and data breaches in general.

3.2 Risk Perception

Mental models have been used to study risk perception, i.e., the perceived chance that an individual will experience the effect of danger [76]. Contrary to technical or objective risk, risk perception is a person's subjective assessment of the probability that a specific event happens and how concerned they feel about its consequences [67]. Early paradigms like the psychometric model [32] interpret risk perception as a process of calculating risks versus benefits. Later theories (e.g., Cultural Theory [24, 21]) place greater emphasis on contextual factors, such as attitudes to the perceived risk and the sensitivity to general risks.

Risk perception can greatly impact individual privacy and security decisions and trigger protective actions [42]. Fagan and Khan [25], using a rational decision model, re-

vealed large differences of risk perception between users who followed common security practices (e.g., update software, use password manager) and those who did not. Altering risk perception was found to be effective in motivating end-users to make better decisions, as demonstrated by Harbach et al.'s study in which end-users behaved more privacy-consciously during the installation of Android applications after seeing personalized examples of personal information use [39].

3.3 Factors in Security and Privacy Behaviors

Risk perception is not the sole determinant of the complexity of privacy and security behaviors. Understanding risks does not automatically make users aware of appropriate countermeasures. Shay et al. [75] find that although users were aware of different account hijacking threats (e.g., malware, phishing, data breaches), most of their countermeasures focused on password management only. Differences in mental models and awareness between security experts and non-experts, are echoed in behavior: experts were found to use two-factor authentication and password managers more frequently, whereas non-experts prefer actions that demand less technical knowledge, such as using anti-virus software, changing passwords frequently, and only visiting known websites [43].

A variety of factors have been identified that prevent people from translating risk perception into protective behavior. Forget et al. [34] note the importance of awareness of technical expertise, as misalignment between estimated and actual expertise can result in insufficient security measures. Acquisti et al. [1] identified main categories that affect privacy and security choices as incomplete or asymmetric information flows; bounded rationality (the general tendency to simplify the decision-making process); and different kinds of cognitive and behavioral biases (e.g., framing effects, optimism bias, loss aversion, and status quo bias). Privacy preferences and behavior are further affected by uncertainty, context, and framing [2]. These factors have been validated in empirical studies [43, 72, 13, 3, 4, 2, 1, 90]. For instance, the belief that information is only secure within the person's own memory (e.g., "no one can hack my mind") explains why non-experts preferred memorizing the passwords themselves, and were skeptical about using expert-advocated password managers [43]. Sawaya et al. [72] showed a similar situation where self-confidence in computer security knowledge had a much greater impact on user behaviors than actual knowledge. Camp [13] pointed out that people tend to underestimate security risks when they have not experienced negative consequences from past risky behaviors.

In addition to individual factors, the source of security advice influences privacy and security behaviors [68, 69, 64, 81]. A representative survey conducted by Redmiles et al. [68] reported that users with lower Internet skill levels and socioeconomic status were less likely to get security advice from readily available sources, hence making themselves more vulnerable to security risks. Another study on security source selection [69] showed that advice from sources with a higher level of perceived trustworthiness was more likely to be taken, whereas sources that included too much marketing content or showed threats to privacy were less favored. Furthermore, Rader and Wash [64], by examining computer security advice from three different sources, discovered that each source uniquely focused on a single aspect of computer security, and it was unlikely that users could get a comprehensive picture of computer security from a single source.

We expand on prior work, by studying the underlying reasons for the suggested gap between consumers' concerns and behaviors following the Equifax data breach.

4. STUDY DESIGN

In our study, we investigated (1) consumers' mental models of how credit bureaus operate, (2) what consumers perceive as risks of the Equifax data breach, and (3) what protective actions consumers took or did not take in reaction to the perceived risks, and the reasons behind their decisions. To understand these questions, we conducted semi-structured interviews with 24 participants in January and February 2018. All interviews were audio-recorded and lasted 40 minutes on average, ranging from 20 minutes to 61 minutes. Each participant was compensated with $10. The study was determined to be exempt by our institution's IRB.

4.1 Interview Procedure

We developed and refined our script for the semi-structured interviews through multiple pilot interviews. The final interview script is included in Appendix A.

In the interviews, we started by asking participants how they manage their personal finances, leading into a discussion about their experiences with and understandings of credit bureaus. Next, we asked about their awareness of Equifax and the 2017 data breach, before providing a basic description for those who had not heard of it. We probed participants' risk perception by asking what they saw as consequences of the breach, reactions when hearing about the breach, and feelings about their data at Equifax. Then we asked whether participants have taken protective actions, and asked about their experiences and interpretations of an action's outcomes. Finally, we asked participants to recall previous experiences with data security issues (e.g., being affected by data breaches) and identity theft (e.g., someone applying for loans under their names). We wrapped up the interview by debriefing participants about the real purpose of the study (we used mild deception in recruitment to mitigate self-selection bias, see Sec. 4.2), and gave them time to ask clarification questions.

At the end of the session, participants were asked to complete two questionnaires that measured their financial decision-making ability [58] and self-determined financial well-being [16]. We collected such financial-related information after the interview to minimize potential priming. For instance, participants might otherwise think the study is about one's financial management and overstate how often they check credit reports. Conversely, the interview questions should have little impact on participants' responses to the exit survey, as they did not touch specifically on the same topics.

4.2 Recruitment

We recruited participants via online platforms (e.g., Reddit, Craigslist, and Facebook) and emails to a university research pool and campus mailing lists. We recruited for a "study on personal finance and credit bureaus" purposefully not mentioning Equifax or identity theft to avoid priming participants and limit self-selection bias. Prospective partic-

ID	Gender	Age	Education	Income	NFEC (0-8)	CFPB (0-100)
P1	F	60-69	Bachelor's	$125-150k	8	88
P2	M	30-39	Master's	$25-50k	6	61
P3	M	60-69	Bachelor's	<$25k	5	35
P4	M	18-29	Some college	$125-150k	7	73
P5	F	50-59	Master's	<$25k	3	41
P6	M	50-59	Bachelor's	$50-75k	6	45
P7	F	18-29	Bachelor's	$25-50k	4	50
P8	F	50-59	Some college	<$25k	6	47
P9	M	60-69	Bachelor's	<$25k	8	48
P10	F	18-29	Some college	$150k+	7	81
P11	F	18-29	Bachelor's	N.A	8	54
P12	M	40-49	Master's	$50-75k	7	65
P13	F	30-39	Professional degree	$50-75k	5	58
P14	F	18-29	Some college	<25k	5	56
P15	M	40-49	Bachelor's	<25k	8	49
P16	M	50-59	Master's	$75-100k	7	57
P17	F	30-39	Master's	$150k+	6	75
P18	M	30-39	Bachelor's	$25-50k	6	57
P19	F	50-59	Master's	$100-125k	7	56
P20	F	18-29	Master's	$50-75k	7	64
P21	M	50-59	Some college	$125-150k	8	82
P22	M	18-29	Bachelor's	$25-50k	6	52
P23	F	40-49	Master's	$75-100k	8	60
P24	F	40-49	Associate's	$50-75k	7	56

Table 1: Demographics of participants, and scores of NFEC financial decision [58] and CFPB financial well-being scales [16].

ipants provided basic demographic information in an online screening survey (see Appendix B). We only recruited U.S. citizens and permanent residents who had lived in the U.S. for more than five years, as recent immigrants might not be familiar with the U.S. credit reporting system or may not be included in credit bureaus' databases, yet. We deliberately selected a diverse sample of 24 participants in terms of age, gender, education, occupation, and income, as prior literature suggests demographic factors can influence people's financial experiences and responsibilities [54, 20].

4.3 Qualitative Data Analysis

With permission of the participants, we audio recorded and then transcribed all interviews. We then conducted thematic analysis [7], a common approach used for qualitative studies in human-computer interaction [50] and usable privacy and security [34, 80, 91]. The initial version of the codebook was developed by two of the authors, who coded a subset of interviews independently and grouped them into initial themes. Through multiple rounds of collaborative refinement, we achieved good inter-coder reliability (Cohen's κ=.79) [33]. The final version of the codebook included 14 overarching themes (e.g., "understanding of credit bureaus," "attitudes toward the breach," and "actions suggested by participants") and a total of 53 unique codes (see Appendix C). One researcher then coded the remaining interviews and recoded previous ones using the final version.

5. RESULTS

We first describe our sample population and then present our results focusing on three areas: mental models of credit bureaus, risk perceptions of the Equifax breach, and protective actions.

5.1 Sample Population

Table 1 summarizes the demographics of our interview participants. Our sample was diverse in terms of age, gender, education, occupation and income. We interviewed 11 male and 13 female participants. Their ages ranged from 21 to 68, with a median age of 44 years. Five (5) participants had no college experience, 10 had a Bachelor's or Associate's degree, and 9 had a graduate degree (e.g., Master's or Professional degree). Eight (8) participants worked in a university setting as students or staff, and the rest had various occupations (e.g., engineering or IT professionals, medical, business, social work, and retired). P16 was the only participant with a cybersecurity background. Our participants' annual household income ranged from less than $25,000 to more than $150,000, with the median income in the range of $50,000 to $74,999. The NFEC financial decision test [58] score ranged from 3 to 8 with a median score of 7 (out of 8); 19 of our 24 participants got a score of 6 or higher, indicating they are financially literate enough to "make entry level financial decisions" [58]. The CFPB financial well-being score [16] ranged from 35 to 88 with median score of 56.5 (out of 100), which suggests average financial well-being in our sample [16].

5.2 Mental Models of Credit Bureaus

Among the 24 participants, 19 of them were aware of the big three credit bureaus, 17 of them correctly interpreted their function as assigning credit scores to individual consumers, yet none of them could fully describe the types of information collected by credit bureaus and corresponding information exchange entities, leading their mental models to be either incomplete or inaccurate.

5.2.1 General awareness of the big three bureaus

While most participants (19) knew that there are three big credit bureaus in the United States, only 7 participants could list the specific names of all three. Four (4) participants mentioned that other smaller-scale credit bureaus also exist, e.g., *"I wouldn't be surprised if there are other smaller companies that track and monitor credit scores and stuff." (P11)*, but none of our participants were able to give specific names. A few (5) participants had difficulty mapping the names they've heard of with the concept of credit bureaus. P15 said: *"I don't know if the credit bureau is separate, or if Equifax, Experian, et al., are considered credit bureaus."* P3 considered Credit Karma, a company that offers free credit monitoring, as a credit bureau, citing his experience of checking credit scores using Credit Karma: *"It is on the same level as those three major ones [...] With Credit Karma, since they're trying to get into the market, I think, you can go to them any day and night, and they're not charging. But they have that same information."*

5.2.2 Purpose of credit bureaus

Seventeen (17) participants described credit bureaus as companies that assign credit scores to individual consumers. Most of them (14) went on to say these scores represent one's creditworthiness and hence help lenders, insurance companies and others make decisions. In contrast, a third of participants gave inaccurate descriptions of credit bureaus. P11 viewed credit bureaus as government-related: *"I think that is basically government agency that tracks and monitors each person's history, financial history."* Some confused credit bureaus with other organizations such as credit unions, debt collectors, and loan companies. P23, for instance, confused credit bureaus with credit rating agencies, who rate creditworthiness of companies and governments rather than individual consumers: *"I guess they need to support the rating [...] and maybe the credibility of that organization. Maybe*

any complaint from the customer. How they use their funding and if it's a bank, how they use the customer's money." P4 referred to credit bureaus as loan companies: *"They loan out money to their credit card that they expect you to pay back. Then if you don't pay back, then they just charge you more interest."*

5.2.3 Incomplete understanding of collected data

Regarding the types of information collected by credit bureaus, PII (e.g., names, addresses, SSN) and financial-related information (e.g., number of credit cards and loans, credit limits, late payments) were noted most frequently, even for participants who did not conceptualize the purpose of credit bureaus correctly. Half of the participants mentioned the collection of employment history, public records (e.g., tax lien and bankruptcy), and inquiries made by creditors in recent years. About one fourth of participants (7) stated that the information collected by credit bureaus is "a lot," "a variety of different things," or "almost everything," yet no participant covered all types of data collected by NCRAs. Participants' knowledge was closely tied with their personal experience with credit bureaus. Those who checked their credit reports recently and more frequently were able to recall more details, but still showed uncertainty sometimes: *"Well I think they use past accounts and maybe employment history. I know they use length of credit. But like I said, I don't know, random guessing."* (P24).

Some participants thought credit bureaus collected certain data, which credit bureaus do not actually collect. For instance, P9 thought credit bureaus checked in with a consumer's relatives and kept tabs on social media profiles such as Facebook: *"Facebook I think would just show things like their hobbies and [...] travel, like to go to Europe or Las Vegas [...] it would give you an idea of their lifestyle, and if they're throwing money around."*

5.2.4 Information providers and customers

Many (19) participants noted that financial institutions are the primary information providers for credit bureaus. *"I guess people who provide information are like banks, loan companies, loan providers, debt collectors and just people who you've rented with before and haven't paid back or stores or credit card companies"* (P13). Some participants mentioned auto dealers, governments, and utility companies as information providers, but these were brought up much less frequently. As for customers of credit bureaus, more participants (19) mentioned creditors and lenders than other businesses (e.g., car dealerships and landlords). Some participants noted that information providers of credit bureaus are simultaneously their customers, and there exists collaboration between these institutions. According to P16: *"What I also imagine is that they also send some of that information back to banks and lenders, it's a two-way street I assume, and there's probably data sharing agreements between the two of them."*

5.2.5 Offerings of credit bureaus

Many (14) participants were aware of their right to obtain a free credit report annually. Only a few (4) mentioned other products and services offered by credit bureaus that are associated with a fee, such as a credit monitoring service. A substantial portion of participants (15) noted that although they knew they could check credit scores directly at credit bureaus, they preferred to check their scores through other means (such as banks or third-party financial management tools like Credit Karma) due to low cost, convenience and frequency of updates. A prominent issue is that participants rarely knew the difference between FICO score and the scores provided by NCRAs, which are calculated using different models. P22 asked: *"As far as I know, I'm not sure how, I guess, the credit bureau interacts with the FICO credit scores, or if they create them?"*

Notably, low income participants generally knew these services were offered, but chose not to take advantage of them, in some cases refusing to interact with credit bureaus altogether. P5 and P15 both said they had no interest in checking their credit reports. According to P15: *"I can find out my credit score [...] there's a website where you can, but of course I have been reluctant to do that because, (a) I know my credit's terrible, (b) I don't want to give them any information."* Participants with higher income who did not use these offerings cited how they did not see the need to apply for credit cards, borrow money, or make big purchases.

5.2.6 Negative perceptions of credit bureaus

Almost half of our participants (10) expressed a moderately or strongly negative sentiment towards credit bureaus and/or the whole credit reporting system. In some instances, negative perceptions stemmed from doubts on whether the credit reporting system was fair to consumers. P19 said: *"I don't like the idea that things like auto insurance and getting an apartment [...] people come up with cash upfront and they still get denied because of a credit report [...] it does make sense that there is something like this, but not the way it's running right now."* P14 described how credit bureaus increased inequality by worsening the financial well-being of people who were less affluent: *"It's really like a bad cycle. If you don't have enough money and then you need a loan, and then you can't get a loan or your interest rate is really high and you can't afford to pay it."* Some (5) participants explicitly stated that credit bureaus and related institutions such as banks took advantage of individual consumers. P24 said that credit bureaus work to serve the interest of lenders, with little concerns about individual consumers: *"For the interest of who? Those in power to make these laws ... I'm assuming they probably all have lobbyists and things that could potentially benefit collaborators of credit bureaus, like lenders, businesses and car companies."*

Other negative perceptions originated from personal experiences with credit bureaus. P1 said that her husband was once denied a credit card, because credit bureaus provided the credit file of another person with the same name to the credit card company. P5 went through a long process of disputing erroneous credit card charges, during which credit bureaus offered little support, leading her to lose faith in the system: *"[The dispute process] It's probably all automated and they only take what people give them. I've been on there for things that I should not have been, but I feel powerless to try to get that stuff off. I just give up. I don't care. That's why I say I don't want to look [up my credit report]. Because how much stress and time that would take?"*

Moreover, some (5) participants expressed confusion and concern over the data collection and aggregation process between credit bureaus and their information providers and

customers. For instance, P3 expressed his frustration when he found out information about transactions between him and other businesses will inevitably fall into the hands of credit bureaus, a process he defined as "breach of confidentiality": *"When it comes to credit bureaus, I don't think there is any such thing as confidentiality [...] Whatever I'm talking to these people [banks], whatever they do, that should be strictly between them and I. Okay? But somehow, in my mind, the credit bureau ends up with this information."*

5.3 Perceived High Risk of the Equifax Breach

More than half of our participants had heard of the Equifax data breach before the interview. They conceptualized identity theft as the primary risk of the breach and described different ways that it could happen. Several (3) participants also noted privacy invasion as a secondary risk. Nevertheless, participants seldom associated these risks with themselves, implying the existence of optimism bias.

5.3.1 Aware but vague memory of the event

Participants showed a high awareness of the occurrence of the 2017 Equifax breach. A majority of participants (20) knew a data breach happened to one of the big three bureaus. 14 of them knew the breach was at Equifax, and the rest either did not remember the name, or attributed the breach to Experian.

Similar to our findings on the perceived types of data collected by credit bureaus, participants generally had a vague idea that the company was hacked, leading to the disclosure of "a lot of" information, but many participants stated that they could not remember a lot of details. P2 said: *"I don't know the specifics, if it was a hacker attack or something like that, but I know that a lot of information got out and millions of people were affected."* As for types of information that were exposed, PII including name, address, date of birth, and SSN, was mentioned most frequently, followed by bank account numbers and credit card numbers. 6 participants, who all included credit card transactions and loan history in their mental models of credit bureaus' data collection, also erroneously assumed these types of information were exposed in the breach whereas in reality they were not.

5.3.2 Identity theft as the primary risk

Most participants (19) mentioned risk of identity theft as a direct consequence of the data breach. Some (10) participants followed up with examples of how identity theft could happen. *"The consequences? Probably a lot of identity theft. It could make it very easy if somebody wants to steal somebody's identity. They could get hold of those big three or four, the name, SSN, and birth date and could just open up a bunch of accounts under their name, and they'd be none the wiser"* (P2). However, most of these examples focused on the opening of new accounts and fraudulent charges on existing accounts; only 2 participants brought up misuses of stolen personal information that did not require credit checks, such as tax fraud. P12 further mentioned that this breach prompted him to consider filing his tax return earlier this year: *"It could lead to some fraud around tax time. I heard the other day where people are... or criminals take other people's tax returns. I'm going to file my tax returns as soon as I can."*

Participants' knowledge of what data was exposed influenced their perception of the risk posed by identity theft. The loss of SSN triggered more identity theft concerns compared to other types of PII (e.g., names, addresses and dates of birth) and financial information (e.g., credit history and credit card numbers). P13 differentiated the sensitivity of exposed information based on how publicly accessible it was: *"You can find someone's date of birth and name online, but the social security number should be harder to find."* P19 was concerned due to how SSN's are hard to replace: *"You can't get a new Social Security Number, the government is not very accommodating about that and all these other things. I would prefer not to think about it because you'll just be screwed."* Both P13 and P19 mentioned that it is the combination of different kinds of data that scared them the most. As P19 said, *"If someone were to steal your identity [...] you would just be helpless. It's not like sometimes someone will take a credit card out in your name or somehow try and use your bank, and you have some recourse, but if they've got everything I have no idea what you would do."*

5.3.3 Privacy invasion as the secondary risk

In addition to identity theft, 3 participants stated that the exposure of such sensitive data is an invasion of privacy. Although P5 did not use the word 'privacy' explicitly, she described her panic when she thought about how much the hackers could know about her: *"The hackers, they would find out my personal information, which really scares me. I don't want people to know where I live. I don't want people to know whatever information they have."* P16, who did not explicitly state his own privacy concern, noted the possibility of knowing one's personal life in detail based on the exposed data: *"As they aggregate that data they can get more and more information about you. For example if there's detailed credit card information, which God I hope not, they would know your shopping habits, they might know where you live, what kinds of cars you drive."* P22 said he would value his financial information as privacy, but did not value it as highly as the loss of his SSN, due to the latter's repercussions for identity theft: *"I guess I would value my Social Security number, number one, because I don't want my identity stolen. I also value my privacy, but I feel like I haven't gotten to a point yet where I've made lots of these kinds of credit-based purchases, so not yet at a point where that's my number one."*

5.3.4 Change of trust

Based on the perceived risks, 9 participants noted that this breach eroded their trust in Equifax's ability to ensure the security of consumers' data. P14 said that consumers had no choice but to trust Equifax because: *"They're gonna get your information whether you wanted them to or not."* P12 claimed his trust in Equifax decreased to the point that he did not accept the free credit monitoring service offered by Equifax: *"Well, you didn't handle the other information, why should I trust you to monitor anymore information?"* Interestingly, a counterexample is provided by P24, who said she would trust Equifax more because Equifax would now have better security practices: *"I'd probably go back to them just because they're probably going to be a little bit more cautious than the one that didn't get hit."*

5.3.5 *Underestimated likelihood of being affected*

While almost all participants demonstrated an understanding of the risks of the breach, the majority (17) did not assume they would be personally affected, exhibiting optimism bias [1]. We identified multiple reasons for the underestimation of personal risk. Four (4) participants mentioned how they checked the Equifax website to see if they were affected and received the message "your personal information was not impacted by this incident." Another reason is the notion of 'I have nothing to lose,' especially for low income participants. P5 said: *"I don't have any credit. I have a bad record so I wouldn't do that [check if were affected]. Nobody can hurt me, it's already at the lowest place."* The third reason is the absence of signals indicating negative repercussions, such as a lack of notifications to individual consumers from Equifax and lack of suspicious account activities since the breach occurred. P7 said: *"They [Equifax] were like there was a breach and if you were directly affected we will let you know. [But then you never received?] No, so I was fine."* The fourth reason is the presumption of not being included in Equifax's database, or having limited information in the database. For instance, P6 asserted he could not be affected because he had never held any credit cards so was not included in credit bureaus' databases. P8, who held a credit card but never checked her credit reports, believed her information shared with credit bureaus was not as extensive as someone who checks their credit reports or interacts with credit bureaus in other direct ways.

Even though some participants thought they might be affected by the breach, none claimed it in an assertive way. Among the 5 participants who received the "Your personal information might have been impacted by this incident" statement from Equifax's site, most were doubtful about the meaning of "might." P13 interpreted it as a public relation strategy which did not necessarily reflect the truth, causing little concern to her: *"If they say no and then you get affected, you might be like: you said I wasn't gonna be affected so I didn't worry and I wasn't monitoring, you know? But if they say yes, then of course you're gonna freak out and start calling them, asking them for advice or services, whatever. But if they say maybe, that's like a safe, middle ground for a company to say."* Other participants who did not check the website but felt they might be affected developed this notion based on the sense that *"[if] these many people were affected, it's likely that I was affected"* (P2).

5.4 Negligence of Protective Actions

Figure 2 lists the frequency of protective actions taken, based on the FTC's suggestions for the Equifax data breach and identity theft in general [83]. More than half of our participants (14) did not actively take any protective measures after the Equifax breach, despite the perceived high risk. Participants were either unaware of available tools, or intentionally avoided using them for various reasons.

5.4.1 *Insufficient knowledge*

The high portion of participants who were unaware of available protective measures suggests insufficient knowledge as a primary reason for inaction. Only 3 participants correctly described fraud alerts, and all of them learned it from being affected by previous data breaches and being offered the service as compensation. The remaining participants either said they did not know what fraud alerts were, or associated

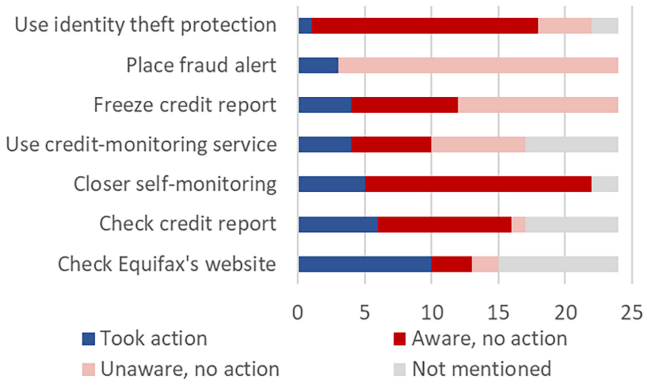


Figure 2: Status of suggested actions taken or not taken by participants.

fraud alerts with alerts sent from banks and credit card companies when fraudulent activities occur. Similarly, credit freezes were incorrectly interpreted as freezing credit cards by half of our participants. Participants generally considered the measures offered by banks and credit card companies as effective and useful in preventing identity theft. However, their unawareness of the nature of fraud alert and credit freeze measures provided by credit bureaus hampers them in utilizing these actions to protect their credit information.

5.4.2 *Costs inhibit action taking*

Cost appears to be a significant issue in determining the likelihood of whether an action was adopted. Actions with no cost were more favored: checking Equifax's website was the one taken by most participants (10), followed by checking credit reports either through the annual credit report site or third-party services for free (6), and a closer self-monitoring of existing accounts (5). For the remaining four options in Figure 2, most actions were initiated prior to the Equifax breach when participants had been affected by previous, separate data breaches and received free services as compensation. In particular, 7 participants expressed their doubts about the effectiveness of identity theft protection in relation to the associated cost. P22 said: *"It feels like you're giving them money for nothing. Also I don't know if I believe them because they can't own all of my data, so how are they actually protecting me?"*

For the 4 participants who had initiated a credit freeze, only P19 paid to have her report frozen at all three NCRAs after the breach. P16 froze his credit at all three NCRAs for free since he had been a documented victim of a previous data breach. P12 placed a credit freeze only at Equifax, which was offered for free. P20 placed only a free credit freeze at ChexSystems, a smaller-scale credit bureau. Almost one third of participants (7) expressed that freezing and unfreezing credit reports should be free, not only for Equifax but also for other credit bureaus.

Some (5) participants viewed identity theft protection services as a waste of money. P3 said: *"I'm poor. I'm having a hard time keeping my head above water or staying. I'm not giving them [credit bureaus] money for their profits."* P8 considered these services as an unwise investment: *"It wouldn't*

be worth paying for something like that, but if I had a lot of assets then I would pay for something like that, because I'd be more likely to lose money."

5.4.3 Optimism bias

A few (5) participants attributed their reason for inaction to the perceived low likelihood of being affected by the breach, making the assumption that whoever had access to the stolen data would target people who were more affluent and had a better credit history. P9 described himself as "a small fish in the pond": *"Why would they come after me? If they're going to go to all the bother of stealing my identity, why don't they go after somebody with some real wealth?"* It is worth noting that this stance was not limited to those with lower income. P1, who claimed to have an affluent income and high credit scores, did not consider herself a target either: *"These days people can be so tricky about applying for things in your name [...] and especially people who had good credit. I would think they would definitely target them. Someone like my son who has a high FICO score they might make that a priority."*

5.4.4 Tendency to delay actions

A fourth reason for inaction is a general retroactive way of dealing with risks. Six (6) participants stated that they have not noticed anything bad happen to them since the breach occurred, and saw this as reassurance that no protective actions were needed. When asked about why she did not do anything in response to the breach, P10 said: *"I haven't had any problems with my credit since that happened, that I heard about, so I'm not too concerned."* Three (3) participants noted that this might not be the most effective approach, but such awareness was not enough to trigger action. For instance, P9 shared his general attitudes towards risks in life: *"Right now, I don't have any problems, so I'm not really going to worry about it, and that's probably a very bad attitude, but I have enough problems in life without looking for trouble."* Similarly, P23 reflected that she might not have a very proactive approach: *"I wait until something bad happens and then I will react to it, so maybe it's not as good as a proactive approach. So far I think I'm okay with all the finance and nobody's stole my identity yet."*

5.4.5 Sources of advice for initiated actions

For participants who did initiate actions, 10 said they were motivated or reminded to take actions after receiving advice from a variety of sources. News media were brought up most often, primarily informing participants about the breach and available options rather than prompting actions. For instance, P9 reflected he heard of the breach from NBC Nightly News but did not follow their recommendations: *"I'm not sure which company it was, Equifax or which one, but I remember, it's been a while [...] consumers were supposed to take action and do something, but I didn't pay much attention to it because I didn't feel threatened."* Four (4) participants mentioned TV advertisements of LifeLock as the information source of identity theft protection, but none of them had signed up for the service.

In contrast, 4 participants who learned about available actions from sources they trusted (e.g., family members, colleagues, and experts) followed the given advice. P19 talked about how she decided to initiate credit freezes after hearing recommendations from a colleague: *"He's our tech guy, he put together an e-mail from various things that he'd seen about how to find out and what to do, and so I finally did something."* P16 followed a security expert's advice to place credit freezes at all three big NCRAs after he was involved in a previous data breach: *"There's a gentleman named Brian Krebs, who is active in the security community, and he gave a very informative article about what's involved with credit freezes, and why he chose to take that path to protect his credit. Given the way I use credit and given the way I have a very good credit rating, and given the data breaches it made a lot of sense for me to do that."* P16 also mentioned that he shared Krebs' article to his family members after knowing his son was affected by the Equifax breach.

5.4.6 False sense of security

Three (3) participants mentioned that taking actions created a *"false sense of security"* (P16) that led to them not taking other actions. P19, for example, after freezing her credit report at all three bureaus, did not continue to monitor her credit reports or accounts: *"I downloaded the reports so I had a copy of it then, but I haven't done anything since then with regards to looking at it, since I assumed that the freeze is working. I guess I am trusting the freeze, and also I just don't want the hassle of having to worry about it all the time."* P16 said how, after he was involved in a prior data breach and froze his report, he checked his reports once a year instead of increasing the frequency at which he checked his reports, which he thought he should do. He was also aware that a credit freeze cannot fully eliminate the risk of identity theft: *"I think the credit freeze can help with some of it, but again it depends on the institution that they're using... let's say for example it was a car loan. If somebody was able to misrepresent themselves as me they might be able to get the loan, and because the person didn't find out that there was a credit freeze, maybe there's an agency or maybe someone else besides the big three that is used to verify someone's credit information. That would be a concern to me."*

5.4.7 Usability issues

Usability issues did not necessarily deter participants from taking actions but still affected their experience. Two (2) participants mentioned the need to use a PIN to lift the freeze was inconvenient. P8 described how a credit freeze created a lot of hassles for her elderly parents: *"I know my father one time I was with him and he wanted to buy something, and he had to call the company, TransUnion, but then he couldn't remember his account, his numbers and it just seemed like it was a lot of trouble, and nowadays since you always have to know so many different passwords it makes it difficult to remember."* Another instance is P20, who initially tried to place credit freeze at NCRAs, but found it "costs money and delays things," so she eventually placed a credit freeze at a smaller bureau ChexSystems.

Participants also offered suggestions on making the information flows around credit bureaus more transparent. For instance, P5 was not satisfied that consumers could only partially check what data credit bureaus collect about them, with limited resources for intervention: *"I think that I can learn what they know about me, but I don't have the power and the access to find out ... well, I guess it should be equal, those reporting should be the same as those who have their name reported, but I'm skeptical."* P23 expressed confusion about different credit scores provided by different bureaus,

and argued that they should all be the same. Some participants stated that Equifax should communicate more openly about its mistake, instead of publishing a website and assigning the responsibility of checking and taking actions to consumers. P21 said: *"They should have reached out to their companies or their customers. Be very open about what exactly happened to the extent possible on a personal basis and communicate that to me personally."* P16 offered a general suggestion to the credit reporting system: *"I think the best way to regulate them is to define the boundaries around privacy and data, and then to come up with means and standards to protect that information. Then on top of that there should be means for consumers to work with those companies to have them respond to errors and misinformation, and to meet consumer needs."*

6. DISCUSSION

Our findings provide insights on the reasons why consumers' concerns and risk awareness did not translate into protective behaviors after the Equifax data breach. Next, we first discuss potential limitations of our study, before summarizing our key insights and discussing the implications of our findings for public policy, technical solutions, and educational efforts.

6.1 Limitations

Our study has certain limitations. First, as is common for qualitative research [50], our sample size cannot support quantitative conclusions about the general U.S. population. We also acknowledge that our sample exhibits a higher level of education on average. However, we believe our study provides rich qualitative insights on people's mental models of credit bureaus and hurdles in taking protective actions after data breaches. Findings like optimism bias and a reactive approach to dealing with risk are unlikely to be specific to more educated people. We recruited a demographically diverse sample to gather a wide range of issues, perceptions and perspectives. Furthermore, while we studied credit bureaus and protective behavior in the context of the 2017 Equifax breach, our findings provide insights beyond this particular data breach.

Second, we conducted our study four months after the 2017 Equifax breach was made public. This may have resulted in a dilution effect — a decrease in awareness of the breach during the time between the breach and our interviews. We chose the timing deliberately to ensure participants had sufficient time to take any protective actions they might want to take. Although most participants had vague memory of the details of the breach, they still remembered clearly that it occurred as well as what actions they did and did not take, and were able to articulate the reasons why.

Third, our study is limited by the self-reported nature of interviews. Participants may overclaim their security and privacy concerns due to social desirability bias. To mitigate this issue, we designed our interview script to avoid biasing participants about security and privacy risks, giving them opportunities to bring up details of their own attitudes and actions before prompting them about protective measures.

6.2 Key Insights

Our findings reveal that protective actions were less influenced by mental models and risk perception, but more influenced by costs, sources of advice, an optimism bias of "the rich will be targeted" and "I've got nothing to lose."

6.2.1 Awareness does not lead to action

Our participants' mental models of credit bureaus and their risk awareness were not the primary factors affecting their protective behaviors. In line with previous work [88, 10, 92, 91], we found connections between certain components of participants' mental models and their identity theft risk perception: for instance, the only 2 participants who mentioned the potential of tax fraud also specified government agencies as information providers of credit bureaus. A majority of participants showed detailed awareness of identity theft risks (regardless of the sophistication of their mental models), and yet most did not translate this awareness into action. For participants who had articulated mental models of credit bureaus but chose not to take action, their decisions seemed to be influenced by the misinterpretation of the outcome of existing tools and their own biases, rather than a lack of knowledge on how credit bureaus operate.

6.2.2 Costs as a barrier for protective action

A striking theme among our findings is how credit reports not only disadvantage people with low income, but on top of that, the fees associated with protective actions (such as freezing or unfreezing one's credit report) further enhance the barriers to take protective actions for people with low income. Identity theft protection, for instance, was perceived as an untrustworthy and unwise investment by about one fourth of the participants. For participants who had initiated a credit freeze, half of them did not place it at all NCRAs due to the costs at the other credit bureaus: hence, their credit freezes were not fully effective. Most participants who took actions in response to the breach chose economical options, such as checking the free annual credit report and keeping a close monitoring on existing accounts themselves.

6.2.3 Security advice as a trigger for action

Our findings confirm the significance of the source as an important factor in security advice adherence, similar to previous studies [69, 68], and provide additional insights about potential effects of different types of sources. Quite a few participants gained the awareness of the breach and certain protective actions from news media, but the awareness was not enough to trigger taking protective actions. Among those who took actions, many of them actually followed recommendations from sources with high perceived expertise and trustworthiness, rather than seeking information themselves. A possible explanation is that participants generally received high-level information of the incident from news media, but were more likely to resonate with the detailed, personal experiences provided by people they trusted. Our finding implies the importance for future research to examine how different characteristics of sources (e.g., social closeness, accessibility, quality, credibility, up-to-dateness) may affect the selection of sources and effectiveness of security advice.

6.2.4 Underestimating risk of being affected

The reasons for inaction are related to factors previously identified as preventing users from using security and privacy measures. For example, optimism bias — the general tendency of underestimating the possibility of being affected by negative events [74] — is a significant factor in affecting

privacy and security decisions [1]. Our study confirms this by showing that, regardless of participants' own income levels, they tended to think the 'rich' were more likely to become the target of identity theft. Our results exhibit similar patterns to Camp's findings [13], in that people tend to underestimate security risks when the negative consequences of previous risky behaviors are unnoticed, which reinforces the notion that protective action is not needed.

6.2.5 The "I've got nothing to lose" fallacy

Among participants with low income, the lack of motivation to take actions is similar to the well-known 'I've got nothing to hide' fallacy in privacy research [78] — some people believe they do not need to be concerned about privacy, as long as they have no secrets to hide. Similarly, several participants in our study did not exhibit strong motivations to take actions because they thought they had nothing to lose, given their limited income or assets. Nevertheless, this notion runs counter to a population survey conducted in 2013 [66]: people in low income households were highly likely to have negative online experiences, such as having email and social media accounts compromised. In addition, this survey pointed out that median households were most likely to be victims of identity theft rather than high income households [66], potentially due to the latter group being more capable of affording identity theft protection services.

6.3 Implications and Recommendations

Our findings have implications for public policy, as well as for technical and educational approaches for improving consumers' reaction to data breaches.

6.3.1 Public policy recommendations

Our findings demonstrate the need to revise or amend the Fair Credit Reporting Act to better protect consumers' sensitive information held by credit bureaus as well as lower the barriers for consumers to take sufficient protective actions.

Free and frequent access to credit reports. We argue that consumers should be able to obtain their detailed credit reports from the NCRAs for free at anytime. Under the FCRA, consumers are entitled to check their credit reports for free once a year at each NCRA. We found that participants preferred to check their credit status through banks and third-party financial management services, due to lower costs, greater convenience and usability, and the ability to more frequently check their credit scores. Nevertheless, many of these third-party offerings only show a credit score and a simplified version of the report, which may lead consumers to overlook important details in a full version report. In some cases, these services aggregate credit scores from different credit bureaus but do not explain it explicitly to participants, leading to confusion. Given these issues, and the impacts of identity theft and erroneous credit reports on someone's life, a free and frequent access to credit reports is needed to lower the barriers for consumers to monitor their credit reports for irregular activities.

Free credit freezes. Similarly, credit freezes — which are currently the most effective way of limiting undesired access to one's credit data — should be free under any circumstance in all states (some U.S. states already have state legislation mandating credit freezes to be free), as also suggested by Bruce Schneier in his congress testimony on the Equifax breach [73]. Currently, a freeze or unfreeze operation can cost up to $10 per NCRA depending on the state of residence, and this credit freeze has to be performed at each NRCA separately.

Stringent and preemptive oversight. The magnitude of the 2017 Equifax breach indicates a need for more stringent oversight of credit bureaus and better auditing credit bureaus' operation and data security. In the past, the FTC has charged both credit bureaus [28] and data furnishers [31] for violating rules of the FCRA. While such measures might be appropriate reactions after a breach occurred, our findings showed that participants in general held a negative sentiment towards credit bureaus on many aspects, such as inaccurate credit files, opaque data aggregation practices, and inappropriate handling of data breaches. In addition to remedial enforcement, more emphasis should be placed on preemptive oversight measures (such as detecting and preventing misconduct through audits), in order to ensure the security and accuracy of consumer credit data.

6.3.2 Technical recommendations

Accompanying public policy reform, better technical solutions should be implemented to ensure that regulatory efforts result in improved protective measures for consumers.

Enhancing usability of protection mechanisms. Our findings revealed the tools consumers use to manage their credit data have severe usability issues: participants experienced hassles when using some of the tools (e.g., forgetting whether a credit freeze had been placed), or avoided using them due to low perceived trustworthiness. Educating consumers about protective actions would not make sense unless these usability issues are addressed. We argue that credit freezes, for instance, should be offered as an integrated, user-friendly system. Similar to fraud alerts, credit freeze requests should be automatically communicated between all three NCRAs, rather than requiring consumers to work through (and pay for) the steps of freezing or unfreezing their credit with each bureau.

Enhancing transparency of information flows. Further research should focus on making credit-related information flows more transparent and on re-thinking how consumers can be integrated into these information flows. Our study shows that the opaque data collection and aggregation process of credit bureaus leads to misconception: some participants believed they were not included in credit bureaus' databases since they had no credit card, whereas in reality credit bureaus can still collect data about them from other information providers such as car dealerships and utility companies. Compounding this is how consumers cannot opt out of Equifax's services. Even though they can avoid using certain paid services, such as credit monitoring, consumers have no control over the information exchange between NCRAs or other credit bureaus and their data providers.

Nevertheless, efforts can be made to make the information flows more transparent with higher engagement from consumers. One possibility is to develop just-in-time notifications informing consumers whenever companies request access to their credit data, new data is added to their credit file, or any credit bureau creates a credit file about them.

Such a notification system could be a centralized offering, similar to the FTC's annual credit report website, or (less ideally) offered by individual credit bureaus. Once a consumer signed up for this service and their identity has been verified, these notifications could be delivered in various formats (e.g., mobile app, text message, email). These notifications should also be quick to read and easy to understand.

Such notifications could even be combined with an approval process between credit bureaus and consumers when a credit request is made by a third party, so that consumers have the agency to allow or deny those requests [73] — similar to permission requests on smartphones. Moreover, given how for some participants the current dispute process is problematic and can erode trust, dispute options could be integrated directly into this notification system. For instance, consumers could immediately raise a red flag when they notice wrong data being added to their files, thus making the dispute process function more timely and efficient. This might lead to higher quality of credit data overall, thus benefiting not only consumers, but credit bureaus and lenders as well.

6.3.3 Educational efforts

Furthermore, the implementation of regulatory and technical measures should be accompanied by the development and assessment of effective consumer education. Similar to previous findings [18], participants in our study showed a limited understanding of existing tools such as credit freeze and fraud alert, and frequently misinterpreted their outcomes.

Aiming educational resources at influencers. Efforts to educate consumers about financial literacy and identity theft protection should aim to enhance not only the understanding of key financial concepts, but also the aptitude in managing personal finances and making reasonable financial decisions [70]. While making resources more widely accessible online is important, our findings suggest that provisioning resources alone is not sufficient to reach the majority of consumers. Our participants tended to act primarily upon advice from people they trusted rather than news and online resources, and the fact that no participant mentioned the abundant free resources on identity theft protection, such as the FTC's identity theft website, illuminates the significant gap between consumers' awareness and available public resources. However, it also suggests an interesting opportunity: enlisting financially-literate or tech-savvy consumers as 'influencers' to educate their community. Rather than creating 'one-size-fits-all' educational materials and resources, help people who are already motivated and well versed in these matters better communicate ideas and recomendations to others.

7. CONCLUSION

We examined consumers' mental models of credit bureaus, risk perceptions, and reasons for taking or not taking protective action in the context of the 2017 Equifax data breach. We found that mental models varied, especially with regards to information flows and information providers of credit bureaus. We also found that identity theft and privacy invasion were perceived as the primary and secondary risks of this breach, with most participants demonstrating a good understanding of how these risks may manifest. But more importantly, we found that, overall, the accuracy or completeness of consumers' mental model, and awareness of the data breach and its risks, did little to explain consumers' inaction; instead, factors such as insufficient knowledge regarding protective actions, optimism bias, a belief that only 'rich' people would be targets, a tendency to delay actions, a false sense of security, usability issues, and associated fees played a much more prominent role.

In line with our findings, we propose directions for future research. One is to confirm and quantify our results through larger-scale surveys, examining the prevalence of our identified reasons for taking or not taking protective measures, and also formalizing the aspects of mental models we identified through structural equation modeling. Another direction is to conduct longitudinal studies to investigate whether there is any significant shift of consumers' attitudes and behaviors in reaction to a data breach over time. Furthermore, future research should analyze other types of non-self-reported data that may better represent consumers' actual behaviors, such as comments regarding the breach on social media, and numbers of credit freezes placed at each NCRA.

We outline implications and recommendations for public policy, technical and educational efforts aimed at enhancing consumer protections and empowering consumers to more effectively protect themselves after data breaches. Efforts in these areas should be pursued simultaneously in order to increase chances of success. Consumer protection regulation needs to be augmented with usable protection mechanisms and systems that make the credit system's information flows more transparent. At the same time, new systems are needed to better integrate consumers into these information flows through just-in-time notifications and integrated approval and dispute capabilities. However, on their own new usability and technology solutions are unlikely to be adopted by NCRAs due to little incentive to provide usable or free measures to consumers, unless mandated through regulatory oversight. Educational efforts, furthermore, are needed to make consumers aware of their rights and available choices and guide them to take actions.

So far, the 2017 Equifax data breach has not resulted in regulatory changes, despite efforts by consumers and policy makers. A class action lawsuit against Equifax by consumers is on-going [5]. The Consumer Financial Protection Bureau (CFPB) has received more than 20,000 complaints regarding the Equifax data breach, but the CFPB has not yet responded [79]. Recently, bills have been introduced in Congress aiming to impose stricter penalties for data breaches [87] and to make fraud alerts and credit freezes more accessible for consumers [86]. Nevertheless, Congress has not been able to translate these proposals into legislation due to conflicting interests, particularly from industry [56], resulting in counter proposals that instead would make it easier for financial institutions to evade responsibilities when a data breach occurs [53].

While we conducted our study in the context of credit bureaus and the 2017 Equifax data breach, we believe that our findings also provide indications as to why people might not act after data breaches in general. The combination of optimism bias, usability issues, and financial hurdles seems to be a powerful deterrent to protective actions, which requires further investigations in other contexts and the development of holistic approaches to address these issues together rather than focusing only on one or a subset of them.

8. REFERENCES

[1] A. Acquisti, I. Adjerid, R. Balebako, L. Brandimarte, L. F. Cranor, S. Komanduri, P. G. Leon, N. Sadeh, F. Schaub, M. Sleeper, et al. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)*, 50(3):44, 2017.

[2] A. Acquisti, L. Brandimarte, and G. Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[3] A. Acquisti, L. K. John, and G. Loewenstein. The impact of relative standards on the propensity to disclose. *Journal of Marketing Research*, 49(2):160–174, 2012.

[4] I. Adjerid, A. Acquisti, L. Brandimarte, and G. Loewenstein. Sleights of privacy: Framing, disclosures, and the limits of transparency. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, page 9. ACM, 2013.

[5] E. Ambrose. How to get in on a class-action lawsuit against equifax. `https://finance.yahoo.com/news/class-action-lawsuit-against-equifax-174234204.html`, 2018. last accessed on: 05.22.2018.

[6] T. Armerding. The 17 biggest data breaches of the 21st century. `https://www.csoonline.com/article/2130877/data-breach/the-biggest-data-breaches-of-the-21st-century.html`, 2017. last accessed on: 02.12.2018.

[7] J. Aronson. A pragmatic view of thematic analysis. *The qualitative report*, 2(1):1–3, 1995.

[8] Association of Corporate Counsel. The fair and accurate credit transactions act (FACTA). `http://www.acc.com/legalresources/quickcounsel/tfaacta.cfm`, 2011. last accessed on: 01.22.2018.

[9] A. Bahney. Credit freeze: What is it and should you do it? `http://money.cnn.com/2017/09/12/pf/what-is-a-credit-freeze/index.html`, 2017. last accessed on: 02.12.2018.

[10] C. Bravo-Lillo, L. F. Cranor, J. Downs, and S. Komanduri. Bridging the gap in computer security warnings: A mental model approach. *IEEE Security & Privacy*, 9(2):18–26, 2011.

[11] J. Bromberg, T. Alexander, S. Kerney, and V. Tarpinian. Identity theft services: Services offer some benefits but are limited in preventing fraud. Technical report, Government Accountability Office, Washington D.C., United States, 2017.

[12] D. Cameron. Equifax has been sending consumers to a fake phishing site for almost two weeks. `https://gizmodo.com/equifax-has-been-sending-consumers-to-a-fake-phishing-s-1818588764`, 2017. last accessed on: 01.22.2018.

[13] L. J. Camp. Mental models of privacy and security. *IEEE Technology and society magazine*, 28(3), 2009.

[14] L. J. Camp, F. Asgharpour, D. Liu, and I. Bloomington. Experimental evaluations of expert and non-expert computer users' mental models of security risks. *Proceedings of WEIS 2007*, 2007.

[15] Consumer Financial Protection Bureau. Key dimensions and processes in the US credit reporting system: A review of how the nation's largest credit bureaus manage consumer data. Technical report, Washington, DC: US Government Printing Office, 2012.

[16] Consumer Financial Protection Bureau. Measuring financial well-being: A guide to using the CFPB Financial Well-Being Scale. `https://www.consumerfinance.gov/data-research/research-reports/financial-well-being-scale/`, 2015. last accessed on: 02.16.2018.

[17] Consumer Financial Protection Bureau. CFPB orders TransUnion and Equifax to pay for deceiving consumers in marketing credit scores and credit products. `https://www.consumerfinance.gov/about-us/newsroom/cfpb-orders-transunion-and-equifax-pay-deceiving-consumers-marketing-credit-scores-and-credit-products/`, 2017. last accessed on: 01.22.2018.

[18] L. K. Cox. Your credit cards keep getting hacked: Only 1% use credit freezes. `https://www.creditsesame.com/blog/credit/credit-cards-hacked-only-one-percent-fight-back-with-credit-freezes/`, 2017. last accessed on: 02.12.2018.

[19] K. Craik. The nature of explanation. *Cambridge University, Cambridge UK*, 1967.

[20] B. Cude, F. Lawrence, A. Lyons, K. Metzger, E. LeJeune, L. Marks, and K. Machtmes. College students and financial literacy: What they know and what we need to learn. *Proceedings of the Eastern Family Economics and Resource Management Association*, 102(9):106–109, 2006.

[21] K. Dake. Orienting dispositions in the perception of risk: An analysis of contemporary worldviews and cultural biases. *Journal of cross-cultural psychology*, 22(1):61–82, 1991.

[22] A. E. Dastagir. Equifax data breach: I tried to freeze my credit. there were problems. `https://www.usatoday.com/story/money/2017/09/13/equifax-data-breach-tried-freeze-my-credit-there-were-problems/663014001/`, 2017. last accessed on: 01.22.2018.

[23] S. Dekker and E. Hollnagel. Human factors and folk models. *Cognition, Technology & Work*, 6(2):79–86, 2004.

[24] M. Douglas and A. Wildavsky. *Risk and culture: An essay on the selection of technological and environmental dangers*. Univ of California Press, 1983.

[25] M. Fagan and M. M. H. 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.

[26] L. Fair. Fraud alerts vs. credit freezes: FTC FAQs. `https://www.ftc.gov/news-events/blogs/business-blog/2017/09/fraud-alerts-vs-credit-freezes-ftc-faqs`, 2017. last accessed on: 01.22.2018.

[27] Federal Reserve System. Credit reports and credit scores. `https://www.federalreserve.gov/creditreports/pdf/credit_reports_scores_2.pdf`, 2017. last accessed on: 01.22.2018.

[28] Federal Trade Commission. Nation's big three consumer reporting agencies agree to pay $2.5 million to settle ftc charges of violating fair credit reporting act. https://www.ftc.gov/news-events/press-releases/2000/01/nations-big-three-consumer-reporting-agencies-agree-pay-25, 2000. last accessed on: 01.22.2018.

[29] Federal Trade Commission. Fair and accurate credit transactions act of 2003. https://www.ftc.gov/enforcement/statutes/fair-accurate-credit-transactions-act-2003, 2003. last accessed on: 01.22.2018.

[30] Federal Trade Commission. Report to congress under section 319 of the fair and accurate credit transactions act of 2003. Technical report, Washington DC: US Government Printing Office, 2012.

[31] Federal Trade Commission. Debt collector settles ftc charges it violated fair credit reporting act. https://www.ftc.gov/news-events/press-releases/2016/05/debt-collector-settles-ftc-charges-it-violated-fair-credit, 2016. last accessed on: 01.22.2018.

[32] B. Fischhoff, P. Slovic, S. Lichtenstein, S. Read, and B. Combs. How safe is safe enough? a psychometric study of attitudes towards technological risks and benefits. *Policy sciences*, 9(2):127–152, 1978.

[33] J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[34] A. Forget, S. Pearman, J. Thomas, A. Acquisti, N. Christin, L. F. Cranor, S. Egelman, M. Harbach, and R. 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.

[35] J. W. Forrester. Counterintuitive behavior of social systems. *Technological Forecasting and Social Change*, 3:1–22, 1971.

[36] D. Gentner and A. L. Stevens. *Mental models*. Psychology Press, 2014.

[37] K. B. Grant. How to protect yourself after the Equifax breach: Assume you're affected. https://www.cnbc.com/2017/09/08/how-to-protect-yourself-after-the-equifax-data-breach.html, 2017. last accessed on: 01.22.2018.

[38] M. Hadi and B. Logan. Equifax: Hackers may have the personal details of 143 million US customers. http://www.businessinsider.com/equifax-hackers-may-have-accessed-personal-details-143-million-us-customers-2017-9, 2017. last accessed on: 01.22.2018.

[39] M. Harbach, M. Hettig, S. Weber, and M. Smith. Using personal examples to improve risk communication for security & privacy decisions. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2647–2656. ACM, 2014.

[40] E. Harrell and L. Langton. *Victims of identity theft, 2014*. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, 2015.

[41] M. Helander, T. Landauer, and P. Prabhu. Mental models and user models. In *Handbook of human-computer interaction*, pages 49–63. Elsevier, 1997.

[42] A. E. Howe, I. Ray, M. Roberts, M. Urbanska, and Z. Byrne. The psychology of security for the home computer user. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 209–223. IEEE, 2012.

[43] I. Ion, R. Reeder, and S. Consolvo. "... no one can hack my mind": Comparing expert and non-expert security practices. In *SOUPS*, volume 15, pages 1–20, 2015.

[44] L. Irby. 8 people who check your credit report. https://www.thebalance.com/people-who-check-credit-report-960517, 2017. last accessed on: 02.12.2018.

[45] A. Johnson. Equifax breaks down just how bad last year's data breach was. https://www.nbcnews.com/news/us-news/equifax-breaks-down-just-how-bad-last-year-s-data-n872496, 2018. last accessed on: 05.22.2018.

[46] P. N. Johnson-Laird. Mental models and reasoning. *The nature of reasoning*, pages 169–204, 2004.

[47] A. Klein. The real problem with credit reports is the astounding number of errors. https://www.brookings.edu/research/the-real-problem-with-credit-reports-is-the-astounding-number-of-errors/, 2017. last accessed on: 01.22.2018.

[48] S. Kokolakis. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security*, 64:122–134, 2017.

[49] T. Kude, H. Hoehle, and T. A. Sykes. Big data breaches and customer compensation strategies: Personality traits and social influence as antecedents of perceived compensation. *International Journal of Operations & Production Management*, 37(1):56–74, 2017.

[50] J. Lazar, J. H. Feng, and H. Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

[51] J. Lin, S. Amini, J. I. Hong, N. Sadeh, J. Lindqvist, and J. Zhang. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 501–510. ACM, 2012.

[52] R. Lipshitz and O. Ben Shaul. Schemata and mental models in recognition-primed decision making. *Naturalistic decision making*, pages 293–303, 1997.

[53] M. Litt. Three bills in congress this week would let equifax off the hook. https://uspirg.org/news/usp/three-bills-congress-week-would-let-equifax-hook, 2018. last accessed on: 05.22.2018.

[54] A. Lusardi and O. S. Mitchell. How ordinary consumers make complex economic decisions: Financial literacy and retirement readiness. *Quarterly Journal of Finance*, 7(03):1750008, 2017.

[55] M. Mahoney. Errors and gotchas: How credit report errors and unreliable credit scores hurt consumers. Technical report, Washington, DC: Consumers Union, 2014.

[56] M. Matishak. After equifax breach, anger but no action in congress.

https://www.politico.com/story/2018/01/01/equifax-data-breach-congress-action-319631, 2018. last accessed on: 05.22.2018.

[57] A. Naini. Equifax and wells fargo reveal what's offensively wrong with forced arbitration. http://www.nydailynews.com/opinion/equifax-wells-fargo-reveal-wrong-forced-arbitration-article-1.3520644, 2017. last accessed on: 02.15.2018.

[58] National Financial Educators Council. Test your knowledge with the nfec financial foundation test. https://www.financialeducatorscouncil.org/financial-foundation-test/, 2017. last accessed on 01.22.2018.

[59] D. A. Norman. Some observations on mental models. *Mental models*, 7(112):7–14, 1983.

[60] K. Olmstead and A. Smith. Americans and cybersecurity. Technical report, The Pew Research Center, 2017.

[61] M. Osakwe. What's the difference between fraud alerts and credit freezes? https://www.huffingtonpost.com/entry/whats-the-difference-between-fraud-alerts-and-credit_us_596e4acde4b07f87578e6c7b, 2017. last accessed on: 01.22.2018.

[62] M. W. Perl. It's not always about the money: Why the state identity theft laws fail to adequately address criminal record identity theft. *J. Crim. L. & Criminology*, 94:169, 2003.

[63] Ponemon Institute. The aftermath of a data breach: Consumer sentiment. Technical report, Ponemon Institute LLC, 2014.

[64] E. Rader and R. Wash. Identifying patterns in informal sources of security information. *Journal of Cybersecurity*, 1(1):121–144, 2015.

[65] E. Rader, R. Wash, and B. Brooks. Stories as informal lessons about security. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, page 6. ACM, 2012.

[66] L. Rainie, S. Kiesler, R. Kang, and M. Madden. Online identity theft, security issues, and reputational damage. Technical report, 2013.

[67] S. Rayner and R. Cantor. How fair is safe enough? the cultural approach to societal technology choice. *Risk analysis*, 7(1):3–9, 1987.

[68] E. M. Redmiles, S. Kross, and M. 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.

[69] E. M. Redmiles, A. R. Malone, and M. 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.

[70] D. L. Remund. Financial literacy explicated: The case for a clearer definition in an increasingly complex economy. *Journal of Consumer Affairs*, 44(2):276–295, 2010.

[71] S. Romanosky, R. Telang, and A. Acquisti. Do data breach disclosure laws reduce identity theft? *Journal of Policy Analysis and Management*, 30(2):256–286, 2011.

[72] Y. Sawaya, M. Sharif, N. Christin, A. Kubota, A. Nakarai, and A. Yamada. Self-confidence trumps knowledge: A cross-cultural study of security behavior. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2202–2214. ACM, 2017.

[73] B. Schneier. Me on the equifax breach. https://www.schneier.com/blog/archives/2017/11/me_on_the_equif.html, Nov 2017. last accessed on: 02.12.2018.

[74] T. Sharot. The optimism bias. *Current biology*, 21(23):R941–R945, 2011.

[75] R. Shay, I. Ion, R. W. Reeder, and S. Consolvo. My religious aunt asked why i was trying to sell her viagra: experiences with account hijacking. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 2657–2666. ACM, 2014.

[76] J. F. Short. The social fabric at risk: toward the social transformation of risk analysis. *American sociological review*, 49(6):711–725, 1984.

[77] C. Skeath. Delaware amends data breach notification law to require credit monitoring, attorney general notification. https://www.insideprivacy.com/data-security/data-breaches/delaware-amends-data-breach-notification-law-to-require-credit-monitoring-attorney-general-notification/, 2017. last accessed on: 01.22.2018.

[78] D. J. Solove. I've got nothing to hide and other misunderstandings of privacy. *San Diego L. Rev.*, 44:745, 2007.

[79] E. Stewart. Consumers have filed thousands of complaints about the equifax data breach. the government still hasn't acted. https://www.vox.com/policy-and-politics/2018/4/30/17277172/equifax-data-breach-cfpb-elizabeth-warren-mick-mulvaney, 2018. last accessed on: 05.22.2018.

[80] E. Stobert and R. Biddle. The password life cycle: user behaviour in managing passwords. In *Proc. SOUPS*, 2014.

[81] P. Szewczyk and S. Furnell. Assessing the online security awareness of australian internet users. 2009.

[82] The Federal Trade Commission. Consumer reports: What information furnishers need to know. https://www.ftc.gov/system/files/documents/plain-language/bus33-consumer-reports-what-information-furnishers-need-know_0_0.pdf, 2016. last accessed on 01.22.2018.

[83] The Federal Trade Commission. The equifax data breach: What to do. https://www.consumer.ftc.gov/blog/2017/09/equifax-data-breach-what-do, 2017. last accessed on 02.12.2018.

[84] The Federal Trade Commission. Identity theft recovery steps. https://www.identitytheft.gov/, 2018. last accessed on: 02.15.2018.

[85] S. Vosniadou and W. F. Brewer. Mental models of the earth: A study of conceptual change in childhood. *Cognitive psychology*, 24(4):535–585, 1992.

[86] Warren, Elizabeth. S.1816 - freedom from equifax exploitation act. Technical report, United States Congress, 2018.

[87] Warren, Elizabeth. S.2289 - data breach prevention and compensation act of 2018. Technical report, United States Congress, 2018.

[88] R. Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, page 11. ACM, 2010.

[89] M. D. White and C. Fisher. Assessing our knowledge of identity theft: The challenges to effective prevention and control efforts. *Criminal Justice Policy Review*, 19(1):3–24, 2008.

[90] H. Xia and J. C. Brustoloni. Hardening web browsers against man-in-the-middle and eavesdropping attacks. In *Proceedings of the 14th international conference on World Wide Web*, pages 489–498. ACM, 2005.

[91] Y. Yao, D. L. Re, and Y. Wang. Folk models of online behavioral advertising. In *CSCW*, pages 1957–1969, 2017.

[92] E. Zeng, S. Mare, and F. Roesner. End user security & privacy concerns with smart homes. In *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

APPENDIX

A. INTERVIEW GUIDE

1. Could you tell me how you manage your personal finance, such as income and credit cards? Has it changed over time?
   (a) If yes, could you tell me any particular points that the change occurred?
   (b) If no, could you explain why?
2. What's the first thing that comes into your mind when you hear the term "credit bureau"?
   (a) From your point of view, what do credit bureaus do?
   (b) You just said credit bureaus do... How do they do this? Could you draw or sketch on the paper to make it clear? (A few prompts listed as below if necessary)
      i. What information do they collect?
      ii. What parties do they share information with?
      iii. What do they know about you?
      iv. What information you can get from them?
      v. What is their purpose?
   (c) Could you name some credit bureaus?
3. Could you tell me your personal experience with credit bureaus?
   (a) Have you ever interacted with credit bureaus directly?
      i. If yes, when was the last time, and how was the experience?
      ii. If no, could you explain why?
   (b) What do you know about your credit history and credit scores?
   (c) Have you ever checked your credit report?
      i. If yes, when was the last time? What prompted you to check it? How did you do it? With which credit bureau? Only one or multiple?
      ii. If no, why not?
   (d) Have you ever checked your credit score? (if they have checked report, ask if credit report included credit score)
      i. If yes, when was the last time? What prompted you to check it? How did you do it? With which credit bureau? Only one or multiple?
      ii. If no, why not?
   (e) Do you feel that credit bureaus have an impact on your life?
      i. If yes, what is the impact?
      ii. If no, could you explain why not?
4. Have you ever heard of Equifax?
   (a) If yes, what do you know about it?
   (b) If no, "Equifax is one of the big three consumer-focused credit bureaus in the United States".
5. Equifax experienced a data breach in 2017. How much do you know about the data breach of Equifax?
   (a) Have you ever heard of this Equifax data breach before this interview?
      i. If yes, could you describe what happened based on your understanding?
      ii. If no, "It happened between May and July in 2017 and compromised the personal information (i.e. names, addresses, birth dates and Social Security Numbers) of over 145 million Americans".
   (b) In your view, what are the potential consequences of this breach?
   (c) What was your reaction when you heard about the Equifax data breach?
   (d) How do you feel about your data at Equifax now? Did it change after the breach?
   (e) Do you know if you were personally affected by this breach?
      i. Do you know if data about you was exposed in the data breach?
         A. If yes, how do you know?
      ii. Did you check if you were affected?
         A. If yes, how did you do it?
      iii. Did you check your credit reports at any point since you learned about the breach?
         A. When did you do it? How often?
         B. How did you do it?
         C. Only at Equifax or also at other credit bureaus?
   (f) Do you know what you could do to protect your credit data in general?
   (g) Did you do anything to protect yourself in response to the breach?
      i. Have you heard of fraud alerts?
         A. Can you describe what it is?
         B. Have you placed a fraud alert before or after the Equifax data breach?
         C. Can you describe how?
         D. With Equifax? With other credit bureaus? Which ones?
         E. Did you pay money for it?
         F. How long has the fraud alert been active for?

ii. Have you heard of a credit freeze?
   A. Can you describe what it is?
   B. Have you placed a credit freeze before or after the Equifax data breach?
   C. Can you describe how?
   D. With Equifax? With other credit bureaus? Which ones?
   E. Did you pay money for it?
   F. How would you unfreeze your credit?
iii. Did you start monitoring your credit and bank accounts more often since then?
   A. Can you describe how?
   B. With Equifax? With other credit bureaus? Which ones?
   C. Do you pay money for it?
iv. Have you heard of identity theft protection?
   A. Can you describe what it is?
   B. Have you signed up for any identity theft protection services?
   C. Can you describe how?
   D. With what company/entity?
   E. Do you pay money for it?
v. Did you do any other things not mentioned previously?

6. Before this breach occurred...
   (a) Have you ever experienced any data security problem, such as someone secretly changed your password?
   (b) Have you ever experienced identity theft, such as someone applying for credit cards under your name? (For each question, if yes, follow up with "Could you tell me more about the experience? Do you feel it has any impact on you?")

B. SCREENING SURVEY

Thank you for your interest in our study! Please answer a few questions about your demographics and availability for the interview.

1. In which year were you born?
2. What is your current gender identity?
   (a) Male
   (b) Female
   (c) Non-binary/third-gender
   (d) Not listed (please specify)
   (e) Prefer not to answer
3. What is the highest level of education you have completed?
   (a) Less than high school
   (b) High schhol degree or equivalent
   (c) Some college but no degree
   (d) Trade, technical, or vocational degree
   (e) Associate's degree
   (f) Bachelor's degree
   (g) Master's degree
   (h) Doctoral degree
   (i) Professional degree (JD, MD, etc.)
   (j) Other (please specify)
   (k) Prefer not to answer
4. Which of the following categories best describes your occupation?
   (a) Administrative support (e.g., secretary, assistant)
   (b) Art, Writing, or Journalism (e.g., author, reporter, sculptor)
   (c) Business, Management, or Financial (e.g., manager, accountant, banker)
   (d) Education or Science (e.g., teacher, professor, scientist)
   (e) Homemaker
   (f) Legal (e.g., lawyer, law consultant, or law professor)
   (g) Medical (e.g., doctor, nurse, dentist)
   (h) Engineering or IT Professional (e.g., programmer, IT consultant)
   (i) Service (e.g., retail clerk, server)
   (j) Skilled Labor (e.g., electrician, plumber, carpenter)
   (k) Unemployed
   (l) Retired
   (m) College student
   (n) Graduate student
   (o) Not listed (please specify)
   (p) Prefer not to answer
5. What was your total household income before taxes during the past 12 months?
   (a) Less than $25,000
   (b) $25,000 to $49,999
   (c) $50,000 to $74,999
   (d) $75,000 to $99,999
   (e) $100,000 to $124,999
   (f) $125,000 to $149,999
   (g) $150,000 or more
   (h) Prefer not to answer
6. What is your citizen status?
   (a) I am a citizen of the United States.
   (b) I am a permanent resident of the United States.
   (c) I am neither a citizen nor a permanent resident of te United States.
   (d) Other (please specify)
   (e) Prefer not to answer

   (If answer to above question was "citizen of the United States" or "permanent resident"
7. How many years have you been living in the United States?
   (a) $<$ 1 year
   (b) 1-2 years
   (c) 2-3 years
   (d) 3-4 years
   (e) 4-5 years
   (f) $>$ 5 years
   (g) Prefer not to say

C. CODEBOOK

Below is the codebook used for interview transcript analysis, grouped into four big categories.

C.1 Category: Financial Management

Financial status: Description of general financial situation, e.g., income, number of checking/saving accounts, number of credit cards currently held, late payment, as well as the mentioning of occupation, big purchases (e.g., cars and mortgages).

Financial tracking: The way to keep track of earnings and spendings, manage different credit cards, use checks or do everything online, the way of paying bills (e.g., set up automatic withdrawals or pay bills whenever it comes).

Financial behavior change: Any particular change in the ways of managing one's finance, how and why it occurred, may also include behavioral change resulting from attitudinal change (e.g., I tried to spend less because I wanted to save money).

C.2 Credit Bureau Related

Understanding of credit status: (1) The knowledge of the meaning and components of credit scores in general, how credit score is generated, whether it costs money to check credit scores, the mentioning that different bureaus may have different scores etc. (2) The impression of whether the participant's own credit score is good or bad, the description of when's the last time checking it and how to check it, where does the credit score come from (e.g., one of the three big bureaus or banks) (3) The impression of one's credit history, things included in the credit report, whether or not they have things like late payments and debts.

Awareness of credit bureaus: The number of credit bureaus, specific names of credit bureaus, also use this when they say they can't remember it or can't give the full name, also include the participant's knowledge or guess about whether there are bureaus other than the big three.

Impact of credit bureaus: "What impacts do credit bureaus have on you": how credit bureaus may impact consumer lives by giving credit ratings/scores or in other ways. Also include cases where participants say credit bureaus have little or no impact on them personally because of various reasons.

Check credit status at credit bureaus: Directly contact credit bureaus to access credit reports or sign up for other credit-related products and services, description of the process (e.g., schedule times to make use of the free opportunity to check credit reports annually).

Check credit status at other places: Usually through banks and third-party financial aggregation app (such as NerdWallet, Credit Karma, and Mint) to check credit history, credit score, or credit status in general, and the reason for doing it (e.g., it's free and more convenient), the frequency of the received updates, whether or not it might be helpful.

Reasons for no interactions: Description of having little or no interactions with credit bureaus, didn't check credit status through either credit bureaus or other places, and the reasons for doing it, e.g., I don't need to make big purchases or I don't want to know my credit status because it's poor.

Dispute process: Anything related to the dispute system within the credit reporting system, can be (1) the general telling that consumers have the right to dispute incorrect information; or (2) the complaint that the current dispute system doesn't work to solve consumers' problems (e.g., they have to spend a lot of time filing the dispute and it's hard to get the error eventually corrected).

Information providers of credit bureaus: Companies and organizations that provide information to credit bureaus, e.g., government, IRS, lending companies.

Customers of credit bureaus: Entities to which credit bureaus share or sell individual consumer's information, who may have the access to consumer credit files at credit bureaus. Also include cases where participants may not explicitly mention it but rather say it's an information exchange process, e.g., "I think that banks quarry them but they would also ask banks about".

Types of information collected: The types of information credit bureaus collect from their providers (e.g., checking accounts, savings, credit history, loans) about individual consumers, usually the answer following "what types of information do credit bureaus collect?" and "what do credit bureaus know about you?"

Offerings of credit bureaus: What information consumers can receive from credit bureaus, such as the annual free credit reports, credit reports that cost money to see credit scores, credit monitoring services.

Purpose of credit bureaus: This will refer to how credit bureaus use the collected information for, what their purposes are, e.g., assessing one's creditworthiness, generating credit scores. Answers following the question "what's the first word that you associate with credit bureaus" and "what are their purpose" might fall under this category.

Errors in mental models: This code encompass any obvious errors that we capture in participants' describing of credit bureaus.

Inaccurate credit files: Specific instance of negative perception - the experience that credit bureaus get errors on consumer credit files or retrieve the file of the wrong person, and hence leading to bad or unpleasant experience for consumers.

Opaque data aggregation process: Specific instance of negative perception - mentioning of the process how credit bureaus collect and aggregate all different types of information as opaque, unclear, not idea about what's going on behind the curtain.

Abusive use of power: Specific instance of negative perception - the mentioning that credit bureaus (and other related institutions such as governments and banks) are in the position of holding great power/have little interest in protecting consumer rights; consumers are in a relatively weak position.

Insidious data collection: Specific instance of negative perception - describing the data sharing between credit bureaus and data furnishers as passive, creepy or scary, without obtaining consent from consumers. As for consumers,

they have limited control and choice over this kind of data collection.

Positive perception of credit bureaus: Positive description of credit bureaus in general, the statement that credit bureaus have a positive image in the participant's mind.

Negative perception of credit bureaus: Negative description of credit bureaus, the statement that credit bureaus have a negative image in the participant's mind, note that if they just say "credit bureaus steal money from people" it doesn't count, there should be specific negative adjectives to describe it being bad or their negative feelings about it.

C.3 Risk Perception

Emotional feelings of the breach: The emotional feelings that participants experienced after heard of the breach (e.g., angry, disgusting, indifferent, not surprised), the emotional/attitude change towards Equifax (or other bureaus) after the breach compared to the time before.

Change of trust: Mentioning that after this breach, Equifax (or other credit bureaus) will have a less reputable image in the mind of consumers, or the participant personally will have less trust in the company.

Expectation of credit bureaus: Expectations towards Equifax, or other companies that have experienced data breaches about what they should do as the countermeasure of the breach, whether they have meet or failed the expectations in the past, as well as their expectations to these companies' future actions.

The class action lawsuit: The specific mentioning of the class action lawsuit against Equifax following the breach, whether participants might have heard of it or joined it, how they feel about it.

Prevalence of data breaches: The mentioning that there are too many previous data breaches in recent years that the occurrence of the Equifax breach doesn't make the participant too surprised, and that there is too much data available online.

Mentioning identity theft: Direct mentioning of identity theft or indirect conceptualization through examples as a consequence of the Equifax breach, or just identity theft in general.

Victims of the breach: Talking of targets that are more likely to be affected by the breach, e.g., people who have good credit.

Likelihood of being personally affected: The knowledge, assumption or assessment of whether participants themselves are personally affected, and if yes, to what extent, can be either an assured response or a guess.

Negative consequences of the breach: Mentioning consequences that's not about identity theft but can still happen after the Equifax data breach, such as invasion of personal privacy when so much personal and financial information was exposed.

Knowledge of Equifax: Impression of Equifax as a company, e.g., it's one of the big three credit bureaus, it's the one that got hacked, also include cases where participants say they've never heard of it.

Cause of the breach: The description that this breach was conducted by people other than hackers, such as governments, and/or it was profit-driven, e.g., some participants assumed that hackers will sell the stolen data to someone else, others believed that it's an internal breach and someone's disclosing the information intentionally.

Types of exposed data: Description of the general impression of some data being exposed in the Equifax data breach (e.g., a lot of personal information released) or specific types of data (e.g., SSN, credit card numbers). Also include cases where participants say they don't know.

Awareness of the breach: Memory of whether or not this participant has heard of the breach, what happened in general in the breach.

Previous data security experience: Previous experience of data security problem, such as being involved in a data breach and having password compromised somewhere.

Previous identity theft experience: Previous experience of being an identity theft victim, such as someone else applying for credit-related products under the participant's name, the effort in solving the related problems, or the reason for not conducting any kind of follow-up investigations.

C.4 Protective Actions

Check Equifax's website: The mentioning of someone (either the participant or other related people) check the Equifax website for his or her own breaching status. Also include this code when participants say they didn't check it.

Credit freeze: The action of placing a credit freeze, the interpretation of what credit freeze means/what's the expected outcome, the cost of credit freeze, why someone may want to initiate a credit freeze, their assumptions of what a credit freeze may do.

Check credit report after the breach: The mentioning of checking credit report following the data breach as a safeguard measure.

Fraud alert: The mentioning of placing a fraud alert on file, either for this breach or previous ones, their assumptions of what a fraud alert might do, the process of how to place a fraud alert.

Credit monitoring service: Enroll in credit monitoring services provided by credit bureaus, governments, or other entities.

Self-monitoring: The action of checking accounts more frequently, keeping an closer eye on them, and the related outcomes.

Identity theft protection: Conceptualization of what this type of service does, why someone may want it.

General security practices: Strategies to protect one's credit data/online privacy in general, e.g., don't disclose personal information such as SSN and passwords to others, avoiding suspicious emails, not using PayPal.

Self-initiated actions after the breach: Things that the participant has done in reaction to the breach or knows that they could have done, also include cases where they say they don't know.

Reasons for taking actions: Any reasons why the participant chose to take any one of the suggested actions above.

Reasons for not taking actions: Any reasons why the participant chose to not taking any one of the suggested actions above.

Triggering new actions: Any places where participants say they will or might consider doing some actions after the interview, the conversation inspires them to do something, and the reasons behind.

Suggestion from participants: The suggestion or proposal made by participants throughout the interview, e.g., credit bureaus shouldn't charge money for their certain offerings such as credit freeze, and there should be a consistent way to calculate credit scores.

Sources of recommendation: Protective actions recommended by anyone who's considered as reputable, trustworthy or expert by the participant, e.g., family member, financial advisor. Also include cases where participants said they provided recommendations for other people and hence became the source of knowledge.

Usability issues: Reporting about problems and hurdles participants encountered (or other people they know) when trying to initiate any one of the suggested actions.

Compensations after data breaches: Description of products and services offered by companies following previous data breaches that the participant or someone he/she knows was involved in (e.g., some companies may offer free or paid credit monitoring services and fraud alerts for victims).

Data Breaches: User Comprehension, Expectations, and Concerns with Handling Exposed Data

Sowmya Karunakaran Kurt Thomas Elie Bursztein Oxana Comanescu
Google Inc.
{sowmyakaru, kurtthomas, elieb, oxana}@google.com

ABSTRACT

Data exposed by breaches persist as a security and privacy threat for Internet users. Despite this, best practices for how companies should respond to breaches, or how to responsibly handle data after it is leaked, have yet to be identified. We bring users into this discussion through two surveys. In the first, we examine the comprehension of 551 participants on the risks of data breaches and their sentiment towards potential remediation steps. In the second survey, we ask 10,212 participants to rate their level of comfort towards eight different scenarios that capture real-world examples of security practitioners, researchers, journalists, and commercial entities investigating leaked data. Our findings indicate that users readily understand the risk of data breaches and have consistent expectations for technical and non-technical remediation steps. We also find that participants are comfortable with applications that examine leaked data—such as threat sharing or a "hacked or not" service—when the application has a direct, tangible security benefit. Our findings help to inform a broader discussion on responsible uses of data exposed by breaches.

1. INTRODUCTION

In recent years, data breaches have exposed the online credentials and personal data of billions of users across the Internet. In 2017 alone, news headlines announced that criminals had stolen usernames and passwords for 3 billion Yahoo users [16], the financial details of 143 million Americans collected by Equifax [10], and private data belonging to 57 million Uber users [17]. Once stolen, this data becomes readily accessible via black markets. Previous studies have identified over 3.3 billion credentials from breaches freely traded on the underground along with credit cards and other financial data [7, 25, 26]. Exposure puts victims at further risk of account takeover, financial theft, identity theft, or worse.

Despite repeated data leaks due to breaches, best practices for how companies should respond to incidents have yet to be formalized. One common remediation step—requested by both victims and increasingly by regulators [1, 3, 19, 22]—is that companies notify any affected victim. However, evidence that notifications influence user behavior is limited. For example, victims do not opt to switch to other, more secure services [1, 4, 14]. Moreover, companies do not always notify victims in a timely manner: Uber waited over a year before disclosing a $100,000 ransom payment in response to a breach [9].

At the same time, there are no clear boundaries for how one should responsibly handle data after it is leaked. Some security systems examine third-party breaches to protect victims from further harm: Google, Facebook, and Netflix automatically reset passwords for victims appearing in password dumps [2, 29]. Others provide information to victims, such as leak aggregation services that collect exposed credentials to help notify victims [13]. Exposed data also plays a role in the construction of password strength meters and investigations of underground market activity [5, 6, 18, 24, 28]. How victims weigh any potential security benefits against other concerns, including their privacy, remains uncertain.

In this paper, we bring users into the discussion of how companies should respond to breaches and how user data should be respected even after it finds its way on to black markets. We do this through two surveys. In the first, we asked 551 participants what actions a company should take upon learning of a data breach including technical solutions such as forcing password resets or enabling two-factor authentication. In the second part of our study, we surveyed 10,212 participants across six countries to assess their level of comfort and concerns towards eight scenarios capturing real-world situations where security practitioners, researchers, journalists, and commercial entities investigated data exposed by breaches. While we focus on lessons for the security community, we include these latter categories to act as a baseline comparison.

We frame our key findings as follows:

Data breaches feature prominently in the public's mind share: Over 93% of participants understood the meaning of a data breach. These participants cited identity theft (52%), the loss of personal information (25%), and monetary loss (9%) as their top concerns.

Notifications remain the most popularly requested remediation step: 83% of participants requested that companies affected by a breach send an immediate notifi-

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

cation to victims. Other more technical requests included enabling two-factor authentication on accounts (63%) and resetting exposed passwords (61%).

Users are supportive of applications that consume exposed data if they provide a direct security benefit: Of the 8 scenarios we examined, users were most comfortable with proactive password resetting in the event of reuse and sharing information with other identity providers.

Past experience as a victim of a breach increases support for security use cases: We observed significant differences in a prior victim's vs. non-victim's level of comfort for security related use cases. For example, 44% of prior victims expressed comfort with proactive password resetting compared to 34% of non-victims.

Users are wary of interacting with criminals (such as purchasing exposed data), but recognize the potential security benefits: For non-security use cases, over 70% of participants negatively expressed that purchasing exposed data was unethical and incentivized criminal behavior. However, with security related use cases only 40–51% participants expressed similar negative concerns.

Support for security use cases is consistent across countries: Although we observed significant differences in the absolute comfort levels between countries, every country consistently weighed security use cases over non-security use cases in terms of comfort.

2. RELATED WORK

Our work builds upon prior research into the experiences of victims of data breaches. In a study similar to ours, Ablon et al. surveyed 6,000 participants from the United States in 2015 and found 44% reported having received a data breach notification [1]. Credit card details topped the list of exposed data (49%), but health information (21%), social security numbers (17%), and account details (13%) also featured prominently. Reactions from participants to breaches were varied: only 11% of those surveyed stopped interacting with the affected company. More commonly, victims changed their password or PIN (51%) or switched to a new account (24%), while another 22% of participants did nothing at all. Other studies have also found that users rarely switch to another service or stop interacting with a company even upon receiving a breach notification [4, 14]. Our study examines in greater detail the expectations of breach victims and the technical remedies they most strongly prefer.

While financial theft features prominently in the concerns of victims, account takeover is also a significant risk. A survey by Shay et al. found that 15.6% of 1,502 survey participants self-reported having their account taken over [23]. A similar study by Rainie et al. found 21% of 1,002 adults experienced a social network or email account being hijacked [21]. These common experiences stem from billions of usernames and passwords exposed due to data breaches, with Thomas et al. estimating that data breach victims are 11.6x more likely to fall victim to account takeover than a random sample of users [25]. The prevalence of account takeover heavily influences the design of our study scenarios.

3. METHODOLOGY

We conducted two online surveys to evaluate user comprehension, attitudes, and expectations around data breaches. We describe each survey in detail. We refer readers to the Appendix for the full structure and text of both surveys.

3.1 Survey on responding to breaches (N=551)

Our first survey gauged user perceptions of risk surrounding breaches and how users would want a company to respond if their data had been exposed. We recruited participants via Amazon Mechanical Turk in July 2017 and administered the survey through Google Forms. Participants were asked to take a "simple task and experience survey." We avoided using the term "data breach" to prevent non-response bias. The survey took approximately 3 minutes to complete and participants were each compensated $0.50, including the screened out participants. In total, we received 604 responses, of which 551 feature in our final analysis.

3.1.1 Survey structure

We began the survey with a single screener question with three possible definitions of a data breach. The ordering of these options was randomized.

- Public exposure of usernames and passwords of millions of users of an online system. (N=564)
- Using large sets of data to aid robots to solve a problem that humans cannot solve. (N=13)
- Web page that is unable to load due to too much data on the page. (N=27)

Overall, 564 of 604 participants chose the correct definition and were allowed to continue through the rest of the survey. We dropped the remaining 40 participants from any further questions.

Following the screener, we asked participants to select the single most important "harm" that might arise from their password being exposed through a data breach and what remediation steps a company should take to protect the participant's account. Finally, we asked participants to rate their level of comfort with six potential actions a company could take in response to a breach. For each action, in addition to rating their level of comfort, participants provided an open-ended reason for their rating.

Outside these core questions, we asked whether participants had ever been the victim of a data breach. We also included two quality control questions, and six demographic questions. In total, we eliminated 13 inattentive responses where participants answered both quality control questions incorrectly, leaving a total of N=551 responses.

We reviewed a small sample (N = 50) of open-ended responses and developed codes. The rest of the open ended responses were then assigned codes through manual inspection. Responses that did not fall into any of the coding buckets were categorized under 'Other'. Roughly 3% of responses were blank which we did not categorize. The researcher not involved in the coding process conducted the quality checks by independently reviewing a sub-sample. The agreement rate was about 90%.

3.1.2 Survey development

Prior to running the survey, we conducted an initial pilot (N=34) where the single most important harm was left as an open-ended question. We then codified the most popular responses, selecting eight possible options for the final survey. We also expanded the list of remediation steps to include new, incorrect steps (e.g., buying a new computer) to gauge comprehension. We also switched from strictly asking each participant's comfort towards certain responses to also requesting their reasoning. We then ran a second pilot (N=31). We used the open-ended responses to clarify the six actions a company might take in response to a breach. Finally, we added a demographic question related to whether participants had ever been a victim of a previous breach.

3.1.3 Participant demographics

For the 551 participants, 52% identified as male, 47% identified as female, and 1% preferred not to answer. Roughly 12% were 18–24, 45% 25–34, 22% 35-44, 13% 45–54, 6% 55–64, and 2% older than 65 or preferred not to say. In terms of education, 47% had a bachelors degree, 19% a masters degree or higher, and 17% some college education. Participants predominantly resided in the United States—69%—with another 23% residing in India and 8% in other countries. In terms of employment, 80% were had some form of employment (53% full-time, 17% self-employed, and 10% part time), 8% were students, and 12% were unemployed, retired, or looking for work.

3.1.4 Limitations

In terms of the study sample, although the user population on Mechanical Turk is relatively diverse for an Internet sample, there is still a bias. For example, the Mechanical Turk workers are considered WEIRD (Western, educated, industrialized, rich, and democratic) [15]. To reduce the effect of this bias, we opened the survey to residents of all countries, not just United States residents. However, the underlying demographics of workers still skews towards the United States and India.

3.2 Survey on breach data use cases (N=10,212)

Our second survey examined user comfort towards a spectrum of use cases that handle data exposed by breaches. We recruited participants through an international panel provider that recruits through online communities, social networks and the web. The panel provider also enforced strict quality controls such as digital fingerprinting to identify duplicate participants and pattern recognition to flag fraudulent responses. As such, we do not embed any quality control questions in the survey questions. We specifically stratified our sample to participants from the United States, Canada, United Kingdom, Australia, India, and Germany. We administered the survey using the online survey platform and panel provider Qualtrics. We paid $6 per response to our panel provider, a portion of which was paid to the participants as incentive.

3.2.1 Survey structure

We used a scenario based survey to frame eight potential use cases of data exposed by breaches. To minimize fatigue, each survey was structured to included only two scenarios randomly selected from our pool of eight. When considering a scenario, we asked users to rate their level of comfort if they knew the data had been purchased from criminals via a black market; to explain their rating in an open-ended question; and finally whether their level of comfort would change if they knew the data was freely available. We also included six demographic questions and one question on whether the participant had previously been a victim of a breach. We outline each scenario and highlight real-world equivalents. In total, we received 10,212 responses, with over 400 responses per scenario and per country.

Security research (S1): In the first scenario, we framed whether it was acceptable for a researcher at a university to use data exposed by a breach to study how users select passwords. Examples of such research in practice include studies of password reuse [5] and the development of better password strength meters from existing, exposed data [28, 6, 18].

Hacked or not service (S2): We asked participants whether it was acceptable for a company to provide a paid service where anyone could query for a "username" to determine whether their data was exposed due to a breach. A multitude of such services currently exist, such as *haveibeenpwned.com*, *breachalarm.com*, and *leakedsources.com*. In practice, some of these services operate on donations and only reveal whether an account was present in a breach. Others require a monthly fee and allow a subscriber to look up any username and its associated passwords, at times running afoul of law enforcement [20].

Threat sharing, finance and social (S3, S4): For two scenarios, we asked whether participants were comfortable with a breached company sharing the email addresses of victims with third-party services to protect against lateral attacks. We offered two, independent scenarios for the third-party service involved: a financial institution and an online social network. These scenarios mimic emerging threat exchange services where companies share information on ongoing attacks.

Proactive password resetting (S5): We asked participants whether they were comfortable with a service finding usernames and passwords exposed in third-party breaches to proactively re-secure the participant's account if they reused an exposed password. This scenario matches how Google, Facebook, and Netflix currently reset passwords for victims appearing in third-party breaches [29, 2].

Journalist, tax fraud (S6): We framed whether participants were comfortable with a journalist writing an article on tax evasion that sourced their materials from private emails exposed due to a breach. Rough equivalents include the Panama Papers [11] and Paradise Papers [8] that exposed millions of email records detailing the financial dealings of offshore investments and entities.

Journalist, dating site (S7): We examined whether it was appropriate for a journalist to use personal information from breached data profiles as source material for an article. Recent examples include the leak of Ashley Madison users, which media outlets used to expose the activities of registered members.

Competitor (S8): We framed whether it was appropriate for a non-breached company to contact victims in order to advertise switching services. For example, after the Equifax breach, one identity theft provider created ads and press released to announce how it could help victims [12].

3.2.2 Survey development

Prior to running our survey, we conducted two pilots. The first involved user researchers at our institution who provided feedback on the framing text of the scenarios. The second pilot involved a small sample of participants (N=40). Based on the responses, we added a follow-up question for every scenario to understand whether a participant's comfort would change if data was freely available.

3.2.3 Participant demographics

For the 10,212 participants, 51% identified as male, 48% female, and 1% preferred not to answer. In terms of age, 11% were 18–24, 28% 25–34, 19% 35-44, 17% 45–54, 13% 55-64, and 9% older than 65. Participants were equally distributed across six countries: 16% in Australia, 18% in Canada, 17% in Germany, 14% in India, 16% in the United Kingdom, and 15% in the United States. 46% indicated to be employed full-time, 13% employed part-time, 13% retired, 5% students, 7% self-employed, 7% home makers, 5% unemployed and 4% other. In terms of education, 5% indicated receiving less than high school education, 17% High School, 18% Some college no degree, 15% Associate's degree, 28% Bachelor's degree, 12% Master's degree, 1% Ph.D and 3% Other.

3.2.4 Limitations

Our surveys were spread across several weeks, however we could not control for respondent's exposure to external information such as news stories and press articles on data breaches. In addition, given that our approach relies on scenarios based assessment, one can argue the presence of availability bias. Availability heuristic is a mental shortcut that relies on immediate examples that come to a given person's mind when evaluating a specific topic, concept, method or decision [27]. In reality, users would have access to many other pieces of input about the scenario at hand which may also play a role in influencing their level of comfort. Gut reactions and framing may also influence the perceived acceptability of the scenarios we explored. Likewise, privacy enhancing technologies might help to allay user concerns with respect to data sharing.

4. RESPONDING TO A DATA BREACH

We report on the results of our first survey, which explored the familiarity of participants with data breaches as both a concept and a personal experience. We present how participants perceived the risk of breaches, what actions they felt companies should take in response to a breach to protect victims, and finally their level of comfort with companies engaging with the press, government, criminals, and other companies as part of remediation.

Comprehension: As a first step towards interpreting our results, we examined whether participants were familiar with data breaches and their accompanying risk. The vast majority of participants (N=564, 93%) correctly identified the definition of a breach from one of three choices. As shown in Table 1, their top concerns of what harm might arise included identity theft (N=287, 52%) and the leak of personal information (N=138, 25%). Monetary loss was a distant third (N=50, 9%), possibly due to our framing of data breaches as relating to usernames and passwords. Impossible harms, such as a participant's computer being infected with a virus, were selected by only 17 participants (3%). More than a hypothetical experience, 232 participants (42%) reported having had their data exposed by a prior breach while 65 participants (12%) reported not knowing. These results suggest that participants are both familiar with the concept of a data breach and the resulting consequences.

Table 1: Top harm that results from a data breach.

Potential harm	Breakdown	N
Identity theft	52%	287
Leak of personal information	25%	138
Monetary loss	9%	50
Loss of access to personal information	5%	28
Phone being monitored by hackers	3%	17
Computer being infected with virus	3%	17
Spam being sent out from your account	2%	11
Other	1%	4
No harm	< 1%	2

Table 2: Ranking of remediation steps companies should take in response to a breach.

Remediation step	Breakdown	N
Send you an immediate notification	83%	457
Enable two-factor authentication	63%	347
Reset your password	61%	336
Provide credit monitoring	56%	309
Issue a refund	39%	215
Give you a new account	32%	176
Change your username	31%	171
Pay users a consolation bonus for breaking their trust	29%	160
Upgrade your web browser	15%	83
Company buys you a new computer	5%	28

Preferred remediation steps: Table 2 provides a breakdown of the remediation steps participants selected as the best ways companies could protect their account in the event of a breach. Participants most frequently requested that companies send an immediate notification to affected users (N=457, 83%). This was followed by more technical account protections such enabling two-factor authentication (N=347, 63%) and resetting an account's password (N=336, 61%). Some of these actions mirror steps that victims self-report taking in response to a breach, such as 51% of victims changing their password or PIN [1]. However, the same is not true for two-factor authentication: fewer than 3% of hijacking victims adopt two-factor authentication after learning their account was compromised [25]. This suggests a disconnect between understanding the protections two-factor authentication provides and actual adoption.

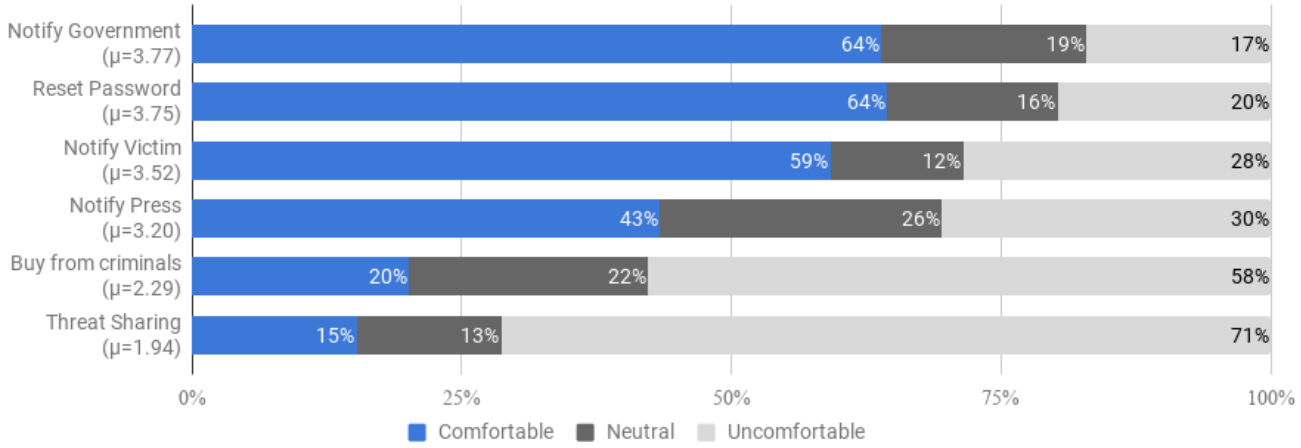


Figure 1: Comfort of participants towards breached companies dealing with victims, criminals, the press, the government, and other companies. We binned ratings of 1 or 2 as uncomfortable, 3 as neutral, and 4 or 5 as comfortable.

Account security measures and communication outranked financial protections, such as credit monitoring (N=309, 56%) or companies issuing a refund (N=215, 39%). This mirrors participants' perception of harm, where monetary loss ranked lower than identity theft or data loss. A small but not insignificant group of participants selected ineffective remediation steps that would provide no security benefit in the context of data breaches. These actions included changing usernames (N=171, 31%) or upgrading web browsers (N=83, 15%). The latter action suggests that users may conflate general security best practices such as keeping software up to date with something that might protect them from a breach.

Remediation and the wider ecosystem: Beyond user-centric remediation steps, we asked participants to rate how comfortable they were with companies taking a range of actions such as communicating with criminals, the press, the government, and other companies in response to a breach. We measured comfort on a scale of 1 to 5, with 1 indicating "Not at all comfortable" and 5 indicating "Very comfortable." We relied on two user-centric actions, namely resetting passwords and notifying victims, as a baseline comparison. Figure 1 shows the spectrum of ratings participants selected.

In the case of notifying victims of the breach, participants in aggregate rated the action with an average comfort level of $\mu = 3.52$. Common themes that correlated with a positive level of comfort—surfaced in the coded open-ended questions—included an obligation on the part of the company to be transparent (N=194, 36%) and that such a notification would allow participants to reset their password (N=46, 8%). Conversely, participants that were uncomfortable frequently cited that notifications made them feel insecure (N=63, 12%) and that it did nothing to make up for the loss of data (N=47, 9%). Neutral participants often cited that companies needed to do something more (N=45, 8%). For example:

> P474: "The notification is important, however, the company must also inform about the corrective measures it intends to take."

In comparison, participants were more favorable with notifying the government ($\mu = 3.77$), though less favorable of notifying the press ($\mu = 3.20$). The positive affinity towards government activity relates to prosecuting criminals and holding companies responsible (N=224, 41%):

> P355: "In order to prevent other breaches I think the government should be involved at helping catch the criminals responsible."

Unique concerns for reaching out to the press included feeling that victims should be contacted directly (N=77, 14%) and that headlines might attract criminals to take advantage of the exposed data (N=28, 5%).

> P406: "...making it too public may inspire others to try and take advantage of the breach".

Beyond notifications, a majority of participants expressed discomfort ($\mu = 2.29$) with companies reaching out to criminals to buy a copy of the leaked data to know what was exposed. Participants commonly cited that it was unethical to deal with criminals (N=89, 16%) and that it would incentivize further attacks (N=110, 20%):

> P24: "That shows the hackers that that company can be bullied, making them future targets for hacks."

Surprisingly, participants rated the prospect of companies sharing exposed usernames and passwords with other identity providers as the least comfortable action a company could take, lower even than dealing with criminals ($\mu = 1.94$). Common concerns included a violation of the participant's trust (N=205, 38%) and feeling it exacerbated the problem by exposing private information further (N=99, 18%):

> P338: "OMG no. I don't want my info shared!!!"
> P399: "The company has no permission to share my data, even if it was already stolen."

Table 3: Comparison of the level of comfort for past breach victims and non-victims. We note statistically significant differences with an astericks.

Remediation step	Comfort (victim)	Comfort (non-victim)	p-value
Notify government	3.90	3.65	0.018**
Reset password	3.79	3.73	0.443
Notify user	3.67	3.37	0.047**
Notify press	3.35	3.06	0.012**
Buy from criminals	2.23	2.35	0.217
Threat sharing	1.88	2.00	0.178

Taken as a whole, our findings indicate that participants are comfortable with actions that lead to better protections or even catching the criminals involved. However, participants expressed a strong degree of discomfort for actions that might further distribute exposed data or encourage future criminal activity. These concerns heavily influenced the design of our second survey (Section 5).

Influence of breach experiences on comfort: As an added dimension, we examined how prior experience with a data breach influenced a participant's level of comfort towards various actions. For our analysis, we treat participants that reported as being unsure if they had been part of a previous data breach as non-victims. Table 3 presents our results. Overall, victims reported a higher level of comfort for notifying users, the press, and the government than non-victims, while actions beyond notification saw no statistically significant difference.

5. HANDLING EXPOSED DATA

Turning to our second survey, we report how participants valued security applications built from exposed data and the trade-offs they perceived. We also examine how demographic variations and past experience with a breach influence a participant's level of comfort.

5.1 Scenarios, in depth

We provide a ranking of participant comfort to all eight scenarios in Figure 2. Participants were most comfortable with scenarios that helped to directly protect them from further risk, such as resetting reused passwords and working with other identity providers to prevent lateral attacks. In contrast, security protections that might help in the abstract, such as a "hacked or not" service or research in password security were rated lower. We explore each scenario (in order of comfort level) and the top concerns that participants surfaced through our open-ended questioner.

Threat Sharing, Finance: Participants were most comfortable when presented with a scenario of a breached company working with another identity provider—in this case a financial institution–to share threat intelligence of victims ($\mu = 2.94$). The stated goal of this sharing was to enable password resetting at the financial institution to protect victims from financial fraud. Based on our coded responses, participants most frequently expressed a lingering fear their financial assets remained at risk (56%) and skepticism resetting a password would dissuade criminals (19%). For one participant, this was an intimate experience:

> P[8920]: "[the breached company] owes explanation how my email got hacked in the first place and why they didn't protect me. This exact scenario happened to me with Yahoo and Paypal and somebody got into my account, took my Paypal credit card number and charged thousands of dollars at Walmart on it."

Despite these concerns, participants still remained neutral or positive on threat sharing as a minimum step towards responding to a breach. For example:

> P[7429]: "They are doing something to help fix a problem and partnering with a trusted company, so I have no objections to their being proactive."

Threat Sharing, Social Network: Similar to the previous threat sharing scenario, participants reported the second highest level of comfort when a social network was the recipient of threat intelligence ($\mu = 2.92$). Overall, participants most frequently cited privacy as their top concern (43%). Others felt that the security benefit outweighed any privacy concerns (20%) or welcomed the extra level of protection (22%):

> P[385]: "Of course which [sic] is also an invasion of my "privacy", but I find it a justified and proper engagement in order to protect other accounts before a hacker attack."
> P[7668]: "A proactive approach on the part of [the breached company] is likely the best means of blocking fraudulent activity and instituting counter-measures."

Proactive Password Resetting: When asked about a company purchasing third-party credential dumps from criminals to proactively protect against password reuse, 35% of participants reported being comfortable with such an activity ($\mu = 2.85$). Of participants, 53% stated this would enhance their security and another 16% that it was good to see proactive activity.

> P[9615]: This is a proactive step from [the company], and one that they are not actually obligated to do. This makes me feel like the company cares about protecting my identity.

However, another 25% of participants were concerned with the legality of such activity even if were beneficial, or whether it might encourage criminals:

> P[8941]: They're paying people who obtained the information illegally. This seems a bit odd, almost like they're encouraging people to hack sites.

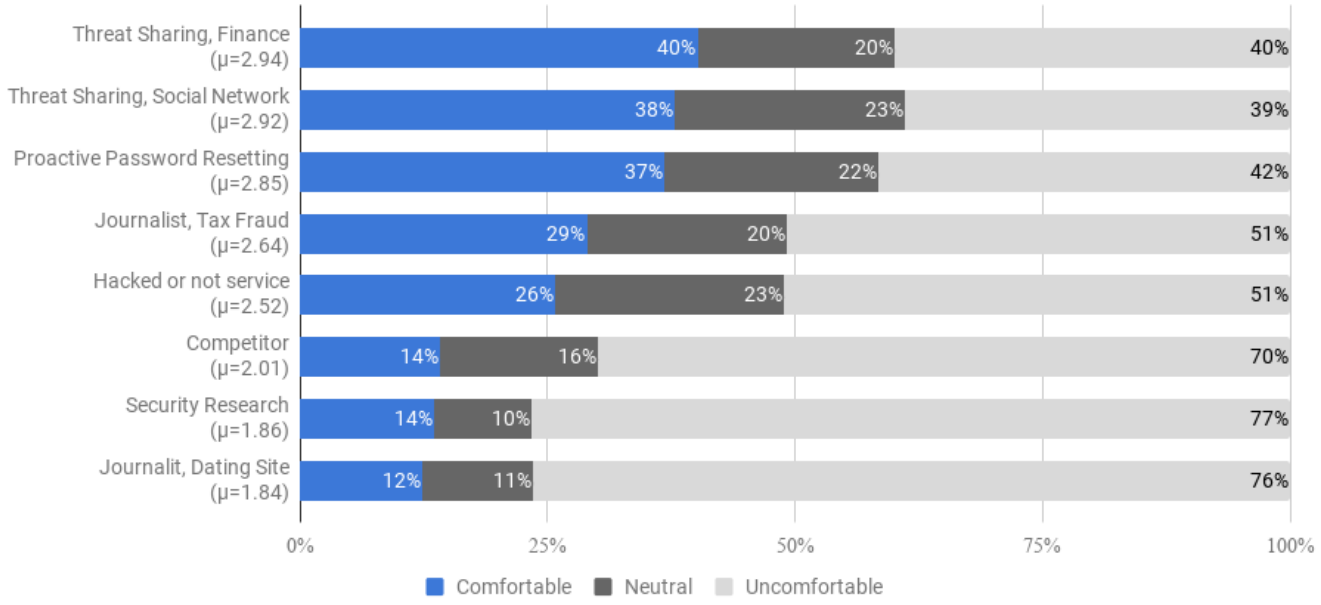


Figure 2: Participant comfort towards the eight scenarios involving purchasing (or where the source of data was not applicable). Participants ranked scenarios that provided direct security benefits higher than all other scenarios.

Less frequent, 4% of participants highlighted ethical concerns with any purchasing of data from criminals:

> P[575]: "I believe it's unacceptable for any company, whether their motives are good, to purchase or otherwise obtain illegally-gained data, especially personal information. ... It's a blatant disregard for people's privacy."

Surprisingly, participants were less comfortable ($\mu = 2.73$) when the data was freely available, a phenomenon also observed with the "hacked or not" service scenario as shown in Figure 3. We did not collect open-ended follow ups in conjunction with asking participants about freely available exposed data, so we cannot definitively state why this is the case. One hypothesis is that participants may have felt the damage is already done if credentials become freely available.

Journalist, Tax fraud: As a source of comparison, we asked participants their level of comfort towards journalists using data exposed by a breach to investigate fraud. Roughly 30% of participants reported being comfortable purchasing data from criminals to conduct such an investigation ($\mu = 2.64$). More participants expressed comfort when the data was freely available ($\mu = 2.77$). Participants frequently raised concerns about the legality of such behavior (56%):

> P[5414]: "Obtaining the information illegally doesn't make me feel comfortable. If it was handed to him for free, this feels a little less immoral."
> P[4607]: "It's important that the information be obtained and revealed, but [the journalist] has done so by potentially breaching the privacy of innocent individuals."

Others supported the journalist's actions, with the ends justifying the means:

> [7830]: "Even though the method is unethical, he is exposing a corruption. I will have to trade my uncomfortableness."
> P[5102]: "Whilst I wouldn't necessarily condone the hacking element, it is now a fact of modern society that these methods of information gathering are available. ... Publishing what was found through that means is in the public interest."

As with purchasing credential dumps, participants fall into a spectrum of ethical frameworks. For some, there is never a justification for using private data. For others, the value extracted from exposed data can override privacy concerns.

Hacked or not service: When asked to rate their level of comfort towards a service aggregating breaches to provide a "hacked or not" service, 25% of participants reported being comfortable ($\mu = 2.52$). As with proactive password resetting, comfort dropped when data was freely available ($\mu = 2.43$). Participants frequently cited the trustworthyness of the "hacked or not" service operator as their top concern (58%). Participants also felt any purchase would encourage criminals. In the words of participants:

> P[7550]: "It makes me fear that they work with the hackers and may not be trustworthy."

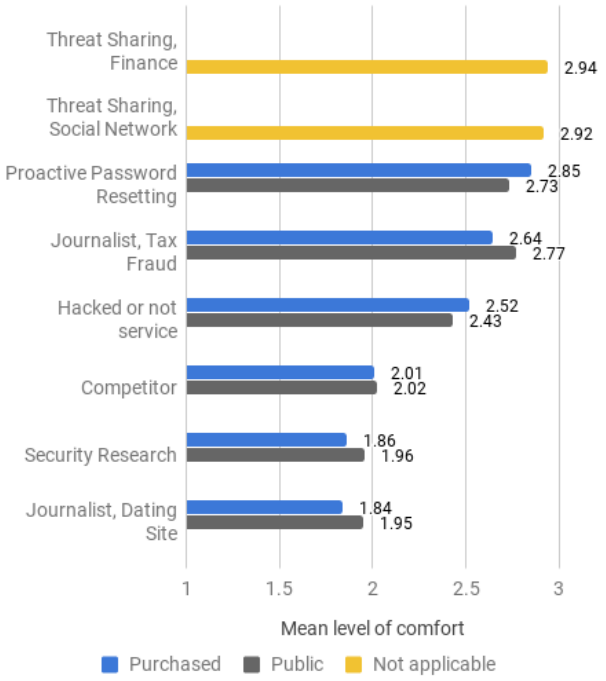


Figure 3: Comparison of comfort towards scenarios that involved purchasing data from the black market versus scenarios where data was freely available. All differences are statistically significant.

> P[4868]: "The fact that they are buying it from the hackers themselves is of concern....how do they know them? How are they getting the info? Do they have a relationship with the hackers?"

Despite these concerns, participants remained neutral on the prospect of such a service. A minority of participants (9%) highlighted the benefit of such a service, given a lack of clear notifications:

> P[10188]: "It is a good idea to see whether your account has been hacked. How else would you find out?"

Competitor: Only 14% of participants reported being comfortable with a competitor purchasing data from criminals in order to identify victims and offer for them to switch services ($\mu = 2.01$). There was little change in comfort if the data was freely available ($\mu = 2.02$). Participants frequently cited ethical concerns (44%) and the illegality of such behavior (39%).

> P[6145]: "This seems very unethical since they plan to gain profit from data that has been stolen."

However, 17% of participants expressed they would be better off in the end:

> P[622]: "This is a cheap shot to get consumers but at this stage I would probably go with [the competitor] as I know they have the proper software to avoid hackers."

Security Research: Faced with the prospect of researchers purchasing stolen credentials to study, participants reported the second lowest level of comfort compared to other scenarios ($\mu = 1.86$)—behind using stolen data for advertising. This comfort increased slightly when researchers obtained credentials from a free source ($\mu = 1.96$). Based on our coded responses, participants' top concern was the legality of the researchers actions (45%):

> P[9507]: "This may help his research and the result may help millions of internet users but the way he acquires the data is illegal and without the permission of account owners."

Other negative reactions included breaching the privacy of the victim (12%) and a sentiment that it was unethical to deal with criminals (9%):

> P[9307]: "It's incredibly unethical for [the researcher] to buy passwords from hackers. It's no different than someone buying a car that was stolen."

Another 23% of participants felt the value of research outweighed other concerns:

> P[7953]: "I have faith that this action will ultimately contribute to research that will make the general population less vulnerable in the long run."

These results suggest that, in the absence of a tangible security benefit, the privacy and ethical concerns of participants outweigh any potential justification. Surprisingly, research is viewed in even lower light than a scenario of a competitor advertising to victims of a breach, yet the latter still provides the prospect of a tangible benefit.

Journalist, Dating site:: A mere 13% of participants reported being comfortable with journalists using exposed dating profiles purchased from criminals to reveal the private lives of entities involved ($\mu = 1.84$). Comfort increased when data was freely available ($\mu = 1.95$). Most participants cited this was illegal (50%), a breach of privacy (27%), and unethical (11%).

> P[562]: "I find it very disconcerting that someone thinks they have a right to invade my space in any way without permission. It makes me want to withdraw from computors [sic] FULLSTOP. ... Makes for an unsafe unstable unfair dog eats dog world. Where has human respect, honesty and compassion gone."

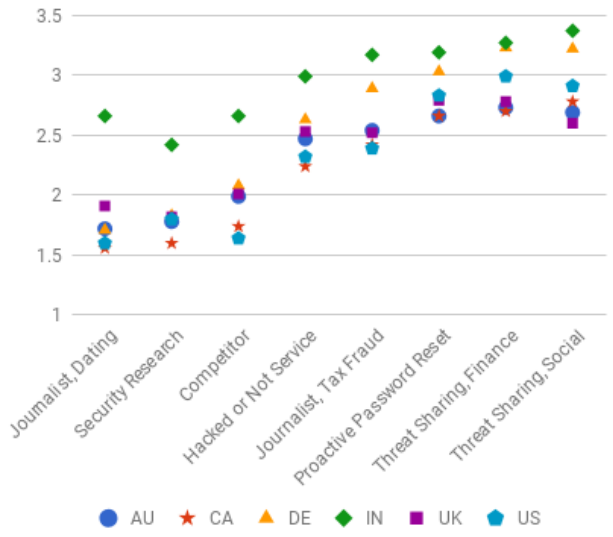


Figure 4: Mean level of comfort per scenario, broken down by country. Where applicable, comfort reflects the sub-scenario of purchasing data from a breach. Participants consistently rank security scenarios as the most comfortable compared to other scenarios.

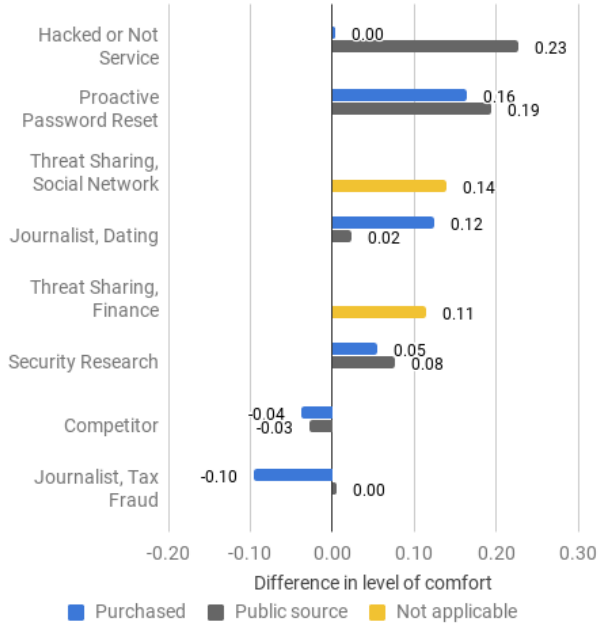


Figure 5: Difference in the comfort between prior victims of breaches and non-victims. Victims are consistently more likely to support security scenarios. All differences are statistically significant.

Some participants put a positive spin on the activity, stating it would help expose the threat of data breaches (6%). Others emphasized freedom of speech above all else (3%):

> P[1613]: "Concrete examples of how the hack has affected the lives of ordinary people makes the story more relatable."
> P[3878]: "Freedom of the press is essential in a well-functioning democracy. Reporters must come in some way to their publications."

5.2 Demographic variations

Table 4 provides a detailed summary of how the level of comfort, measured as a mean, compared across genders, age-groups and countries alongside the results from tests for statistical significance ($p = 0.05$). We corrected for multiple testing for Age-groups and Countries, using Bonferroni correction (adj.$p = 0.008$).

Differences across age groups: Across age groups, younger participants had a higher level of comfort with handling exposed data than older participants. The difference in comfort ranged between 0.14–0.68 for all scenarios, with the exception of the scenario of journalists reporting tax fraud via a public source. With research suggesting that younger adults are more likely to be victims of breaches [30], this may reflect a greater desire for solutions to a common experience.

Differences across country: Among the six countries we surveyed, participants from India and Germany indicated the highest level of comfort towards scenarios where entities purchased exposed data. Canada and Australia expressed the lowest levels of comfort (Figure 4). The difference in comfort for each scenario ranged between 0.53–1.10 per country. Despite absolute differences in comfort, participants universally rated security applications as more comfortable relative to other scenarios.

Differences based on breach experience: Of the respondents we surveyed, 24% self-reported having experienced a data breach. Among countries, 40% of US respondents reported experiencing a breach. Between genders, we did not see a significant difference between men (25%) and women (23%). Among age-groups we observed that more respondents between the ages 25-44 years reported experiencing a data breach (28%).

Figure 5 shows the difference in the level of comfort for prior victims of breaches and non-victims for each scenario. A positive value indicates victims were more comfortable than non-victims. Overall, victims reported consistently higher levels of comfort for all scenarios, including purchasing from criminals. The only exceptions were the competitor and tax fraud scenarios. The highest change in comfort related to a "Hacked or not" service. One explanation is that victims are more familiar with the harms that can result from a breach and are thus more supportive of security applications.

Differences across genders: Across genders, men had a higher level of comfort than women ($\mu = 2.57$ vs. $\mu = 2.41$). This was true in all scenarios, other than a "Hacked or Not" service where men had the same comfort level as women.

Table 4: Comparison of mean level of comfort across demographics: Gender, Age-group, and Country.

Metric	Gender		Age-Group						Country					
	Male	Female	18-24	25-34	35-44	45-54	55-64	65+	AU	CA	DE	IN	UK	US
Research (Purchased)														
Mean	1.96	1.75	2.06	2.08	1.75	1.74	1.76	1.54	1.78	1.60	1.83	2.42	1.82	1.80
Test Statistic	$p < 0.001$; Mann-U		$p < 0.001$; KW test; $\chi^2 = 64.8$						$p < 0.001$; KW test; $\chi^2 = 87.62$					
Research (Public)														
Mean	2.09	1.83	2.21	2.19	1.93	1.88	1.76	1.53	1.84	1.80	1.92	2.44	1.89	1.96
Test Statistic	$p < 0.001$; Mann-U		$p < 0.001$; KW test; $\chi^2 = 79.62$						$p < 0.001$; KW test; $\chi^2 = 55.61$					
Hacked or Not (Purchased)														
Mean	2.51	2.52	2.68	2.82	2.50	2.37	2.25	2.11	2.47	2.24	2.63	2.99	2.53	2.32
Test Statistic	$p = 0.597$; Mann-U		$p < 0.001$; KW test; $\chi^2 = 89.27$						$p < 0.001$; KW test; $\chi^2 = 80.77$					
Hacked or Not (Public)														
Mean	2.48	2.38	2.40	2.68	2.41	2.36	2.30	2.05	2.33	2.24	2.39	2.73	2.40	2.47
Test Statistic	$p = 0.122$; Mann-U		$p < 0.001$; KW test; $\chi^2 = 51.53$						$p < 0.001$; KW test; $\chi^2 = 23.71$					
Threat Sharing, Finance														
Mean	3.00	2.88	3.05	3.10	2.99	2.93	2.76	2.49	2.73	2.70	3.23	3.27	2.78	2.99
Test Statistic	$p = 0.023$; Mann-U		$p < 0.001$; KW test; $\chi^2 = 48.19$						$p < 0.001$; KW test; $\chi^2 = 80.47$					
Threat Sharing, Social Network														
Mean	2.98	2.87	3.05	3.07	2.97	2.89	2.69	2.67	2.69	2.78	3.22	3.37	2.60	2.91
Test Statistic	$p = 0.035$; Mann-U		$p < 0.001$; KW test; $\chi^2 = 32.69$						$p < 0.001$; KW test; $\chi^2 = 114.13$					
Proactive Password Reset (Purchased)														
Mean	2.91	2.80	3.13	3.04	2.70	2.83	2.69	2.56	2.66	2.66	3.03	3.19	2.79	2.83
Test Statistic	$p < 0.001$; Mann-U		$p < 0.001$; KW test; $\chi^2 = 51.58$						$p < 0.001$; KW test;$\chi^2 = 53.97$					
Proactive Password Reset (Public)														
Mean	2.83	2.63	2.92	2.82	2.66	2.80	2.61	2.40	2.60	2.61	2.89	2.77	2.57	2.95
Test Statistic	$p = 0.122$; Mann-U		$p < 0.001$; KW test; $\chi^2 = 29.7$						$p < 0.001$; KW test; $\chi^2 = 29.5$					
Journalist, Tax Fraud (Purchased)														
Mean	2.76	2.52	2.69	2.84	2.59	2.58	2.43	2.55	2.47	2.24	2.63	2.99	2.53	2.32
Test Statistic	$p < 0.001$; Mann-U		$p < 0.001$; KW test; $\chi^2 = 36.05$						$p = 0.001$; KW test; $\chi^2 = 112.6$					
Journalist, Tax Fraud (Public)														
Mean	2.86	2.68	2.78	2.88	2.84	2.73	2.55	2.78	2.33	2.30	2.39	2.73	2.40	2.47
Test Statistic	$p = 0.001$; Mann-U		$p = 0.003$; KW test; $\chi^2 = 18.22$						$p = 0.001$; KW test; $\chi^2 = 21.6$					
Journalist, Dating Site (Purchased)														
Mean	1.97	1.69	1.98	2.12	1.86	1.62	1.60	1.51	1.72	1.56	1.71	2.66	1.91	1.60
Test Statistic	$p < 0.001$; Mann-U		$p < 0.001$; KW test; $\chi^2 = 87.60$						$p < 0.001$; KW test; $\chi^2 = 214$					
Journalist, Dating Site (Public)														
Mean	2.08	1.81	2.19	2.16	2.01	1.78	1.72	1.57	1.88	1.78	1.81	2.60	2.01	1.74
Test Statistic	$p < 0.001$; Mann-U		$p < 0.001$; KW test; $\chi^2 = 97.2$						$p < 0.001$; KW test; $\chi^2 = 127.8$					
Competitor (Purchased)														
Mean	2.06	1.94	2.28	2.23	2.05	1.87	1.72	1.68	1.99	1.74	2.08	2.66	2.01	1.64
Test Statistic	$p = 0.008$; Mann-U		$p < 0.001$; KW test; $\chi^2 = 92.30$						$p < 0.001$; KW test; $\chi^2 = 160.20$					
Competitor (Public)														
Mean	2.06	1.98	2.30	2.24	2.00	1.92	1.77	1.69	1.90	1.80	2.08	2.53	2.04	1.85
Test Statistic	$p = 0.185$; Mann-U		$p < 0.001$; KW test; $\chi^2 = 79.56$						$p < 0.001$; KW test; $\chi^2 = 85.04$					

Figure 6: Tree map of open-ended responses from second study suffixed with the term "my consent".

6. DISCUSSION

User expectations after a breach: Over 40% of participants from the United States and 25% across the United Kingdom, Germany, Australia, Canada, and India reported being former victims of a breach. Our results indicate participants have strong expectations of being notified when their data is exposed. In the case of credentials, participants expressed this allows them to take precautionary measures such as resetting their password for all their affected accounts. Participants also emphasized that transparency remained important, even when a notification alone was viewed as an insufficient response. Other proactive measures included force resetting passwords or otherwise hardening security around accounts, such as with two-factor authentication. Outside of a company's responsibilities to users, participants also strongly favored interacting with government authorities as a means of holding criminals accountable—as well as the breached company.

Community responses to a breach: Digital identity is not an island; a breach at one company may allow criminals to access other resources due to reused passwords or recovery questions. Our results indicate that participants are supportive, or at least neutral towards, emerging security strategies by identity providers. Between 37–40% of participants expressed comfort towards threat sharing between identity providers as well as proactively resetting passwords exposed by third-party breaches. Another 20–23% of participants expressed a neutral opinion towards these activities. Lingering concerns for participants fell into two categories: skepticism any such actions would help secure their accounts, and strong expectations about privacy and ethical behavior—even when companies can acquire data without engaging with black markets.

Conversely, participants expressed a greater degree of concern towards "Hacked or not" services, and even more concern for research based on data exposed from breaches. Consent was a consistent theme, with common strains of feedback shown in Figure 6. Here, participants valued their privacy over abstract security benefits, though victims of prior breaches reported higher levels of comfort. A key takeaway is that security professionals and researchers need to articulate how any services or investigations can provide a direct benefit to the victims of breaches given the sensitive nature of the data involved.

Responsibly handling exposed data: Our study provides a perspective of user expectations and concerns with respecting breached data. However, before establishing a line in the sand for best practices of responsibly handling exposed data, it is also vital to consider the views of journalists, security experts, and researchers. We leave building such a broad perspective to future work.

7. CONCLUSION

In this work, we presented the results of two surveys that gauged user comprehension, expectations, and concerns with both responding to breaches and how one handles exposed

data. Our results indicate that data breaches are a highly topical issue in the minds of participants. As such, participants have clear expectations for remediation steps: Breached companies need to be transparent and notify victims; proactively reset passwords and lock down accounts from further damage; and engage with law enforcement to identify the criminals involved.

Zooming out to the wider community, our results show that participants are supportive of emerging security practices including threat sharing between companies in the event of a breach, as well as proactively resetting reused passwords found in password dumps. Other use cases that have less direct or tangible benefits to users, such as research or "hacked or not" services, were viewed less favorably due to overriding privacy concerns. Our findings also help to inform a broader discussion within the community of how to responsibly handle exposed data while respecting concerns around privacy and consent.

8. REFERENCES

[1] L. Ablon, P. Heaton, D. C. Lavery, and S. Romanosky. Consumer attitudes toward data breach notifications and loss of personal information. In *Proceedings of the Workshop on the Economics of Information Security*, 2016.

[2] B. Benko, E. Bursztein, T. Pietraszek, and M. Risher. Cleaning up after password dumps. `https://security.googleblog.com/2014/09/cleaning-up-after-password-dumps.html`, 2014.

[3] Council of the European Union. Notification of a personal data breach to the supervisory authority. `https://gdpr-info.eu/art-33-gdpr/`, 2017.

[4] L. F. Cranor. Giving notice: why privacy policies and security breach notifications aren't enough. *IEEE Communications Magazine*, 43(8):18–19, 2005.

[5] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. Wang. The tangled web of password reuse. In *Proceedings of the Network and Distributed System Security Symposium*, 2014.

[6] X. D. C. De Carnavalet, M. Mannan, et al. From very weak to very strong: Analyzing password-strength meters. In *Proceedings of the Network and Distributed System Security Symposium*, 2014.

[7] J. Franklin, A. Perrig, V. Paxson, and S. Savage. An inquiry into the nature and causes of the wealth of internet miscreants. In *Proceedings of the Conference on Computer and Communications Security*, 2007.

[8] J. Garside. Paradise papers leak reveals secrets of the world elite's hidden wealth. `https://www.theguardian.com/news/2017/nov/05/paradise-papers-leak-reveals-secrets-of-world-elites-hidden-wealth`, 2017.

[9] A. Greenberg. Hack brief: Uber paid off hackers to hide a 57-million user data breach. `https://www.wired.com/story/uber-paid-off-hackers-to-hide-a-57-million-user-data-breach/`, 2017.

[10] S. Gressin. The equifax data breach: What to do. `https://www.consumer.ftc.gov/blog/2017/09/equifax-data-breach-what-do`, 2017.

[11] L. Harding. What are the panama papers? a guide to history's biggest data leak. `https://www.theguardian.com/news/2016/apr/03/what-you-need-to-know-about-the-panama-papers`, 2016.

[12] M. Hiltzik. Did TransUnion Increase Cost Of Credit Monitoring In Wake Of Equifax Breach? `http://beta.latimes.com/business/hiltzik/la-fi-hiltzik-lifelock-equifax-20170918-story.html`, 2017. [Online; accessed 20-January-2018].

[13] T. Hunt. The impact of "Have I been pwned" on the data breach marketplace. `https://www.troyhunt.com/the-impact-of-have-i-been-pwned-on-data/`, 2016.

[14] R. Janakiraman, J. H. Lim, and R. Rishika. The effect of data breach announcement on customer behavior: Evidence from a multichannel retailer. *Journal of Marketing*, 2017.

[15] M. G. Keith and P. D. Harms. Is mechanical turk the answer to our sampling woes? *Industrial and Organizational Psychology*, 9(1):162–167, 2016.

[16] S. Larson. Every single yahoo account was hacked - 3 billion in all. `http://money.cnn.com/2017/10/03/technology/business/yahoo-breach-3-billion-accounts/index.html`, 2017.

[17] S. Larson. Uber's massive hack: What we know. `http://money.cnn.com/2017/11/22/technology/uber-hack-consequences-cover-up/index.html`, 2017.

[18] W. Melicher, B. Ur, S. M. Segreti, S. Komanduri, L. Bauer, N. Christin, and L. F. Cranor. Fast, lean, and accurate: Modeling password guessability using neural networks. In *Proceedings of the USENIX Security Symposium*, 2016.

[19] R. M. Peters. So you've been notified, now what: The problem with current data-breach notification laws. *Ariz. L. Rev.*, 56:1171, 2014.

[20] R. Price. A site that tracked massive hacks has disappeared after being allegedly raided by the cops. `http://www.businessinsider.com/hack-tracking-site-leakedsource-disappears-allegedly-raided-law-enforcement-2017-1`, 2017.

[21] L. Rainie, S. Kiesler, R. Kang, M. Madden, M. Duggan, S. Brown, and L. Dabbish. Anonymity, privacy, and security online. *Pew Research Center*, 2013.

[22] P. M. Schwartz and E. J. Janger. Notification of data security breaches. *Michigan Law Review*, pages 913–984, 2007.

[23] R. Shay, I. Ion, R. W. Reeder, and S. Consolvo. My religious aunt asked why i was trying to sell her viagra: experiences with account hijacking. In *Proceedings of the Conference on Human Factors in Computing Systems*, 2014.

[24] D. R. Thomas, S. Pastrana, A. Hutchings, R. Clayton, and A. R. Beresford. Ethical issues in research using datasets of illicit origin. In *Proceedings of the Internet Measurement Conference*, 2017.

[25] K. Thomas, F. Li, A. Zand, J. Barrett, J. Ranieri, L. Invernizzi, Y. Markov, O. Comanescu, V. Eranti, A. Moscicki, et al. Data breaches, phishing, or malware?: Understanding the risks of stolen credentials. In *Proceedings of the Conference on Computer and Communications Security*, 2017.

[26] K. Thomas and A. Moscicki. New research:

Understanding the root cause of account takeover. https://security.googleblog.com/2017/11/new-research-understanding-root-cause.html, 2017.

[27] A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

[28] D. L. Wheeler. zxcvbn: Low-budget password strength estimation. In *Proceedings of the USENIX Security Symposium*, 2016.

[29] V. Woollaston. Facebook and netflix reset passwords after data breaches. http://www.wired.co.uk/article/facebook-netflix-password-reset, 2016.

[30] K. Zickuhr. Generations and their gadgets. *Pew Internet & American Life Project*, 20, 2011.

Appendix

Study 1: Questionnaire:

Which of the following according to you is a data breach?
o Public exposure of usernames and passwords of millions of users of an online system *[screened in]*
o Web page that is unable to load due to too much data on the page *[screened out]*
o Using large sets of data to aid robots to solve a problem that humans cannot solve *[screened out]*

According to you, what is the most important harm that can happen due to a data breach?
o Spam being sent out from your account / Receiving Spam
o Your computer will be infected by a virus
o Identity theft
o Monetary Loss
o No harm
o Loss of access to personal information
o Leak of personal information
o Your phone will be monitored by hackers
o Other, Please specify

In response to being hacked, which of the following actions should a company take to protect your account? (Choose all that Apply)
o Upgrade your web browser
o Change your username
o Issue a refund
o Send you an immediate notification about the breach
o Enable two-factor authentication
o Provide credit monitoring
o Pay users a consolation bonus for breaking their trust
o Reset your password
o Give you a new account
o Company buys you a new computer
o Other, Please specify

What is the shape of a red ball?
o Red o Blue o Square o Round

How comfortable would you be with a company taking the following actions after the company had a data breach?

ACTION 1: Company that was hacked and experienced the data breach notifies you that your password was stolen
Not at all Comfortable | 1 2 3 4 5 | Very Comfortable
Please explain your rating.
open ended response

ACTION 2: Company that was hacked and experienced the data breach notifies the press about the incident
Not at all Comfortable | 1 2 3 4 5 | Very Comfortable
Please explain your rating.
open ended response

ACTION 3: Company that was hacked and experienced the data breach reports the incident to the government
Not at all Comfortable | 1 2 3 4 5 | Very Comfortable
Please explain your rating
open ended response

ACTION 4: Company that was hacked and experienced the data breach buys a copy of stolen usernames and passwords from the hacker to know what was exposed
Not at all Comfortable | 1 2 3 4 5 | Very Comfortable
Please explain your rating
open ended response

ACTION 5: Company that was hacked and experienced the data breach resets your password

Not at all Comfortable | 1 2 3 4 5 | Very Comfortable
Please explain your rating
open ended response

ACTION 6: Company that was hacked and experienced the data breach shares a copy of all stolen usernames and passwords with other companies to protect your other accounts where you may have reused your username/password
Not at all Comfortable | 1 2 3 4 5 | Very Comfortable
Please explain your rating
open ended response

Have you ever been a victim of data breach?
o Yes o No o Don't know

What is your gender?
o Female o Male o Transgender o I prefer not to answer o Other:

What is your age-group?
o 18-24 years old o 25-34 o 35-44 o 45-54 o 55-64 o 65 or older o I prefer not to answer

Which country do you live in?
drop-down with list of countries

What is the highest degree or level of school that you have completed?
o Professional doctorate (for example, MD, JD, DDS, DVM, LLB) o Doctoral degree (for example, PhD, EdD) o Masters degree (for example, MS, MBA, MEng, MA, MEd, MSW) o Bachelors degree (for example, BS, BA) o Associates degree (for example, AS, AA) o Some college, no degree o Technical/Trade school o Regular high school diploma o GED or alternative credential o Some high school o I prefer not to answer o Other, Please Specify

Which of the following describes your current employment status?
o Employed full-time o Employed part-time o Self-employed o Care-provider o Homemaker o Retired o Student - Undergraduate o Student - Masters o Student - Doctoral o Looking for work / Unemployed o Other, Please specify

What is the color of a red ball?
o Red o Blue o Square o Round

Study 2: Questionnaire:

WELCOME
The goal of this study is to get your feedback on a few scenarios that are detailed in the next set of screens. You will be presented with two hypothetical scenarios. For each of these scenarios, you will be asked to rate your level of comfort.

Your contribution to this study would be of great value to us as we are always looking for ways to improve the experience of internet users like yourself.

The study will take approximately 5-10 minutes to complete.

When you are ready to proceed, please click ■.

Introduction: Global Inc is a leading online social network that provides services such as Email, Chat, Blogs, and Profile pages with over 100 million users using its services everyday. Global Inc just suffered a data breach resulting in hackers gaining access to the data of every user, including their username and password. The hackers responsible for the attack are now selling the stolen data online for a price.

(Randomly show each respondent 2 out of the 8 questions below)

Security Research
John is a researcher from the Southern University investigating online security. John's research focuses on identifying how internet users select passwords, including the most commonly selected passwords. When John hears that he can buy millions of passwords exposed by the Global Inc hack via the online black market, he decides to buy a copy to use for his research.

Q1: Imagine you are one of the users of Global Inc who was affected by the breach. Rate your level of comfort with John buying the hacked data which may contain your credentials too?
Extremely Uncomfortable | Somewhat Uncomfortable | Neither comfortable nor Uncomfortable | Somewhat Comfortable | Extremely Comfortable

Q2: Please explain your rating.
– *open ended response* –

Q3:If the hacked data was publicly available on the internet for free download, rate your level of comfort with John downloading and using the hacked data for his research.
Extremely Uncomfortable | Somewhat Uncomfortable | Neither comfortable nor Uncomfortable | Somewhat Comfortable | Extremely Comfortable

---

Hacked or not service
LMN Tech is an online security service provider. They provide a paid service where anyone can look up their username to see whether it was exposed as part of a major data breach. Their service relies on archives of data collected from a variety of data breach incidents. LMN Tech hears about Global Inc's recent data breach and is planning to buy the hacked data from the hackers to be able to add it to their huge archive of breached datasets.

Q1: Imagine you are one of the users of Global Inc. Rate your level of comfort with LMN Tech's purchase.
Extremely Uncomfortable | Somewhat Uncomfortable | Neither comfortable nor Uncomfortable | Somewhat Comfortable | Extremely Comfortable

Q2: Please explain your rating.
open ended response

Q3: If the hacked data was publicly available on the internet for free download, what will be your level of comfort with LMN Tech downloading and using them to support their service?
Extremely Uncomfortable | Somewhat Uncomfortable | Neither comfortable nor Uncomfortable | Somewhat Comfortable | Extremely Comfortable

---

Threat sharing: Finance
PayPool is an industry leading online payments provider that supports making online purchases. Global Inc knows that its users often logged into other online services such as PayPool, with their Global Inc email address. The Security team at Global Inc suspect that the hackers will use the hacked data to further hack accounts of PayPool users to commit financial fraud. Global Inc believes that sharing a list of hacked email ids that are definitely linked to PayPool accounts will enable PayPool to guard those user's account on PayPool through proactive password resets.

Q1: Imagine you are one of the users of Global Inc and you use your Global Inc email to login to PayPool for online purchases, rate your level of comfort with the security team's plan.
Extremely Uncomfortable | Somewhat Uncomfortable | Neither comfortable nor Uncomfortable | Somewhat Comfortable | Extremely Comfortable

Q2: Please explain your rating.
open ended response

Threat Sharing: Social
LoopedIn is a popular professional networking site where job seekers post their CVs and employers post jobs. Global Inc knows that its users often logged into services such as LoopedIn, with their Global Inc email address. The Security team at Global Inc suspect that the hackers will use the hacked data to further hack accounts of LoopedIn users and may leak their job search information. Global Inc believes that sharing a list of hacked email ids that are definitely linked to LoopedIn accounts will enable LoopedIn guard those user's account on LoopedIn through proactive password resets.

Q1: Imagine you are one of the users of Global Inc and you use your Global Inc email to login to LoopedIn for professional networking. Rate you level of comfort with the security team's plan.
Extremely Uncomfortable | Somewhat Uncomfortable | Neither comfortable nor Uncomfortable | Somewhat Comfortable | Extremely Comfortable

Q2: Please explain your rating.
open ended response

Proactive Password Reset
Doodle is a large internet company with billions of users. Many of Doodle's users reuse their passwords on third party services. In an attempt to proactively protect users from someone breaking into their account, Doodle regularly buys hacked datasets that are sold on the black market to scan and re-secure accounts of users who have had their password exposed on other third party services.

Q1: Imagine you are a Doodle user. Rate your level of comfort with Doodle‘s proactive security measure.
Extremely Uncomfortable | Somewhat Uncomfortable | Neither comfortable nor Uncomfortable | Somewhat Comfortable | Extremely Comfortable

Q2: Please explain your rating.
open ended response

Q3: If the hacked data was publicly available on the internet for free download, rate your level of comfort with Doodle downloading and using them as a proactive security measure?
Extremely Uncomfortable | Somewhat Uncomfortable | Neither comfortable nor Uncomfortable | Somewhat Comfortable | Extremely Comfortable

---

Journalist, tax fraud
Jerry is a prominent journalist working at That‘s Correct Media and Publishing company. After Global Inc‘s hack Jerry buys the hacked data via the online blackmarket and gets access to personal emails of some of Global Inc‘s users. These emails reveal a major public scam involving several public officials using offshore financial centers to avoid taxes. Jerry publishes a news article to disclose the tax avoidance scam, using the hacked email as proof.

Q1: Imagine you are one of the users of Global Inc who was affected by the breach. Rate your level of comfort with John buying the hacked data which may contain your credentials too?
Extremely Uncomfortable | Somewhat Uncomfortable | Neither comfortable nor Uncomfortable | Somewhat Comfortable | Extremely Comfortable

Q2: Please explain your rating.
open ended response

Q3: If the hacked data was publicly available on the internet for free download, rate your level of comfort with John downloading and using the hacked data for his research.
Extremely Uncomfortable | Somewhat Uncomfortable | Neither comfortable nor Uncomfortable | Somewhat Comfortable | Extremely Comfortable

---

Journalist, dating site
Mark a journalist at TownNews Today learns about a recent data breach of a dating site GoDate.com. He decides to purchase the hacked data to look up names of people from his town and publish information about their private dating profiles. Mark feels that the profiles would make up interesting subject matter for his articles and is planning to publish some articles based on these profiles.

Q1: Imagine you are reading JerryâĂŹs article, rate your level of comfort with JerryâĂŹs source of proof.
Extremely Uncomfortable | Somewhat Uncomfortable | Neither comfortable nor Uncomfortable | Somewhat Comfortable | Extremely Comfortable

Q2: Please explain your rating.
open ended response

Q3: If the hacked data was publicly available on the internet for free download, rate your level of comfort with Jerry downloading and using it as a source of proof for his article.
Extremely Uncomfortable | Somewhat Uncomfortable | Neither comfortable nor Uncomfortable | Somewhat Comfortable | Extremely Comfortable

---

Competitor
Moon Inc is a competitor of Global Inc. After hearing about Global Inc‘s data breach, the marketing team at Moon Inc plans to buy the hacked data via the online blackmarket to learn about Global Inc‘s users who were hacked. The marketing team plans to target these Global Inc users with an offer to switch services.

Q1: Imagine you are one of the users of Global Inc who was affected by the breach. Rate your level of comfort with being approached by Moon Inc offering you to switch to their service.
Extremely Uncomfortable | Somewhat Uncomfortable | Neither comfortable nor Uncomfortable | Somewhat Comfortable | Extremely Comfortable

Q2: Please explain your rating.
open ended response

Q3: If the hacked data was publicly available on the internet for free download, rate your level of comfort with Moon Inc downloading and using them to approach hacked Global Inc users to switch to their service.
Extremely Uncomfortable | Somewhat Uncomfortable | Neither comfortable nor Uncomfortable | Somewhat Comfortable | Extremely Comfortable

---

Please specify your gender: o Male o Female o Other o Prefer not to say

Please select your age group: o 18-24 o 25-34 o 35-44 o 45-54 o 55-64 o 65+ o Prefer not to say

What is the highest degree or level of school that you have completed? o Less than high school o High school graduate (includes equivalency) o Some college, no degree o Associate‘s degree o Bachelor‘s degree o Master‘s degree o Ph.D. o Other, Please specify

Please specify your current employment status o Employed Full-time o Employed Part-time o Self-employed o Care-provider o Homemaker o Retired o Student - Undergraduate o Student - Masters o Student - Doctoral o Looking for work / Unemployed o Other, Please Specify

Have you ever been a victim of data breach? o Definitely yes o Probably yes o Might or might not o Probably not o Definitely not

---

User Comfort with Android Background Resource Accesses in Different Contexts

Daniel Votipka, Seth M. Rabin, Kristopher Micinski*, Thomas Gilray,
Michelle M. Mazurek, and Jeffrey S. Foster
University of Maryland, *Haverford College
dvotipka,srabin,tgilray,mmazurek,jfoster@cs.umd.edu; *kmicinski@haverford.edu

ABSTRACT

Android apps ask users to allow or deny access to sensitive resources the first time the app needs them. Prior work has shown that users decide whether to grant these requests based on the context. In this work, we investigate user comfort level with resource accesses that happen in a *background* context, meaning they occur when there is no visual indication of a resource use. For example, accessing the device location after a related button click would be considered an interactive access, and accessing location whenever it changes would be considered a background access. We conducted a 2,198-participant fractional-factorial vignette study, showing each participant a resource-access scenario in one of two mock apps, varying what event triggers the access (*when*) and how the collected data is used (*why*). Our results show that both *when* and *why* a resource is accessed are important to users' comfort. In particular, we identify multiple meaningfully different classes of accesses for each these factors, showing that not all background accesses are regarded equally. Based on these results, we make recommendations for how designers of mobile-privacy systems can take these nuanced distinctions into account.

1. INTRODUCTION

Android apps potentially have access to a range of sensitive resources, such as location, contacts, and SMS messages. As a result, Android and similar systems face a critical privacy and usability trade-off: when should the system ask the user to authorize an app to access sensitive resources? Requesting permissions too often can overburden the user; requesting permission too infrequently can lead to security violations.

There has been significant research into this question, much of which shows that users' access-control decisions depend on the context, including when and why the access attempt is made [7, 27, 30, 38, 43, 45]. However, this prior work has typically focused on individual aspects of context in isolation, such as app behavior at the point of resource-access [30, 45, 46], or the reason the app requires access to the sensitive resource [27]. In particular, much of this work relies on a binary distinction between foreground and background accesses—sometimes defined as whether the app is visible on the screen [45,46], and sometimes defined as whether the resource access is explicitly triggered by a specific user interaction [30, 36]. (Section 2 discusses related work in more detail.)

In this paper, we investigate more deeply how users understand resource uses that occur *in the background*, which we broadly define as not explicitly and obviously caused by a user interaction. We examine whether different kinds of background uses are viewed similarly, or whether more fine-grained distinctions are required for user comprehension.

In our investigation, we consider a broad range of possible background accesses, drawn in part from existing literature and in part from reverse-engineering the behavior of popular apps. We examine the context of these background accesses along two key axes: *when* and *why* the resource access is triggered. We consider four cases for *when*: after an *Interaction* (this is a non-background case, as a control), due to *Prefetching*, by a *Change* to a resource such as the device's location changing, or after an unrelated UI action, which we refer to as a *UI Background* access. We also consider five cases for *why*: to *Personalize* the app, to get data from an app *Server*, to support *Analytics* to improve the app, to provide *Ads*, or for no given reason (*NA*). We consider these cases for a mock *Dating* app and a mock *Ride Sharing* app, and for three sensitive resources: *Location*, *Contacts*, and *SMS* messages.

We performed a 2,198-participant, between-subjects online vignette survey investigating users' comfort across 52 conditions selected from the full-factorial set. Each participant viewed a slideshow of a mock app being used and then a diagram illustrating *when* and *why* the app accessed a selected sensitive resource. The participant was then asked whether they would be comfortable using an app that behaved similarly and whether they would recommend such an app to friends. (Section 3 describes our methodology, and Section 4 reports participant demographics.)

We found that both the *why* and *when* aspects of context played a significant role in users' expressed comfort with background accesses. Background accesses that shared data with third parties for advertising and analytics were more objectionable than accesses providing personalized features, even when data was sent off device to the app developer. Perhaps unsurprisingly, of the third-party accesses, those

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

associated with with advertising were the least acceptable. Additionally, if no reason for access is provided, participants were likely to assume that the data is accessed for personalization. However, perhaps due to uncertainty about whether this assumption is valid, participants were less comfortable in this case than when personalization was explicitly specified.

While all background accesses were viewed as less acceptable than interactive accesses, participants did not react to all background accesses equally. Participants were significantly more comfortable with background accesses when the app is on-screen than when it is off-screen, even when the resource access is not clearly tied to the app's UI. (Section 5 discusses our results.)

Based on these results, we make several design recommendations: that apps explicitly differentiate uses in different contexts, that systems provide better incentives to explain benign uses to users (e.g., when the data is used only for personalization), and that privacy policies track not just *when* data is used, but also where it flows (to explain *why*). (Section 6 presents our design recommendations.)

2. RELATED WORK

In early versions of Android, users were asked to authorize permissions whenever a new app was installed. Multiple studies showed that users did not understand the privacy risks associated with permissions under this model. Felt et al. found that only 17% of their study participants paid attention to the permissions they granted and just 3% fully understood what those permissions could be used to access [15]. Kelley et al. showed that users found the terms and wording of Android permissions hard to understand [22]. Additionally, other researchers demonstrated that users were unable to make informed decisions on whether to install an app because they did not know the context of the resource use, instead relying on their expectations of the app's behavior [5, 25, 38, 43].

Android M [16] and later versions prompt users to grant or deny access to a permission the first time it is required by the app. This model is commonly referred to as Ask-On-First-Use (AOFU). Andriotis et al. found that users feel they have more control of their privacy with AOFU [1, 2]. Bonne et al. showed that users commonly deny a permission and subsequently run the app to determine whether it is truly required [7].

Unfortunately, further work has shown that even under AOFU, users are still not provided with enough context to make informed decisions [7, 30, 38, 43]. In some cases, AOFU may lead the user to make incorrect decisions due to broken assumptions [30]. On the other hand, users experience warning fatigue when presented with too many permission dialogs [6]. Multiple researchers have shown that including a permission's purpose (e.g., feature personalization, advertising) has a significant effect on user comfort [5, 26–28, 38, 41]. Of this prior work, the study by Lin et al. [27] is most similar to ours. In their study, participants were told that a popular app accesses a specific sensitive resource, and participants were given the purpose of that access. Lin et al. collected comfort ratings from 725 MTurkers for 1,200 different combinations of 837 apps, 6 resources, and 4 purposes (participants could provide responses for multiple combinations). They found that the purpose shown had a significant effect on user comfort. We build on their study of user comfort by testing additional purposes, and we compare each purpose to the case where none is given to determine the effect of not informing the user. Also, we add additional variation in our conditions by testing both *why* and *when* the access occurs to study the relative strength of their effects and determine whether there is some interaction between these variables.

Other work has sought to study how user comfort is affected by the timing of resource uses (e.g., after a button is clicked, whenever the resource changes) [30, 43, 45, 46]. Wijesekera et al. study users *in situ*, measuring the effect of the app, whether the app is on screen, and the resource on whether users grant or deny resource access [45, 46]. Wijesekera et al. found that users were more likely to grant access when the request occurred whenever the app was being used. Our prior work studied what resources users expect apps to access as they interact with the app (i.e., on startup, after a button is clicked, when no interaction is shown) [30]. They find that users expect resources to be accessed directly after a related interaction (i.e., camera is accessed after pressing a button labeled "Take a picture"), but do not always expect accesses that are not tied to an interaction. We expand on these findings by investigating comfort with the latter category of accesses.

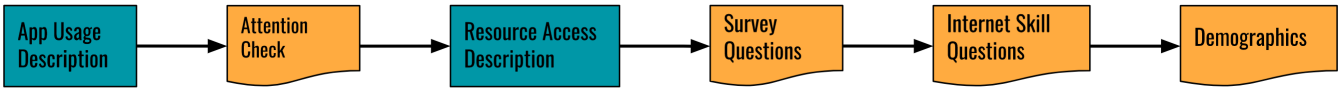
Finally, there has been extensive work on Android permissions more broadly. User comfort has been studied in the context of app recommendation systems [27,27,28,50], which use algorithms to help recommend apps to users based on their privacy preferences. Context has been used to drive static analyses and measure app behavior [9, 12, 19, 47–49]. Lastly, Roesner et al. [36, 37] present Access Control Gadgets, which allows specific buttons in an app to authorize access to specific permissions.

3. METHODOLOGY

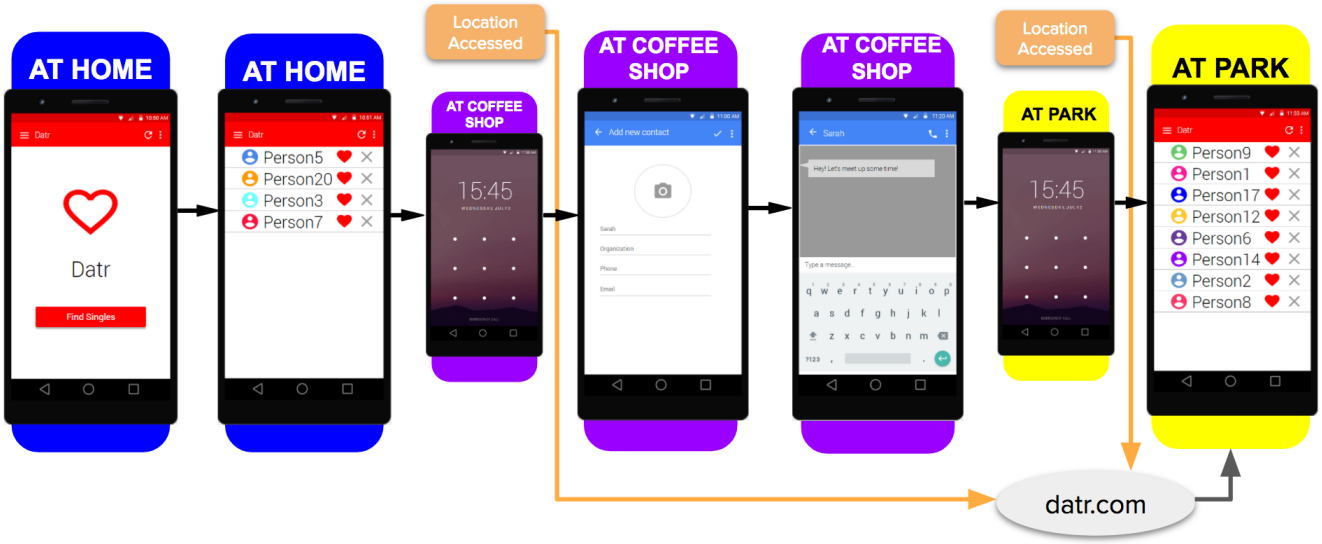
Our study focuses on *background* resource accesses, by which we mean accesses with no obvious, immediate triggering action by the user. For example, accessing device location whenever it changes is a background use because it occurs without the user's direct, immediate intervention. Whereas accessing device location after the user clicks a button is a foreground or *interactive* access. We describe the exact background usage scenarios we study in Section 3.2. From Nissenbaum's theory of Privacy as Contextual Integrity [31] and prior work [5, 26–28, 30, 38, 41, 43, 45, 46], we expect that users' comfort should be significantly affected by the context of an access, including whether it is in the background.

There are potentially many different kinds of background accesses. To determine which accesses to study, we reviewed prior work on common app behaviors [5, 25, 27, 33, 35, 38]. We also manually reverse engineered a small set of Android apps and investigated their background access patterns. In particular, we selected 20 popular apps from our prior analysis that we identified as having background resource accesses [30]. For each app, we used our tool, AppTracer, to locate those background accesses. We then manually examined the app's code, as decompiled with JEB [39], to understand the background access patterns.

Based on this analysis, we decided to study two dimensions of background accesses: *when* the event is triggered and *why*



(a) Survey procedure. Blue rectangles represent description portions of the survey. Orange, curved boxes represent question portions.



(b) Sample vignette for Datr app. In the app usage description, the orange boxes and arrows, and the gray circle, are not shown. They are added in in the resource access description step, along with the follow textual description: "While Jane was using Datr, the app behaved in the following way: Whenever Jane's location changed, Datr learned about the change to her location and sent her updated location to datr.com. datr.com then used her updated location along with other updates it had collected on Jane previously to create a list of recommended singles based on places she has traveled in the past."

Figure 1: User study survey procedure and sample vignette.

the data is accessed. For simplicity, we refer to *when* and *why* as the *access context*. As an example, consider an app that accesses device location every time the device moves and sends this data to a third party advertiser. The *when* is changing device location, and the *why* is advertising.

3.1 Study Overview

We performed a between-subjects, fractional-factorial vignette study [4]. Participants were recruited from the Amazon Mechanical Turk crowd-sourcing service. All participants were at least 18 years old and located in the United States. After completing the vignette study, participants were paid $1.20. Participants took on average 4 minutes and 46 seconds to complete the survey. This study was approved by our organization's ethics review board. Participants were asked for their opinions regarding a given app's functionality and behavior, but we did not explicitly mention privacy or the possible sensitivity of specific resources.

Before beginning the main study, we piloted the survey with nine participants selected from a convenience sample, chosen in part for varying levels of technical knowledge. For each pilot, we asked participants to "think aloud" as they read the prompts and answered each question. We iteratively updated our survey following each pilot, eventually reaching the final instrument detailed below.

Figure 1a describes the survey procedure. First, the participant is shown a short description of the app. We used two mock apps in our study: Datr (*Dating*) and Ridr (*Ride Sharing*). Datr presents users with a list of singles they might like to meet, similar to popular dating apps like Tinder and Bumble. Ridr allows users to request rides to a selected destination, similar to popular ride sharing apps like Uber and Lyft. We do not attempt to fully study the effect of app type on user comfort, but instead simply include two apps from different categories to provide some insight into whether an effect may exist.

We begin each survey by describing how Jane, a fictional character, might use the app. We do so by showing a sequence of screenshots in which Jane first uses the app at home; then travels to a coffee shop and uses other apps; and then travels to the park and uses the app again. Figure 1b shows an example. Note that in this first step, the orange boxes and gray circle in the figure are omitted from the vignette. In this example, Jane opens the Datr app at home, presses the "Find Singles" button, and sees a list of recommended singles. Then Jane travels to a coffee shop to meet a friend. While at the coffee shop, she adds the friend's contact information to her address book and receives a text message. Finally, Jane travels to a park and re-opens Datr, which shows a new list of recommended singles. Vignettes for Ridr are similar, except Jane presses "Request Ride" and is shown a list of recommended destinations.

After viewing the description of the app's use, participants answer a simple question about the app's functionality as an attention check (second item in Figure 1a). We include this check to reduce the risk of invalid responses.

Next, participants are informed of the app's "behind-the-scenes" access context where they are shown the same series of app screens with additional indicators showing *when* and *why* the resource access occurred. Figure 1b shows that Datr collects Jane's location whenever it changes (i.e., *when* she goes to the coffee shop and later to the park) and sends it to `datr.com`. The screenshots are also accompanied by a textual description, listed in the caption of Figure 1b, explaining the scenario. The server then returns a list of recommended singles, to be displayed the next time she opens the app, based on Jane's location. For each vignette, the set and order of app pages shown is the same, but we vary the *when* and the *why* (i.e., the orange boxes, gray circles, arrows, and text explanations).

For all scenarios where information is sent to a server, we show an arrow from the app to a circle labeled with a suggestive domain name. Kang et al. found that this was the most common convention used by non-technical users to draw information transmitted over a network [21].

After describing the app's access context, we ask the participants a series of five-point Likert-scale questions.[1] We begin by asking whether participants believe the access context is likely ("Very unlikely" to "Very likely") to appear in popular apps and whether they agree ("Disagree" to "Agree") the behavior makes the app more useful. We ask these questions because we expect that participants' comfort with an access context is likely affected by their prior exposure to similar scenarios and perceived usefulness of the behavior to the user [32, pg. 133-140].

Next, we directly assess the participant's level of comfort by asking whether they would feel "Very uncomfortable" to "Very comfortable" using an app with the described access context. Additionally, we ask whether they would be "Very unlikely" to "Very likely" to recommend an app with the described access context to a friend who is looking for an app with the given functionality. We add this question to indirectly measure participant comfort.

Some participants were assigned to a condition in which the *why* is not given. In these conditions, we ask participants to provide a short, open response description of why they think the access occurred.

Finally, we conclude with a set of questions about the participants' Internet skill level and demographics. We measure Internet skill using the seven-question scale proposed by Hargittai and Hsieh [17]. Each question on the scale asks participants to rate their familiarity with a different Internet-related term, from "No understanding" to "Full understanding."

3.2 Conditions and Hypotheses

Within this study design, we developed a set of conditions varying over four variables: the *app*, the *resource* being accessed, *why* the app accessed the resource, and *when* the resource was accessed. Table 1 lists the levels for each variable. Conditions consist of one level from each column. As detailed below, we selected a subset of possible combinations to arrive at a final set of 52 conditions, which were assigned round-robin to participants. The condition levels we selected for *when* and *why* map directly to the hypotheses we investigate.

App	Resource[1]	Why[2]	When[3]
Dating	*Loc*	*Personalize*	*Int* [5,6]
Ride Sharing	*Con*	*Server*	*Pre* [5,6]
	SMS [5]	*Analytics* [5,6,7]	*UI-Bg*
		Ads	*Change*
		NA [4,5,6,7]	

[1] *Con* - Contacts, *Loc* - Location, *SMS* - SMS
[2] *Ads* - Advertising, *Analytics* - Debugging/Analytics, *Server* - Server, *Personalize* - Personalize, *NA* - Not Given
[3] *Change* - On Change, *UI-Bg* - UI Background, *Int* - Interactive, *Pre* - Prefetch
[4] Never used with *Int*, and *Pre*
[5] Never used with *Ride Sharing*
[6] Never used with *SMS*
[7] Never used with *Int*

Table 1: Possible values for each variable in tested conditions.

Reasons for resource access. We used five variations for *why* the app collected the sensitive resource. In the personalize (*Personalize*) case, users were told the app collected data to provide personalized features. Additionally, this case stated that no data was sent off device (i.e., to the app's server or any third party). Server (*Server*) is similar, but users were told data was first sent to the app's own server to support personalization. For example, the *Dating* app sends the user's information to the server to retrieve a list of personalized dating matches. Debugging/Analytics (*Analytics*) stated that the app shared data with a third-party for debugging crashes and collecting analytics to improve the app. For Advertising (*Ads*), participants were told the app sent their collected data to a third-party advertiser to improve ad targeting.

From these variations, it can be seen that this context variable also implicitly includes a *who* component. For simplicity, we only consider the general function of the *who* data is shared with and use generic domain names (e.g., *ads.com* for *Ads*). An investigation of the effect of a specific advertisement or analytics provider on user comfort is beyond this scope of this paper.

We also include a Not Given (*NA*) case, in which the participant was not given a reason for data collection.

These scenarios map to our first two hypotheses, as follows.

H1. The provided reason for resource access affects participants' comfort levels.

Within this broad hypothesis, we test three sub-hypotheses concerning specific categories of possible reasons for resource access.

H1a. Users are more comfortable if their information is kept on their device.

[1]The exact wording for each question is in Appendix A.

H1b. Users are more comfortable if their information is not shared with a third party.

H1c. Users are more comfortable if their information is only shared with a third party to improve general app functionality, as opposed to advertising.

Notice that each of these hypotheses represents an increasing degree of willingness to share information.

We test *H1a–c* by searching for divergence among our *why* levels. If *H1a* is true, then we would expect to see a gap in comfort between *Personalize* and *Server*. Similarly, *H1b* indicates a divide between first party (i.e., *Personalize* and *Pre*) and third party (i.e., *Analytics* and *Ads*) sharing. Finally, *H1c* is true if there is a significant difference in comfort between *Ads* and the other levels.

H2. Users are more comfortable if given a reason for background use.

While *H1* investigates how different reasons for resource access compare in terms of user comfort, *H2* asks how these different explanations compare to the lack of an explanation. Prior work in psychology has shown that people are generally more accommodating when given a reason for a request, no matter how vague [24]. *H2* tests whether this is the case for background resource accesses. If *H2* does not hold, then perhaps in some cases the reason for access can be omitted from access notifications or requests, reducing cognitive burden on users without causing undue discomfort.

Triggers for resource access. We considered five variations in *when* the app requests resources. UI-Interactive (*Int*) describes the case where an app accesses a resource after a directly related UI event (e.g., a button click). We include this case, which is not a background access, as a control that mimics the interactive resource use patterns described in our previous work, which led users to expect resource accesses [30].

Prefetch (*Pre*) is similar to *Int*, as the UI indicates that the resource is accessed. However, the actual resource access occurs prior to the UI event (on startup of the application), so that the accessed data is ready to present when the user performs the UI action. In the *Pre* case, there is no visual indication of access when the data is collected, but the user is eventually made aware. UI-Background (*UI-Bg*) presents the same behavior as *Int*—access after a UI event—except the UI event is unrelated to the resource access. Finally, On Change (*Change*) describes an app that accesses a sensitive resource directly after that resource has been modified (e.g., the user changes location or adds/deletes a contact). Note that a *Change* access—unlike *UI-Bg* and *Int*—can occur whether the app is or is not currently in use.

These variations map to our final hypothesis:

H3. Users have different comfort levels when resource accesses are triggered by different events.

Prior work has considered two dichotomous categorizations of access triggers: On-screen vs. off-screen [45] and interactive vs. non-interactive [30]. We use the following sub-hypotheses to understand user comfort across and between these categorizations, with more fine-grained distinctions.

H3a. Users are more comfortable with sensitive resource accesses when they are interactive.

H3b. Users are more comfortable with sensitive resource accesses when there is an explicit foreground visual indicator of use, even if the use occurs before the indicator.

H3c. In the absence of an explicit foreground visual indication of use, users are more comfortable when the app is on-screen than when it is off-screen.

To examine *H3a*, we compare *Int* to all the other levels. To examine *H3b*, we compare *Pre* to the other background levels. Finally, to examine *H3c* we compare *UI-Bg* to *Change*.

Apps and resources. As stated previously, we use two mock apps, Datr (*Dating*) and Ridr (*Ride Sharing*). We selected three resources which we found in our prior work to be used with both foreground and background interaction patterns: Location (*Loc*), Contacts (*Con*), and Text Messages (*SMS*) [30]. We test multiple resources because prior work has shown that grant/deny rates varied between permission types [7].

Final condition set. Because the full-factorial combination of all levels of each variable creates too many conditions to be feasibly tested, we discarded combinations that were redundant, logically inappropriate, or less relevant to our hypotheses. After this reduction, we were left with 52 final conditions.

First, we removed any condition that includes *Int* or *Pre* together with *NA*. Since a reason for the resource access is directly presented to the user through the UI (i.e., the button text clearly states that the resource is accessed to provide personalization of a feature), *Int-NA* and *Pre-NA* are redundant with *Int-Personalize* and *Pre-Personalize*, respectively. Therefore, *NA* is only included with *UI-Bg* and *Change*. This is shown in Table 1 by the orange highlight of *NA* and indicated by the superscript 4.

As we do not intend to completely investigate the effect of app type and resource, we restrict the *Ride Sharing* and *SMS* conditions to only include levels where we expect to observe the largest variation in comfort. Specifically, with *Ride Sharing*, we do not test the resource *SMS*; the *why* levels *Analytics* and *NA*; and the *when* levels *Int* and *Pre*. In Table 1 all the highlighted levels are never considered with *Ride Sharing* as indicated by the superscript 5.

For *SMS*, we do not test the *why* levels *Analytics* and *NA* or the *when* levels *Int* and *Pre*. The levels that are never associated with *SMS* are highlighted in blue, orange, and yellow and indicated by the superscript 6.

Finally, due to the similarity in presentation between *Pre* and *Int*, we limit the levels included with *Int* to only those where we expect to observe the largest variation in comfort. Therefore, we do not consider *Analytics* with *Int*. In Table 1, we highlight in blue—and indicate with the superscript 7—the levels that are never included with *Int* due to this rule.

3.3 Statistical Analysis

For all Likert-scale questions, we use an ordered logistic regression (appropriate for ordinal data) [29] to estimate the

effect of the assigned condition on the participant's comfort, likelihood to recommend the app to others, perceived usefulness of the app behavior, and perceived likelihood that this behavior occurs in popular apps.

For each question, our initial regression model included all the factors and interactions given in Table 2. We applied the standard technique of centering the numerical factor (Internet skill) around its mean before analysis to promote interpretability [10]. To determine the optimal model, we calculated the Bayseian Information Criterion (BIC)—a standard metric for model fit [34]—on all possible combinations of the given factors. To avoid overfitting, we selected the model with the minimum BIC. This process was completed for each regression separately.

Additionally, to understand what participants believed about the reason for data collection when none was explicitly given (i.e., the *NA* level of the *when* factor), we performed an open coding of participants' free responses. Two researchers individually reviewed each response in sets of 30 and iteratively developed the codebook. The coders reached a Krippendorff's α of 0.831 after three rounds of pair coding (i.e., 90 responses), which is within the recommended bounds for coding agreement [18]. The remaining responses were divided evenly and each coded by a single researcher.

3.4 Limitations

Our reliance on mock apps for our controlled experiment limits the ecological validity of our study. We chose this setting because it allows us to reason about the statistical effect of specific factors on participant comfort. Additionally, using mock apps allows us to disregard possible confounding factors such as participants' prior experience with an app or its developer's reputation. However, in this controlled setting, users may be less concerned about their privacy than if their real data were at risk. They may also overstate their discomfort because they are not actually using the app and therefore not placing emphasis on the functionality benefits gained by allowing access to their personal data [40]. To partially account for this, we ask about comfort both directly and indirectly (i.e., would they recommend the app to a friend) and include a description of the app functionality that is dependent on the given access context. Additionally, we only rely on comparative, rather than absolute, results when analyzing responses.

Limiting our study to two types of apps and restricting the resources and access contexts tested is likely to cause us to miss potential factors, especially interactions between factors that affect user comfort. For example, users are likely to expect different types of apps to use resources differently depending on the app's functionality, and these differences in expectations are likely to affect comfort. In an attempt to reduce this problem, we selected conditions based on a review of prior work and manual app reverse engineering.

For each finding from our open-response questions, we report the percentage of participants that expressed a concept. However, a participant not mentioning a specific idea does not necessarily indicate disagreement. Instead, they may have simply failed to state it, or they may not have thought it the most likely possibility. Therefore, our results from open-response questions should be interpreted as measuring what was at the front of participants' thoughts as they responded to the questions.

Since we ask participants to consider app behaviors that occur in the background, it is possible that participants may not completely understand the scenario. However, we attempted to mitigate this issue by using diagrams similar to those drawn by non-technical users to represent network communication [4]. During our pilot interviews, we specifically asked participants to describe what was occurring in the displayed scenario to ensure comprehension, and revised the diagrams accordingly. Finally, all participants who were not shown a reason for the resource access (i.e.,*NA* level of the *why* variable) were asked to state why they thought the app accessed the resource. We did not observe any responses indicating participants misunderstood the scenario.

As is common for any online studies and self-reported data, it is possible that some participants do not approach the survey seriously, and some may try to make multiple attempts at the survey. We limit repeat attempts by collecting participants' MTurk ID and compare these to future attempts to restrict access. Though MTurk has been found to produce high-quality data generally [8, 11, 23, 44], the U.S. MTurker population, from which we drew participants, is slightly younger and more male, tech-savvy, and privacy-sensitive than the general population [20]. This restricted population may affect the generalizability of our results.

However, we consider comparisons between conditions to be valid because each of these limitations apply similarly across all conditions.

4. PARTICIPANT DEMOGRAPHICS

A total of 2,797 participants attempted our survey. Of these, 2,328 (83.2%) finished. From these, we removed several participants who had previously taken the survey. We also removed 121 participants (5.2%) who failed an attention check. We ultimately had 2,198 total responses, with between 40 and 45 responses per condition.

Demographics for our participants are summarized in Table 3. Participants were more male and more white than the U.S. population, as is expected from MTurk. Additionally, our participants' average Internet skill of 32.2 was slightly higher than the mean score of 30.5 recorded by Hargittai and Hsieh on a more general population several years ago [17]. The vast majority of participants use smartphones regularly. The proportion of accepted participants who own a smartphone (99%) is well above the reported U.S. average of 79% reported by Pew [42]. The majority of participants (97%) also considered themselves to have at least "Average" smartphone expertise on a five-point scale from "Far below average" to "Far above average."

5. RESULTS

In our online vignette study, we found that both *why* and *when* resource accesses occurred had a significant effect on user comfort. Additionally, we found that there are several meaningful classes of accesses for each part of the access context.

With respect to *why* the access occurred, we observed that users were more comfortable when data was shared with the app developer (*Personalize* and *Server*) than a third-party (*Analytics* and *Ads*). Further, within third-party sharing, users are more comfortable when data is shared for app an-

Factor	Description	Baseline
When	The context regarding when the sensitive data is accessed	*Int*
Why	The reason the app collected the sensitive data	*NA*
App type	The type of app displayed in the vignette	*Dating*
Resource	The sensitive resource accessed in the vignette	*Loc*
Internet skill	Participant's score on Hargittai and Hsieh's Internet skill scale [17]	0
Smartphone Use	Time per day using a smartphone	0-3 hrs/day
Resource:When	The interaction between the Resource and When variables	*Loc*:*Int*
Resource:Why	The interaction between the Resource and Why variables	*Loc*:*NA*
When:Why	The interaction between the When and Why variables	*Int*:*NA*

Table 2: Factors used in regression models. We compared categorical variables individually to the given baseline. Candidate models were defined using all possible combinations of factors. The final model was selected by minimum BIC.

Metric	Percent
Gender	
Male	54
Female	46
Education	
B.S. or above	49
Some college	39
H.S. or below	13
Age	
18-29 years	34
30-49 years	55
50-64 years	9
65+ years	1

Metric	Percent
Ethnicity	
Caucasian	78
African Am.	10
Asian	1
Hispanic	7
Smartphone Use	
9+	10
6-9	13
3-6	38
0-3	39
No smartphone	<1

Table 3: Participant demographics. Percentages may not add to 100% because we do not include "Other" or "Prefer not to answer" percentages for brevity and selection of multiple options was possible for some questions (i.e., ethnicity).

alytics (to improve the functionality of the application) as opposed to sharing data for advertising. Additionally, if no reason for access was provided, we found that users were less comfortable than they would be if told the data never left their device (*Personalize*), but slightly more comfortable than having their data shared with advertisers (*Ads*).

For *when*, as expected, users are the most comfortable when accesses occur interactively, directly after a UI event (*Int*). Non-interactive (background) accesses can further be divided into two classes: participants were more comfortable if the access occurred when the app was on-screen (*Pre* and *UI-Bg*) compared to off- screen (*Change*). Detailed descriptions of these results are given below.

Interpreting regression results. The majority of our key findings are drawn from our regression analysis over the users' comfort (Table 4a). We also discuss regression analyses for willingness to recommend an app with a given behavior (Table 4b) and belief that the app's behavior is useful (Table 5). Overall, these regressions produced very similar significance results. Our discussion will therefore focus primarily on comfort results.

All three regression tables show (as groups of rows) the variables included in the final selected model. For each categorical variable, we present the base case first. We selected base cases that we expected to produce the highest levels of comfort. For *why*, we selected *Personalize* because it involves the least data sharing. For *when*, we selected *Int* because it is the most interactive, which has been shown to correlate with user expectation of resource access [30]. For resource, we selected *Loc* based on prior work that suggests users are more comfortable with apps accessing location than other sensitive resources [14, 27].

In the odds ratio (OR) column, we show the variable's observed effect. For categorical variables, the OR is the odds of comfort increasing one unit on our Likert scale when changing from the base case to the given parameter level. For the numeric variable (Internet skill), the OR represents the odds of comfort increasing one unit on our Likert scale, per one-point increase in Internet skill. The OR for the base case (categorical) and the average Internet skill (numeric) is definitionally 1.0. For each value, we also give the 95% confidence interval for the odds ratio (CI) and the associated p-value.

As an example, the odds ratio for *Pre* in Table 4a indicates that a user who is assigned to *Pre* rather than *Int*—assuming all other variables are the same—would lead to a 0.64× likelihood of increasing one unit in comfort. Because this effect is less than one, participants are less likely to report higher comfort levels for *Pre* than *Int*. In short, users are less comfortable with *Pre*. Furthermore, *Pre*'s CI indicates that the "true" odds ratio is between 0.48 and 0.87 with 95% confidence. The p-value of 0.004 is less than our significance threshold of 0.05, so we consider this difference between *Int* and *Pre* to be significant.

5.1 H1 and H2: Reasons for resource access

For *H1* and *H2*, we primarily focus on the *why* variable, shown in the first section of Tables 4a and 4b.

Data leaving the device (H1a). We first consider whether resource accesses in which data remains on the device (*Personalize*) are more comfortable for users than those in which data is transferred to the app company's server (*Server*). The first two rows of each Tables 4a and 4b indicate that *Personalize* and *Server* are not significantly different from each other. This is illustrated in Figure 2, which shows participants' Likert responses to the main comfort question, grouped according to the *why* scenario they were shown. In the *Personalize* condition, 44% selected comfortable or very comfortable, compared to 42% in the *Server* condition.

Variable	Value	Odds Ratio	CI	p-value
Why	*Personalize*	–	–	–
	Server	0.88	[0.72, 1.09]	0.240
	Analytics	**0.49**	**[0.37, 0.64]**	**< 0.001***
	Ads	**0.34**	**[0.28, 0.42]**	**< 0.001***
	NA	**0.58**	**[0.43, 0.80]**	**< 0.001***
When	*Int*	–	–	–
	Pre	**0.64**	**[0.48, 0.87]**	**0.004***
	UI-Bg	**0.72**	**[0.55, 0.94]**	**0.014***
	Change	**0.34**	**[0.26, 0.44]**	**< 0.001***
Resource	*Loc*	–	–	–
	Con	**0.33**	**[0.28, 0.39]**	**< 0.001***
	SMS	**0.12**	**[0.09, 0.16]**	**< 0.001***
Internet Skill	0	–	–	–
	+1	**0.95**	**[0.94, 0.97]**	**< 0.001***

*Significant effect – Base case (OR=1, by definition)

(a) Comfort

Variable	Value	Odds Ratio	CI	p-value
Why	*Personalize*	–	–	–
	Server	0.91	[0.74, 1.11]	0.351
	Analytics	**0.51**	**[0.39, 0.67]**	**< 0.001***
	Ads	**0.27**	**[0.22, 0.33]**	**< 0.001***
	NA	**0.58**	**[0.43, 0.80]**	**< 0.001***
When	*Int*	–	–	–
	Pre	**0.62**	**[0.46, 0.84]**	**0.002***
	UI-Bg	**0.68**	**[0.52, 0.88]**	**0.004***
	Change	**0.30**	**[0.23, 0.39]**	**< 0.001***
Resource	*Loc*	–	–	–
	Con	**0.33**	**[0.28, 0.38]**	**< 0.001***
	SMS	**0.13**	**[0.10, 0.18]**	**< 0.001***
Internet Skill	0	–	–	–
	+1	**0.96**	**[0.94, 0.98]**	**< 0.001***

*Significant effect – Base case (OR=1, by definition)

(b) Likelihood to Recommend

Table 4: Summary of regressions over participant comfort and likelihood to recommend apps with different access contexts.

Variable	Value	Odds Ratio	CI	p-value
Why	*Personalize*	–	–	–
	Server	0.93	[0.76, 1.15]	0.502
	Analytics	**0.47**	**[0.36, 0.61]**	**< 0.001***
	Ads	**0.22**	**[0.18, 0.27]**	**< 0.001***
	NA	**0.44**	**[0.33, 0.61]**	**< 0.001***
When	*Int*	–	–	–
	Pre	**0.69**	**[0.51, 0.94]**	**0.018***
	UI-Bg	**0.63**	**[0.48, 0.83]**	**< 0.001***
	Change	**0.33**	**[0.25, 0.43]**	**< 0.001***
Resource	*Loc*	–	–	–
	Con	**0.41**	**[0.35, 0.49]**	**< 0.001***
	SMS	**0.27**	**[0.21, 0.36]**	**< 0.001***
Internet Skill	0	–	–	–
	+1	**0.97**	**[0.95, 0.99]**	**0.002***

*Significant effect – Base case (OR=1, by definition)

Table 5: Summary of regression over participant beliefs regarding the usefulness of different access contexts.

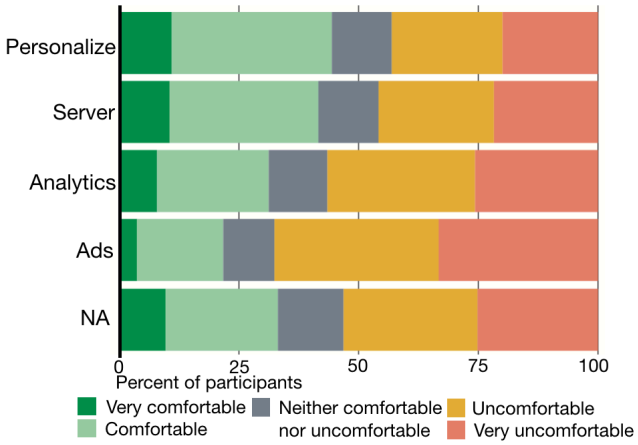


Figure 2: Likert-scale comfort organized by reason for resource access.

Table 5 shows the results of our logistic regression for whether the app's behavior is useful. This provides additional insight into participant preferences, as users may be more willing to tolerate uncomfortable behavior if it is useful. As shown in this table, the *Personalize* and *Server* conditions also do not differ significantly from each other in perceived usefulness.

We therefore conclude that *H1a* does not hold. This corroborates, at a larger scale, the findings of Shklovski et al., who showed that users were comfortable sharing information off device if it was only used by the app's developer [38].

First vs. third parties (H1b). We next consider whether participants responded to first-party accesses (*Personalize* and *Server*) differently than third-party accesses (*Analytics* and *Ads*). Figure 2 shows that participants were overall less comfortable with the third party accesses. Across our two first-party conditions, 43% of participants responded comfortable or very comfortable, compared to only 25% across our two third-party conditions.

As shown in Table 4a, differences between first- and third-party explanations are statistically significant. The *Analytics* and *Ads* conditions are associated with significantly less comfort than the base *Personalize* case. Further, the confidence intervals for *Analytics* and *Ads* do not overlap with that for *Server*, indicating that the two third-party conditions are each significantly different from the first-party *Server* condition as well. The same significance relation holds for app recommendations, as shown in Table 4b. In terms of effect size, the relative odds ratios among the first- and third-party conditions indicate that participants in third-party conditions were between one-third and two-thirds as likely to report a higher level of comfort than were the first-party participants. For example, participants were 0.6× as likely to report a higher level of comfort for *Analytics* than for *Server* (0.49/0.88), and 0.4× as likely for *Ads* than *Server* (0.34/0.88). The effect sizes for willingness to recommend were similar: 0.6× (0.51/0.91) and 0.3× (0.27/0.91), respectively.

A similar analysis of Table 5 shows that participants found

the behavior of apps in the third-party conditions (*Analytics*, *Ads*) to be significantly less likely to be seen as useful than the behavior of apps in the first-party conditions (*Personalize*, *Server*).

Overall, we conclude that *H1b* holds, and that the difference between first- and third-party accesses is meaningful.

Analytics vs. advertising (H1c). We find partial evidence to support *H1c*, which concerns the difference between our two third-party conditions. With respect to our main comfort question (Table 4a), the confidence intervals between *Analytics* and *Ads* overlap, indicating no significant difference between the two. However, comparing confidence intervals in Table 4b does show that participants were significantly more likely to recommend the app in the *Analytics* condition than in the *Ads* condition. *Ads* participants were only 53% (.27/.51) as likely to report a higher level of recommendation. Perhaps unsurprisingly, a parallel reading of Table 5 indicates that participants also found *Analytics* more useful than *Ads*.

Perception when no *why* is provided (H2). To test *H2*, we compare the *NA* condition, in which no reason is provided, to all the other *why* conditions. Overall, we find that *H2* holds partially; a lack of explanation is more comfortable than some explanations, but less comfortable than others.

Inspection of Figure 2 suggests that the *NA* condition falls in the middle of the pack in terms of expressed comfort; 33% of participants in this condition reported being comfortable or very comfortable with this behavior.

Referring again to the top section of Table 4a, we see that this "middle" impression is reflected in our statistical analysis. The *NA* condition is worse than the most comfortable case, with a point estimate of 0.58× the likelihood of higher comfort compared to the baseline *Personalize* condition. On the other hand, comparison of odds ratios suggests that the *NA* condition is slightly (but significantly) better than the worst case (*Ads*). Comparing odds ratios with the other *why* levels, we see that *NA* is not significantly different from *Server* or *Analytics*. The same trend—worse than *Personalize* but better than *Ads*—holds as well for responses to the recommendation question (Table 4b). With respect to the usefulness of the app's behavior, Table 5 indicates that *NA* scenarios were seen as less useful than *Personalize* and *Server*, but not different than *Analytics* or *Ads*.

We asked participants in the *NA* condition to provide an open-ended explanation for the resource accesses they were shown. Figure 3 shows how many participants (grouped according to the type of resource access they were shown) provided each of the most common reasons, according to our manual coding. (Note that an individual participant could provide more than one reason, so totals are greater than 100%.)

By far the most common response (76% of all *NA* participants) was that resource accesses were used for personalization. For example, one participant said Jane's location was accessed to "find singles that are nearby." The second-most common response was advertising (24% of all *NA* participants).

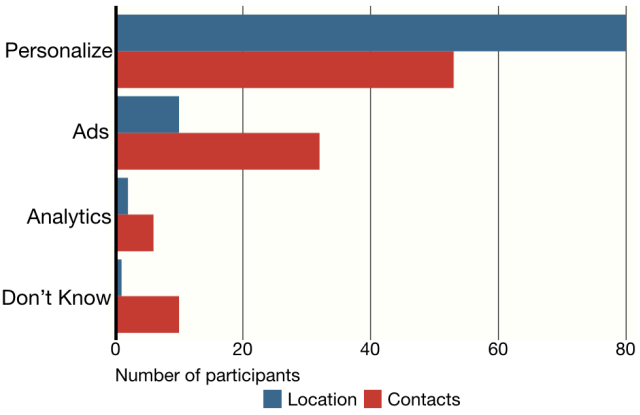


Figure 3: Number of participants who believed the app was collecting their data. Note, these codes are not mutually exclusive, so one participant could express multiple reasons for data access.

Our regression results suggest that a lack of explanation (*NA*) is less comfortable and useful than *Personalize*, even though most participants' assumed the resource access was actually for personalization. Because participants generally did not distinguish between on- and off-device personalization, these personalization responses can be considered roughly equivalent to either our *Personalize* or *Server* conditions. One potential explanation is that the uncertainty associated with a lack of explanation creates some discomfort, even when participants assume that the underlying explanation is acceptable.

Summary of *why* results. Overall, our results for *H1* and *H2* suggest that both who sensitive data is shared with and why matter: accesses used only by the app company for personalization are most comfortable, followed by third-party accesses associated with analytics, with third-party accesses for advertising least comfortable.

5.2 H3: Triggers for Resource Access

We next examine the effect of our *when* variable on users' responses, shown in the second group of results in Tables 4a, 4b, and 5, labeled *when*.

Interactive vs. non-interactive accesses (H3a). We first compare our three non-interactive triggers to the *Int* control condition, to validate that interactive accesses are more comfortable. We find that, as expected, *H3a* does hold. As shown in Tables 4a, 4b, and 5, we find that *Int* is associated with statistically significantly higher levels of comfort, willingness to recommend, and usefulness compared to every other *when* condition. Point estimates range from 1.4× (1/0.72, *UI-Bg*) to 2.9× (1/0.34, *Change*) more likely to report a higher comfort level.

Figure 4, which shows participants' Likert responses to the comfort question organized by *when* condition, illustrates this comfort gap between *Int* and the other *when* conditions.

Importance of visual indicator (H3b). We next consider whether an explicit foreground indication of use can increase user comfort, even if the indication happens after

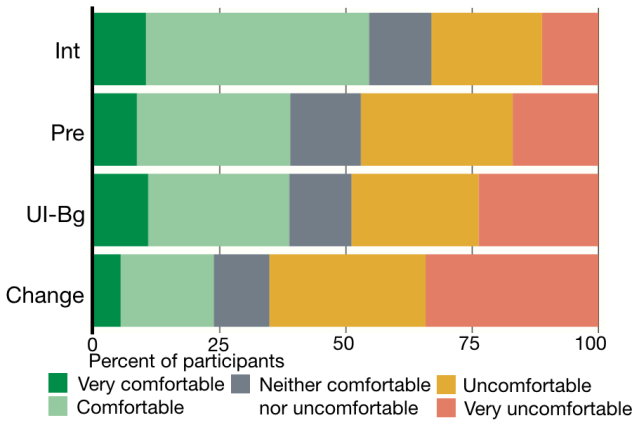


Figure 4: Likert-scale comfort organized by when the resource access occurred.

the access. In particular, we compare the *Pre* condition to the other background *when* conditions.

Comparison of odds ratios in Table 4a suggest that *H3b* holds partially: *Pre* is associated with significantly higher comfort levels than *Change* (1.9×, 0.64/0.34), but is not significantly different from *UI-Bg*. The same pattern holds for willingness to recommend and for usefulness, shown in Tables 4b and 5.

On-screen vs. off-screen (H3c). Finally, we compare the two *when* conditions without visual indicators: *UI-Bg*, which only includes background accesses while the app is on-screen, and *Change*, which includes accesses while the app is off-screen. We find that, as might be expected, off-screen accesses are significantly less comfortable, meaning *H3c* holds.

This finding can be observed visually in Figure 4, which shows that only 24% of participants were comfortable or very comfortable with the *Change* scenario. As shown by comparing odds ratios in Table 4a, this difference is significant: the point estimate suggests that *Change* is only 0.47× as likely to be associated with higher comfort as *UI-Bg* (0.34/0.72). Tables 4b and 5 exhibit the same significance relation for willingness to recommend and usefulness, respectively.

Summary of *when* results. Taken together, our results for *H3a–H3c* suggest three distinct classes of access triggers: interactive accesses, non-interactive (background) accesses that occur when the app is on screen, and background accesses that occur when the app is off screen.

5.3 Other Findings

As described in Section 3.3, our regression analysis included several other covariates beyond *why* and *when*. The final two groups of results shown in Table 4a indicate that the resource shown (*Loc*, *Con*, or *SMS*) and the participants' Internet skill both had significant effects on comfort. In particular, participants reported the highest levels of comfort with the baseline *Loc* resource. Access to *Con* was also significantly (2.8×, 0.33/0.12) more likely to comfortable than access to *SMS*. This aligns with prior work from Felt et al. [14]. Additionally, we observed that users who scored higher on Hargittai and Hsieh's Internet skill scale [17] were significantly likely to be less comfortable (OR 0.95, $p < 0.001$). This means that a participant with the maximum possible score of 35 would be about 0.87× ($0.95^{2.8}$) as likely to express increased comfort as a participant with the mean score of 32.2. This result is analogous to Liccardi et al.'s finding that users with less understanding of how apps operate were more likely to download apps requiring additional significant permissions [25].

Resource type and Internet skill exhibited similar significance relations in willingness to recommend and usefulness (Tables 4b and 5, last two sections), with one exception: accesses to *SMS* were not viewed as significantly less useful than accesses to *Con*.

Notably, none of the interactions we considered (Table 2) appeared in the final minimum-BIC model for any of our outcome variables. This suggests that these variables—most importantly, the *when* and *why* context factors—can be considered independent from each other.

Similarly, app type was not included in any of the final models, meaning we did not observe a significant difference between participants' responses to the Datr and Ridr apps. However, because we only tested two apps, we cannot conclude that the app has no effect on user comfort.

Finally, we observed that a large percentage of participants stated they were uncomfortable or very uncomfortable in all the tested *why* and *when* conditions. In fact, *Int* was the only condition where the majority of participants expressed comfort. In practice, of course, users do use apps with these sorts of background uses. One explanation could be that participants tend to over-report the magnitude of their privacy concerns. Alternatively, users may in practice continue to use apps that violate their privacy preferences because the utility outweighs the cost.

6. DESIGN RECOMMENDATIONS

Based on the results of our study, we make several recommendations for app developers, designers of mobile-privacy systems, and third-party app auditors:

Developers should provide context-sensitive access descriptions. When no reason for an access is given, we found that users are too generous in their assumptions about access context. For example, in absence of explanation, users will tend towards assuming data is being used for personalization (although with slightly lower comfort, perhaps due to uncertainty). If an access is actually used for advertising (or worse, for both advertising and personalization), users might authorize more access than they are actually comfortable with. On the other hand, if data is actually used only for personalization or remains on the device, providing this information could allow the user to feel more comfortable allowing a request than they otherwise would.

Both in Android and iOS, by default whenever an app requests permission to access a sensitive resource (i.e., on first use of the resource), no reason is given for that access. Both systems allow developers to provide a reason, but in practice very few developers take advantage of this feature [41]. Users should be skeptical of any access presented without an explanation, since developers are disincentivized to explain accesses that are used for advertising. Perhaps the Android

API could require an explanation from a fixed set of options, or even default to a "may be used for advertising" explanation if the developer fails to provide a reason. Legitimate developers could presumably be incentivized to provide accurate information to avoid charges of fraud or deceptive practices (analogous to privacy policies).

Further, Tan et al. found that many developers did not include description strings because they did not think they were useful [41]. Our results provide evidence for the utility of these descriptions and could be used to inform design of description strings to ensure users are only shown information relevant to their decisions.

Privacy support agents should consider nuanced variations of context. Because it is unlikely that all app developers will act altruistically, several systems have been proposed to help users make informed decisions according to their privacy preferences, with context in mind [28, 30, 46]. In each, the authors group various access contexts together. We found that such groupings may be insufficiently nuanced. For example, Wijesekera et al. learn user preferences based on whether the app is on- or off-screen at the time of access [46]. This grouping conflates *Int*, *Pre*, and *UI-Bg*, which all occur when the app is on-screen, but were associated with significantly different user comfort levels in our study. Our prior work makes another split, recommending that interactive accesses be treated differently from those that are not associated with user interaction [30]. Again, this oversimplifies user comfort with non-interactive accesses: our results show significant differences between *Pre* and *UI-Bg*, in one class, and *Change* in another.

As a positive example, the privacy assistant developed by Liu et al. divides reasons for resource access into first-party, analytics, and advertisement bins [28]. This grouping accounts for the differences in user comfort we observe in the *why* context. Future such systems should attempt to accurately capture nuanced resource-access classes across both *when* and *why*.

Third-party app auditors should focus on our presented tiers of context. Finally, we believe the job of app auditors can be simplified by concentrating on the most significant contextual classes when investigating app behavior. For example, auditors can focus their efforts on data that is shared off device, because this is most likely to cause user discomfort.

This also highlights the need for tools to support helping auditors answer questions specific to the tiers of context we found. For example, our results underscore the importance of data flow analyses such as Taintdroid [13] and FlowDroid [3].

7. CONCLUSION

In this work, we used a 52-condition, 2,198-participant vignette study to examine how the context of a sensitive resource access in Android—defined as both *when* and *why* the access occurs—affects user comfort with that access. In particular, we examined whether users think similarly about different kinds of background resource accesses, or whether there are important distinctions that determine users' comfort with those accesses.

We found that both *when* and *why* a sensitive resource access occurs have a statistically significant effect on user comfort, and that there are meaningful differences between classes of accesses within both access context variables. While users are most comfortable with interactive accesses, they also make a distinction between non-interactive accesses occurring when an app is on- compared to off-screen. Similarly, users are more comfortable with first-party than third-party accesses, but also make a distinction between third-party accesses for analytics as compared to advertising. We recommend that designers of mobile-privacy systems not only consider both *when* and *why* a resource access is requested, but also respect nuanced distinctions that influence user comfort.

8. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful feedback. This research was supported in part by a UMIACS contract under the partnership between the University of Maryland and DoD, and by a Google Research Award.

9. REFERENCES

[1] P. Andriotis, S. Li, T. Spyridopoulos, and G. Stringhini. A comparative study of android users' privacy preferences under the runtime permission model. In T. Tryfonas, editor, *Proceedings of the 5th International Conference on Human Aspects of Information Security, Privacy and Trust*, HAS '17, pages 604–622. Springer International Publishing, 2017.

[2] P. Andriotis, M. A. Sasse, and G. Stringhini. Permissions snapshots: Assessing users' adaptation to the android runtime permission model. In *Proceedings of the 8th IEEE International Workshop on Information Forensics and Security*, WIFS '16, pages 1–6, Dec 2016.

[3] S. Arzt, S. Rasthofer, C. Fritz, E. Bodden, A. Bartel, J. Klein, Y. Le Traon, D. Octeau, and P. McDaniel. Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps. In *Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI '14, pages 259–269, New York, NY, USA, 2014. ACM.

[4] C. Atzmüller and P. M. Steiner. Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6(3):128, 2010.

[5] R. Balebako, J. Jung, W. Lu, L. F. Cranor, and C. Nguyen. "little brothers watching you": Raising awareness of data leaks on smartphones. In *Proceedings of the 9th Symposium on Usable Privacy and Security*, SOUPS '13, pages 12:1–12:11, New York, NY, USA, 2013. ACM.

[6] R. Böhme and J. Grossklags. The security cost of cheap user interaction. In *Proceedings of the 2011 New Security Paradigms Workshop*, pages 67–82. ACM, 2011.

[7] B. Bonné, S. T. Peddinti, I. Bilogrevic, and N. Taft. Exploring decision making with android's runtime permission dialogs using in-context surveys. In *Proceedings of the 13th Symposium on Usable Privacy*

and Security, SOUPS '17, pages 195–210, Santa Clara, CA, 2017. USENIX Association.

[8] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[9] K. Z. Chen, N. M. Johnson, V. D'Silva, S. Dai, K. MacNamara, T. R. Magrino, E. X. Wu, M. Rinard, and D. X. Song. Contextual policy enforcement in android applications with permission event graphs. In *Proceedings of the 20th Annual Network and Distributed System Security Symposium*, NDSS '13, page 234, San Diego, CA, 2013. Internet Society.

[10] J. L. Devore. *Probability and Statistics for Engineering and the Sciences*. Cengage Learning, 2015.

[11] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor. Are your participants gaming the system?: Screening mechanical turk workers. In *Proceedings of the 28th ACM Conference on Human Factors in Computing Systems*, CHI '10, pages 2399–2402, New York, NY, USA, 2010. ACM.

[12] K. O. Elish, D. Yao, and B. G. Ryder. User-centric dependence analysis for identifying malicious mobile apps. In *Proceedings of the 1st Workshop on Mobile Security Technologies*, MoST '12, San Jose, CA, 2012. IEEE Press.

[13] W. Enck, P. Gilbert, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth. Taintdroid: An information-flow tracking system for realtime privacy monitoring on smartphones. In *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation*, OSDI'10, pages 393–407, Berkeley, CA, USA, 2010. USENIX Association.

[14] A. P. Felt, S. Egelman, and D. Wagner. I've got 99 problems, but vibration ain't one: A survey of smartphone users' concerns. In *Proceedings of the 2nd ACM Workshop on Security and Privacy in Smartphones and Mobile Devices*, SPSM '12, pages 33–44, New York, NY, USA, 2012. ACM.

[15] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the 8th Symposium on Usable Privacy and Security*, SOUPS '12, pages 3:1–3:14, New York, NY, USA, 2012. ACM.

[16] Google. *Requesting Permissions at Run Time*, 2016.

[17] E. Hargittai and Y. P. Hsieh. Succinct survey measures of web-use skills. *Social Science Computer Review*, 30(1):95–107, 2012.

[18] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[19] J. Huang, X. Zhang, L. Tan, P. Wang, and B. Liang. Asdroid: Detecting stealthy behaviors in android applications by user interface and program behavior contradiction. In *Proceedings of the 36th International Conference on Software Engineering (ICSE)*, ICSE 2014, pages 1036–1046, New York, NY, USA, 2014. ACM.

[20] R. Kang, S. Brown, L. Dabbish, and S. Kiesler. Privacy attitudes of mechanical turk workers and the u.s. public. In *Proceedings of the 23rd USENIX Security Symposium*, USENIX Security '14, pages 37–49, San Diego, California, USA, 2014. USENIX Association.

[21] R. Kang, L. Dabbish, N. Fruchter, and S. Kiesler. "my data just goes everywhere:" user mental models of the internet and implications for privacy and security. In *Proceedings of the 11th Symposium On Usable Privacy and Security*, SOUPS '15, pages 39–52, Ottawa, 2015. USENIX Association.

[22] P. G. Kelley, S. Consolvo, L. F. Cranor, J. Jung, N. Sadeh, and D. Wetherall. A conundrum of permissions: Installing applications on an android smartphone. In *Proceedings of the 16th International Conference on Financial Cryptography and Data Security*, FC '12, pages 68–79, Berlin, Heidelberg, 2012. Springer-Verlag.

[23] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the 26th ACM Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.

[24] E. J. Langer, A. Blank, and B. Chanowitz. The mindlessness of ostensibly thoughtful action: the role of "placebic" information in interpersonal interaction. *Journal of personality and social psychology*, 36(6):635, 1978.

[25] I. Liccardi, J. Pato, D. J. Weitzner, H. Abelson, and D. De Roure. No technical understanding required: Helping users make informed choices about access to their personal data. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MOBIQUITOUS '14, pages 140–150, ICST, Brussels, Belgium, Belgium, 2014. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[26] J. Lin, S. Amini, J. I. Hong, N. Sadeh, J. Lindqvist, and J. Zhang. Expectation and purpose: Understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 14th ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 501–510, New York, NY, USA, 2012. ACM.

[27] J. Lin, B. Liu, N. Sadeh, and J. I. Hong. Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Proceedings of the 10th Symposium On Usable Privacy and Security*, SOUPS '14, pages 199–212, Menlo Park, CA, 2014. USENIX Association.

[28] B. Liu, M. S. Andersen, F. Schaub, H. Almuhimedi, S. A. Zhang, N. Sadeh, Y. Agarwal, and A. Acquisti. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Proceedings of the 12th Symposium on Usable Privacy and Security*, SOUPS '16, pages 27–41, Denver, CO, 2016. USENIX Association.

[29] P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142, 1980.

[30] K. Micinski, D. Votipka, R. Stevens, N. Kofinas, M. L. Mazurek, and J. S. Foster. User interactions and permission use on android. In *Proceedings of the 35th*

ACM on Human Factors in Computing Systems, CHI '17, New York, NY, USA, 2017. ACM.

[31] H. Nissenbaum. Privacy as Contextual Integrity. *Washington Law Review*, 79:119–157, 2004.

[32] H. Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

[33] P. Pearce, A. P. Felt, G. Nunez, and D. Wagner. Addroid: Privilege separation for applications and advertisers in android. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, ASIACCS '12, pages 71–72, New York, NY, USA, 2012. ACM.

[34] A. E. Raftery. Bayesian model selection in social research. *Sociological methodology*, pages 111–163, 1995.

[35] A. Razaghpanah, R. Nithyanand, N. Vallina-Rodriguez, S. Sundaresan, M. Allman, C. Kreibich, and P. Gill. Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem. In *Proceedings of the 24th Network and Distributed System Security Symposium*, NDSS '18, San Diego, California, USA, 2018. Internet Society.

[36] T. Ringer, D. Grossman, and F. Roesner. Audacious: User-driven access control with unmodified operating systems. In *Proceedings of the 23rd ACM Conference on Computer and Communications Security*, Vienna, Austria, oct 2016. ACM.

[37] F. Roesner, T. Kohno, A. Moshchuk, B. Parno, H. J. Wang, and C. Cowan. User-driven access control: Rethinking permission granting in modern operating systems. In *2012 IEEE Symposium on Security and Privacy*, pages 224–238, May 2012.

[38] I. Shklovski, S. D. Mainwaring, H. H. Skúladóttir, and H. Borgthorsson. Leakiness and creepiness in app space: Perceptions of privacy and mobile app use. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2347–2356, New York, NY, USA, 2014. ACM.

[39] P. Software. Jeb decompiler, 2017. (Accessed 5-19-2017).

[40] M. Spence and R. Zeckhauser. Insurance, information, and individual action. In *Uncertainty in Economics*, pages 333–343. Elsevier, 1978.

[41] J. Tan, K. Nguyen, M. Theodorides, H. Negrón-Arroyo, C. Thompson, S. Egelman, and D. Wagner. The effect of developer-specified explanations for permission requests on smartphone user behavior. In *Proceedings of the 32nd ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 91–100, New York, NY, USA, 2014. ACM.

[42] P. R. C. I. . Technology. March 7 - april 4, 2016 – libraries, 2018. (Accessed 1-15-2018).

[43] C. Thompson, M. Johnson, S. Egelman, D. Wagner, and J. King. When it's better to ask forgiveness than get permission: Attribution mechanisms for smartphone resources. In *Proceedings of the 9th Symposium on Usable Privacy and Security*, SOUPS '13, pages 1:1–1:14, New York, NY, USA, 2013. ACM.

[44] M. Toomim, T. Kriplean, C. Pörtner, and J. Landay. Utility of human-computer interactions: Toward a science of preference measurement. In *Proceedings of the 29th ACM Conference on Human Factors in Computing Systems*, CHI '11, pages 2275–2284, New York, NY, USA, 2011. ACM.

[45] P. Wijesekera, A. Baokar, A. Hosseini, S. Egelman, D. Wagner, and K. Beznosov. Android permissions remystified: A field study on contextual integrity. In *Proceedings of the 24th USENIX Security Symposium*, USENIX Security '15, pages 499–514, Washington, D.C., Aug. 2015. USENIX Association.

[46] P. Wijesekera, J. Reardon, I. Reyes, L. Tsai, J.-W. Chen, N. Good, D. Wagner, K. Beznoso, and S. Egelman. Contextualizing privacy decisions for better prediction (and protection). In *Proceedings of the 36th ACM on Human Factors in Computing Systems*, CHI '18, New York, NY, USA, 2018. ACM.

[47] W. Yang, X. Xiao, B. Andow, S. Li, T. Xie, and W. Enck. Appcontext: Differentiating malicious and benign mobile app behaviors using context. In *Proceedings of the 37th IEEE International Conference on Software Engineering*, volume 1 of *ICSE '15*, pages 303–313, Florence, Italy, May 2015. ACM.

[48] Z. Yang, M. Yang, Y. Zhang, G. Gu, P. Ning, and X. S. Wang. Appintent: Analyzing sensitive data transmission in android for privacy leakage detection. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS '13, pages 1043–1054, New York, NY, USA, 2013. ACM.

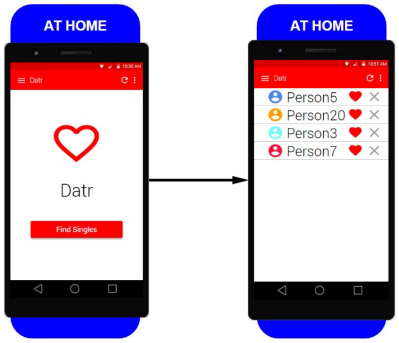
[49] Z. Yang, M. Yang, Y. Zhang, G. Gu, P. Ning, and X. S. Wang. Appintent: Analyzing sensitive data transmission in android for privacy leakage detection. In *Proceedings of the 2013 ACM Conference on Computer and Communications Security*, pages 1043–1054. ACM, 2013.

[50] H. Zhu, H. Xiong, Y. Ge, and E. Chen. Mobile app recommendations with security and privacy awareness. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 951–960, New York, NY, USA, 2014. ACM.

APPENDIX

A. SURVEY QUESTIONNAIRE

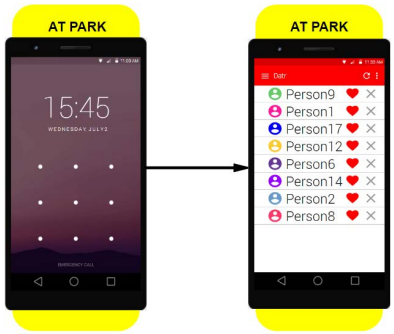
App usage description and attention check.



While at home, Jane decides to use Datr to look for other singles. She opens the app and presses the button "Find

Singles". The app then shows her a screen with a list of recommended singles.

Jane closes the app and travels to a nearby coffee shop where she meets her friend, Sarah. As they get ready to leave, Jane realizes she does not have Sarah's contact information. Jane adds Sarah to her phone's contacts and Sarah sends Jane a text message to remind her that they should meet again some time.



After leaving the coffee shop, Jane heads to the park. She decides to check Datr again and is presented with a new list of singles.

1. Which of the below options best describes the set of steps Jane would have to take to indicate that she is interested in Person 9?
   (a) Press the heart-shaped icon next to Person 9
   (b) Press the X icon next to Person 12
   (c) Press the reload symbol

Resource access description and study questions.

While Jane was using Datr, the app behaved in the following way:

Whenever Jane's location changed, Datr learned about the change to her location and sent her updated location to datr.com. datr.com then used her updated location along with other updates it had collected on Jane previously to create a list of recommended singles based on places she has traveled in the past.

For the remaining questions, we're going to ask you about an app like Datr that collects your location whenever your location changes and sends it to its server to provide personalized features and does not send your location to other parties.

1. Do you think popular dating apps collect your location whenever your location changes and send it to its server to provide personalized features and do not send your location to other parties?
   (a) Very likely
   (b) Likely
   (c) Neither likely nor unlikely
   (d) Unlikely
   (e) Very unlikely
2. Please indicate your level of agreement with the following statement: A dating app like Datr is more useful when it collects your location whenever your location changes and sends it to its server to provide personalized features and does not send your location to other parties?
   (a) Agree
   (b) Somewhat agree
   (c) Neither agree nor disagree
   (d) Somewhat disagree
   (e) Disagree
3. Suppose you were interested in using a dating app like Datr. How would you feel about using a dating app that collects your location whenever your location changes and sends it to its server to provide personalized features and does not send your location to other parties?
   (a) Very comfortable
   (b) comfortable
   (c) Neither comfortable nor uncomfortable
   (d) Uncomfortable
   (e) Very uncomfortable
4. Suppose you know someone who wants to use a dating app like Datr. If you had to recommend an app for them to use, would you recommend an app that collects your location whenever your location changes and sends it to its server to provide personalized features and does not send your location to other parties?
   (a) Very likely
   (b) Likely
   (c) Neither likely nor unlikely
   (d) Unlikely

(e) Very unlikely

Note: For the none case we also included the following free response question

5. Please provide a short description of why you think Datr is interested in knowing Jane's location.

Internet skill questionnaire.

1. How familiar are you with the following computer and Internet-related items? (Items: Reload, Bookmark, Advanced Search, Favorites, Tagging, Preference Settings, PDF) (Choices: No Understanding, Little Understanding, Good Understanding, Full Understanding)

Demographics.

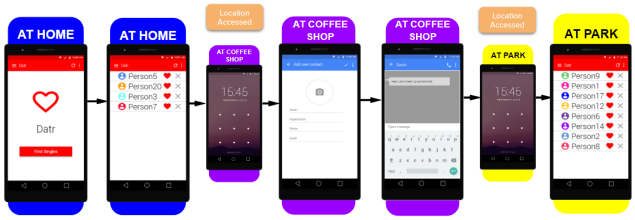
1. What is the highest level of school you have completed or the highest degree you have received? (Choices: Less than high school degree, High school graduate (high school diploma or equivalent including GED), Some college but no degree, Associate degree (2-year), Bachelor's degree (4-year), Master's degree, Doctoral degree, Prefer not to answer)
2. Please specify the gender with which you most closely identify. (Choices: Male, Female, Other, Prefer not to answer)
3. Please specify your ethnicity. (Choices (may choose multiple): Hispanic or Latino, Black or African American, White, American Indian or Alaska Native, (Asian, Native Hawaiian, or Pacific Islander), Other, Prefer not to answer)
4. Please specify your age.
5. Please select the response option that best describes your household income in 2017, before taxes. (Choices: Less than $5,000, $5,000 - $14,999, $15,000 - $29,999, $30,000 - $49,999, $50,000 - $74,999, $75,000 - $99,999, $100,000 - $149,999, $150,000 - $199,999, $200,000 or more, Prefer not to answer)
6. Please select the response option that best describes your current employment status. (Choices: Working for payment, Unemployed, Looking after home/family, Student, Retired, Unable to work due to permanent sickness or disability, Other(specify), Prefer not to answer)
7. How many hours a day do you use your smartphone? (Choices: 0-3, 3-6, 6-9, 9+, Unsure, I do not own a smartphone)
8. Rate your expertise using a smartphone. (Choices: Far above average, Somewhat above average, Average, Somewhat below average, Far below average)

B. EXAMPLE SCENARIOS

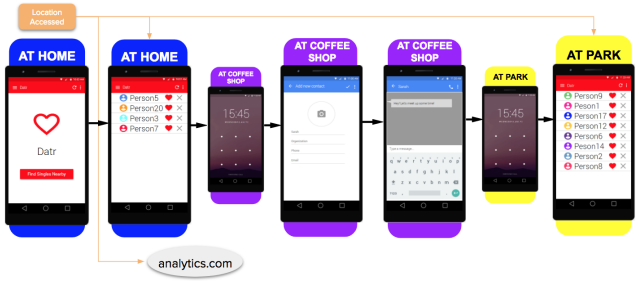
Here, we give a few representative examples of the different resource access scenarios shown to users along with the description of app behavior provided.

(Rideshare, Contacts, UI Background, Advertising)

When Jane presses the button "Request Ride", Ridr learned the contacts in her contact list and sent her contact list to a third party advertiser (advertising.com). advertising.com then used her contact list to better target advertisements to her in the future.



(Dating , Location, On Change, Not Given)
Whenever Jane's location changed, Datr learned about the change and her new location.



(Dating, Location, Interactive, Debugging/Analytics)
When Jane pressed the button "Find Singles Nearby", Datr learned her current location and sent her location to a third party website (analytics.com). analytics.com then used this location information along with other location data it had collected on Jane previously to fix bugs and other problems in the app. Datr also used her location to create a list of recommended singles based on places she has traveled in the past.

(Dating, Contacts, Prefetch, Personalize)
When Jane opened Datr, Datr learned the contacts in her contact list and used her contact list to create a list of recommended singles nearby. Datr creates this list ahead of time so that the list can be displayed quickly if Jane presses the "Find Singles Based On Contacts" button (instead of having to wait a few seconds after the button is pressed). Datr only uses Jane's contact list to personalize her recommendations and does not send her contact list to any other parties (i.e., datr.com or advertisers).

Let Me Out! Evaluating the Effectiveness of Quarantining Compromised Users in Walled Gardens

Orçun Çetin, Lisette Altena, Carlos Gañán, and Michel van Eeten

Department of Multi-Actor Systems, Delft University of Technology
{F.O.Cetin,E.M.Altena,C.Hernandezganan,M.J.G.Vaneeten} @tudelft.nl

ABSTRACT

In the fight to clean up malware-infected machines, notifications from Internet Service Providers (ISPs) to their customers play a crucial role. Since stand-alone notifications are routinely ignored, some ISPs have invested in a potentially more effective mechanism: quarantining customers in so-called walled gardens. We present the first empirical study on user behavior and remediation effectiveness of quarantining infected machines in broadband networks. We analyzed 1,736 quarantining actions involving 1,208 retail customers of a medium-sized ISP in the period of April-October 2017. The first two times they are quarantined, users can easily release themselves from the walled garden and around two-thirds of them use this option. Notwithstanding this easy way out, we find that 71% of these users have actually cleaned up the infection during their first quarantine period and, of the recidivists, 48% are cleaned after their second quarantining. Users who do not self-release either contact customer support (30%) or are released automatically after 30 days (3%). They have even higher cleanup rates. Reinfection rates are quite low and most users get quarantined only once. Users that remain infected spend less time in the walled garden during subsequent quarantining events, without a major drop in cleanup rates. This suggests there are positive learning effects, rather than mere habituation to being notified and self-releasing from the walled garden. In the communications with abuse and support staff, a fraction of quarantined users ask for additional help, request a paid technician, voice frustration about being cut off, or threaten to cancel their subscriptions. All in all, walled gardens seem to be a relatively effective and usable mechanism to improve the security of end users. We reflect on our main findings in terms of how to advance this industry best practice for botnet mitigation by ISPs.

1. INTRODUCTION

Fighting the scourge of malware-infected end user machines is an ongoing challenge that involves many different actors, from software vendors, incident response organizations, antivirus vendors, network operators and, last but not least, the end users themselves. Some efforts are more focused on preventing infections, others on remediation – i.e., cleaning up the compromised hosts. In the context of cleanup, the role of Internet Service Providers has become more salient over time, as it became clear that many end users struggle to detect and remediate infections. The ISPs are a critical control point providing the infected machines with access to the rest of the Internet. In the past 5-10 years, a range of best practices and code of conducts have been published by leading industry associations [22, 24], public-private initiatives [11, 17] and governmental entities [12, 16]. These documents share a common set of recommendations for ISPs around educating customers, detecting infections, notifying customers, and remediating infections.

The effectiveness of these best practices is disputed. When the U.S. National Institute of Standards and Technology (NIST) was developing its own guidance on 'Models To Advance Voluntary Corporate Notification to Consumers Regarding the Illicit Use of Computer Equipment by Botnets and Related Malware', it considered using the Australian iCode as an example [26]. The SANS Institute and other stakeholders criticized this idea, arguing the Australian code had not managed to significantly improve cleanup rates of infected users [26]. Academic research has also questioned the effectiveness of these efforts [3, 4].

There are a variety of reasons for the limited impact of botnet remediation efforts by ISPs. At the core, however, is a usability problem: notifying customers that one of their machines is infected does not translate into actual cleanup. As we know from other areas in security, notifications are routinely ignored, especially if the step towards action is complicated and disrupts ongoing activities.

The lack of effectiveness of mere notifications has led some of the more security-minded ISPs to adopt what is arguably the most costly measure: putting infected customer machines into a quarantine network, also known as a 'walled garden', which only gives access to a small set of white-listed sites. Users are required to perform cleanup to get their connection restored – i.e., to be released from the walled garden. While the use of walled gardens is identified as a security best practice [25], it is also controversial. The ITU's Anti-Botnet Toolkit cites 'technical, financial, legal and customer satisfaction-related disincentives' that may be raised by an ISP [15].

Quarantining infected users is contested, but also one of the few measures that could improve cleanup rates and help end users to remediate and secure their machines. Remarkably, there has been no publicly available study on the effectiveness of walled gardens. Do they actually help end users to clean up? How often do users get reinfected? How much time do users spend in quarantine? How much support do they need? How much pushback do ISPs face from their users?

We present the first empirical study on the usability and effective-

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

ness of walled gardens as a notification and remediation mechanism. We analyzed 6 months of data (April-October 2017) from a real-world implementation of a walled garden at a medium-sized ISP that we collaborated with. The ISP is a market leader in its home market that serves retail broadband to several million customers. The ISP took 1,736 quarantining actions involving 1,208 retail customers. In collaboration with the ISP, we correlated these quarantining actions with independent observations from botnet sinkhole data to track remediation success. We also analyzed anonymized communications with quarantined users. In combination, these datasets allow us to estimate cleanup rates, recidivism rates, and user engagement with the walled garden environment.

In short, we make the following contributions:

- We present the first empirical study of a real-world 'walled garden' system to notify and quarantine end users with malware-infected machines – a widely-recognized security best practice for ISPs.
- We measure the effectiveness of the walled garden notifications in terms of end user cleanup efforts and find that the majority of users spend a relatively short time in quarantine, while still successfully removing the infection.
- We provide insight into the experiences of users by analyzing their communication with ISP employees and find that a fraction of them are frustrated about their access being cut off. This is especially true for users who turn out to operate business services over their consumer broadband connection.

The rest of this paper is structured as follows. Section 2 reviews prior work. Section 3 outlines the properties of walled garden systems and Section 4 presents the data collection methodology. Next, Section 5, we shed light on the effectiveness of the real-world walled garden and relationship between cleanup success and other factors. Section 6 presents key insights gathered from communications. Section 7 presents the ethical considerations and Section 8 discusses the limitations of the study. We conclude by covering the main lessons learned for the use of walled garden systems in securing end-user machines.

2. RELATED WORK

As far as we are aware, there is no prior work on the effectiveness of notifying end users in an access network and asking them to clean up malware infections on their machines. Here, we briefly survey four related areas of work. The work on abuse and vulnerability notifications has studied similar mechanisms , but typically with a different type of end user, namely webmasters, server admins and network operators, not home users. This makes the effectiveness of those mechanisms difficult to compare with malware notifications and cleanup by consumers. Another area of related work concerns the design of the notifications and warnings for regular end users. These notifications and warnings are mostly meant to prevent compromise, trying to steer the user back to safety. In contrast, we study a notification mechanism where the action is not avoiding danger, but dealing with the damage that has already occurred. Also, the action required of the user in case of compromise is not a single decision for or against a potentially dangerous action, but the execution of a rather complicated set of steps to resolve the incident that has already manifested itself. Finally, there is related work that studies whether and how end users understand the security situations they face and how they behave in those contexts. In our study, we do not observe the users directly, nor elicit their thoughts about the situation, but we do have data on some of their actions, as well as some visibility into their experiences through their communications with the ISP.

2.1 Abuse notifications

A range of studies has focused on if and how abuse notifications can expedite cleanup of compromised websites. Notifications can be sent to the affected owners of the site or to their hosting provider. An early study by Vasek *et al.* [1] indicated that more verbose abuse notifications to hosting providers resulted in higher cleanup rates than notifications with minimal information. Çetin *et al.* [10] found that around half of all compromised sites got cleaned up after a notification to the hosting provider. The reputation of the sender of the notifications had no observable impact on the cleanup rate. Li *et al.* [21] showed that direct notifications to webmasters via Google's Webmaster Console increased the likelihood of cleanup by over 50%. They report that 6.6% of sites cleaned up within a day of detection, 27.9% within two weeks, and 41.2% within one month. In a qualitative study, Canali *et al.* [8] set up vulnerable web servers on 22 hosting services, ran different attacks on them that simulated infections and then notified the providers about these attacks. Only one hosting provider notified their customers about a potential compromise of their website after the first notification and only half of the providers after the second notification. Additionally, around 13% of the notified providers warned the user of being compromised upon receiving abuse notifications.

2.2 Vulnerability notifications

Various studies have looked into the feasibility and efficacy of vulnerability notification mechanisms. For example, Kührer *et al.* [20] issued notifications to administrators of vulnerable Network Time Protocol (NTP) servers, in collaboration with CERTs, clearinghouses and afflicted vendors. Though their study lacks a control group to assess the impact of the campaign itself, they found that 92% of NTP server were remediated in 13 weeks. Stock *et al.* [29] studied large-scale vulnerability notification campaigns and found that only around 6% of the affected parties could be reached. Of that small fraction, around 40% were remediated upon notification. Similarly, in a study by Çetin *et al.* [9], the authors concluded that the deliverability of email-based notifications was very poor. They proposed searching for other mechanisms. Stock *et al.* [28] later tested the effectiveness of other channels such as postal mail, social media, and phone and concluded that the slightly higher remediation rates of these channels do not justify the additional work and costs.

2.3 Design of notifications and warnings

A large body of literature explored user responses to different types of security notifications and warnings, focusing on why users ignore warnings and how this could be avoided. A study conducted by Krol *et al.* [19] showed that users' misunderstanding of warnings and notifications is a reason for ignoring them. Almuhimedi *et al.* [2] studied user reactions to Google Chrome malware warnings. Up to half of the warnings were ignored under certain circumstances. Some users confused the malware warnings with SSL warnings. Sunshine *et al.* [30] examined users' reactions to existing and newly designed SSL warnings and suggested that, although existing SSL warnings can be improved, minimizing the use of SSL warnings by blocking users from making insecure connections proves to be more effective. Finally, Mathur *et al.* concluded that one of the reasons why users ignore software updates is that updates regularly interrupt users who often lack sufficient basic information to decide whether or not to update [23]. A closely related topic is

the problem of habituation of users to ignore warnings after they have learned that this does not seem to cause any harm [18, 27]. Bravo-Lillo *et al.* tested the effectiveness of user-interface modifications to draw users' attention to the most important information required for decisions [6, 7].

2.4 End user security behavior

Multiple studies have demonstrated that end users have difficulty securing their computers, either because of lack of knowledge or ignoring security advice that is hard to understand. In a study conducted by Wash *et al.* [31] on how users perceive automated software updates, the authors observed that the majority of users do not correctly understand the automatic update settings on their computer and cannot manage software updates the way they intend to. This mismatch between intention and behavior frequently led to computers being more or less secure than intended. Fagan *et al.* [13] studied user motivations regarding their decisions on following common security advice (i.e., update software, use password manager, change passwords) and concluded that the majority of users follow the usability/security trade-off. Finally, Forget *et al.* [14] developed a Security Behavior Observatory to collect data on users' behavior and their machine configurations. Their findings highlighted the importance of content, presentation, and functionality of security notifications provided to users who have different expertise, expectations, and computer security engagement.

3. WALLED GARDEN

The concept of a "walled garden" stems from the early days of the web, when ISPs implemented closed networks to control the applications, content and media that their subscribers could access. Some ISPs extended the capabilities of these networks to exclude rival content from the heavily curated garden. This model has all but disappeared.

These days, walled gardens are a method to notify subscribers about malware infections and restrict their access to the Internet while infected, so as to protect the infected user from further harm as well as preventing the user's machine from harming other users or networks. More precisely, a walled garden is a quarantined environment that restricts the information flow and services of an end user inside a network. Besides keeping the infected users safely in quarantine, the walled garden also plays an important role in informing the user. While the user tries to browse the Web, she or he will be redirected to a landing website with information about the type of infection and how to clean it up. Whereas emails or letters with the same content can be ignored relatively easily, this mechanism cannot.

There are different ways of implementing and deploying walled gardens to fight malware infections. RFC6561 [22] describes 2 different types: *strict*, a walled garden environment that restricts almost all services, except those to a whitelist of malware mitigation services; and *leaky*, an implementation that permits access to all Internet resources, except those that are deemed malicious, and ensures access to those that can be used to notify users of infections. In this paper, we focus on a strict implementation, which is what was installed at our partner ISP. A strict implementation is potentially more effective, but also more contested.

The quarantine period of an infected user mainly depends on three different processes: (i) the malware detection process; (ii) the infection notification and quarantining process; and (iii) the release process. The flow chart in Figure 1 shows the overall quarantine process in place at our partner ISP. It starts with the ISP realizing that a subscriber is infected and ends with the subscriber leaving

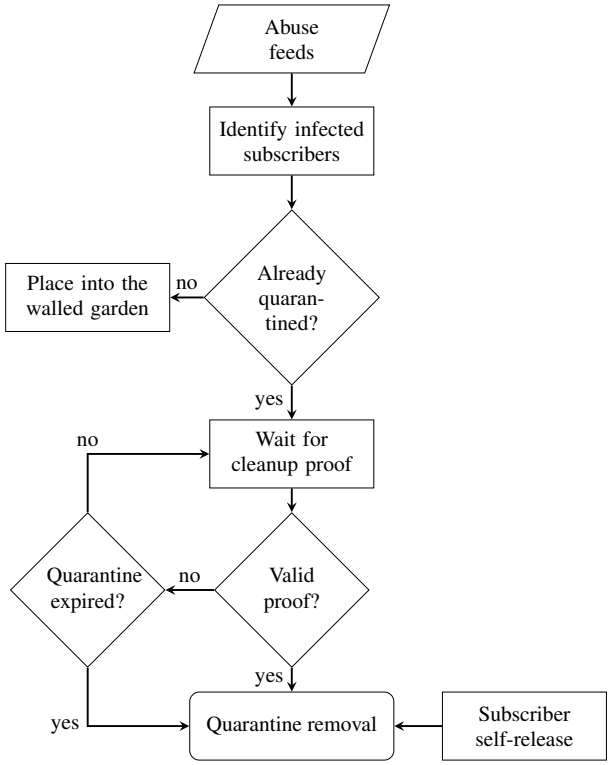


Figure 1: Quarantine flow chart

the walled garden. The starting point, i.e., the infection detection, is independent of the walled garden environment. Typically, this detection is not based on their own network monitoring, but on third-party notifications, e.g., from botnet sinkhole operators and security intelligence providers. The processing of abuse feeds varies per ISP, ranging from manually checking incoming notifications to highly automated systems that consume the feed and push the relevant incidents into abuse ticketing systems. When certain abuse data fits a predefined policy, on data trustworthiness, timeliness, the affected customer type and other criteria, the ISP places the connection of that particular customer into the walled garden.

In order to leave the walled garden, the customer is requested to provide proof of the cleanup actions that were taken to mitigate the infection. This proof might consist of the log of an anti-virus scan or some description of the steps taken by the user. To facilitate the cleanup, the walled garden can provide access to a range of white-listed services. Typically these services include free antivirus tools and trusted software suppliers. Other white-list entries may be added to protect critical services for the user, such as webmail services and online banking. Thus customers can perform basic remediation steps and communicate with the abuse desk, even though they are quarantined.

After leaving the walled garden, there is no guarantee that the malware infection was actually remediated. There are several reason by which a user could get out of the quarantine network while being still infected. First of all, certain walled garden implementations allow users to self-release at any time. Normally, this option is only available for the first and perhaps second infection event during a specific period of time. When a user is placed in quarantine for a third time, because of a reinfection or because the earlier infection was not actually removed, the option of self-release

is no longer available. The quarantine removal can now only be executed by the ISP's abuse or support staff. Second, a user can provide erroneous cleanup proofs. For instance, with an increasing number of connected devices in subscriber networks, it is possible for a non-savvy user to perform cleanup actions on a non-infected device and provide the wrong cleanup proofs to the ISP. It is also possible that advanced malware could remain undetected by common antivirus or removal tools. This will allow infected users to leave temporarily the walled garden until the same infection is detected again. Third, some walled garden implementations have an expiration period after which any user in quarantine is released. Fourth, and last, ISP staff might decide to release the user without cleanup. Infected users might request to leave the walled garden for other reasons, like an urgent need for certain online services or because the malware infection cannot be remediated while being in the walled garden. The ISP might allow the user to access the Internet to gather a non-whitelisted cleanup tool.

Our study has been conducted on a walled garden environment deployed for the home users of a medium-sized ISP. Their enterprise and mobile customers are not quarantined. The walled garden follows a strict implementation that redirects users to a landing page (see Appendix A) and limits the access to a set of 41 white-listed websites, including cleanup tools, antivirus solutions, Microsoft updates, webmail providers and online banking. Their implementation of the walled garden provides users with two chances to self-release within a period of 30 days. With the third quarantine action, the option to self-release is revoked and the intervention of the ISP's abuse staff is required. After a period of 30 consecutive days in quarantine, the walled garden automatically releases those quarantined customers who did not self-release or contact abuse staff.

4. DATA COLLECTION

In this section we describe the data that was provided by an ISP to analyze the effectiveness of a particular implementation of a strict walled garden. Our study consists of 1,736 quarantine events associated with 1,208 unique subscribers of a medium-sized European ISP's network during a 6 months period. The data was gathered from four different sources that support the ISP's abuse management process: (*i*) abuse feeds providing security incident data to ISPs; (*ii*) walled garden logs recording details of quarantine events in the ISP's network; (*iii*) help desk logs containing the ISP's help desk communication with customers; and (*iv*) abuse desk communication logs providing email exchange between abuse desk employees and customers.

4.1 Abuse feeds

In order to detect botnet-related infections, the ISP under study leverages abuse feeds provided by the Shadowserver Foundation. For our analysis, we gathered the Shadowserver botnet reports, collected over a time frame of 9 months between April 10th, 2017 and December 30th, 2017. Three different types of reports are analyzed:

- *Drone Reports*: Drone reports contain detailed information on infected machines discovered through monitoring sinkhole traffic, malicious scans and spam relays. We observed a total of 1,620 number of malware infected customers in the network managed by the ISP under review.
- *Sinkhole Reports:* Sinkhole reports contain information about sinkhole servers that did not use the conventional bot signatures such as HTTP referrers. Due to lack of conventional bot signatures, many IP addresses mentioned in this reports do not have a specific infection name. During our study period, we observed 1,598 unique infected users who had a subscription with the ISP under review.
- *Shadowserver's Microsoft Sinkhole:* Microsoft shares via Shadowserver the intelligence gathered from some of their sinkhole servers. Throughout our data collection period, a small number of malicious IP address related to our ISP were captured by Microsoft sinkholes. We only found 8 IP addresses during our study period.

	Sinkhole	MS sinkhole	Drone
# infected users	1,598	8	1,620
% quarantined	22%	63%	59%

Table 1: Infections per feed and quarantined users

As shown in Table 1, we observe a total of 1,620 unique infected users in the Drone feed, 1,598 unique infected users in Sinkhole and 8 unique infected users in MS sinkhole feeds. Not all of these infections trigger a quarantine action, as Table 1 illustrates. There are several reasons why infected users are not quarantined: (i) the user is a mobile or enterprise customer; (ii) the abuse staff decides that quarantining would make matters worse (as in the case of ransomware, where users are by definition already aware of the infection and the lack of Internet access means they might have no viable way to recover their files); (iii) the walled garden environment was undergoing maintenance; and (iv) there are no quarantining actions during the weekend. Figure 2 shows the daily number of unique IP addresses seen in the feeds.

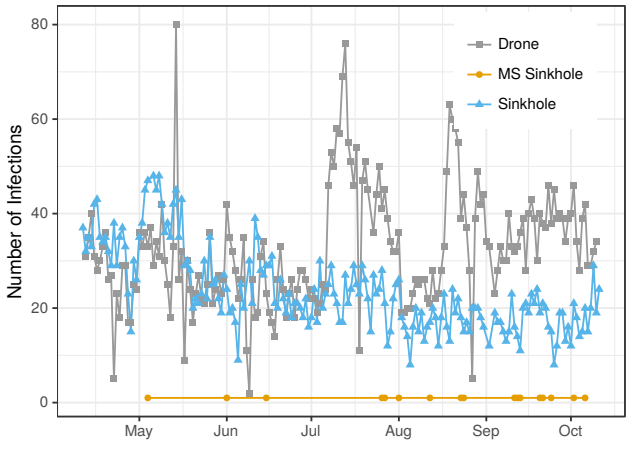


Figure 2: Daily unique infected customers per abuse feed

4.2 Walled garden logs

During our study period, 1,208 retail customers were placed into the walled garden based on the abuse feeds provided by Shadowserver. As some customers were quarantined more than once, this corresponds to 1,736 quarantining events. For each one of these events, several factors were recorded: (i) quarantine time-stamp; (ii) quarantine release mechanism; (iii) quarantine removal time-stamp; (iv) infection type; (v) quarantine event number; and (v) self-release option.

Beside the logs created by the walled garden itself, the quarantined users also have the possibility to submit a form through the walled garden landing page (see Appendix B). This form allows users to explain what cleanup actions they have taken, as well as any other feedback they might have. During the study period, 1,575 forms were received from 831 different infected customers (see Table 2).

	Walled garden form	Abuse desk emails	Help desk phone calls
# Users	831	600	468
# Messages	1,575	2,027	966

Table 2: Messages and users per communication channel

4.3 Help and abuse desks logs

In addition to the walled garden forms (i), customers can also contact the ISP in other ways. We also collected data on (ii) emails between infected customers and the abuse desk; and (iii) phone calls, store visits and social media chat calls between the help desk and the infected customers. Quarantined customers contacted the abuse desk twice as often as the help desk. Table 2 shows that the abuse desk received 2,027 emails, from 600 unique users while help desk employees reported 966 conversations associated with 468 quarantined users.

5. WALLED GARDEN EFFECTIVENESS

We evaluate the impact of the walled garden notification on remediation by looking at the percentage of users that managed to clean the infected machine and at the time an end user remains in the walled garden. We also analyze the relationship between cleanup success and other factors, most notably the type of malware infection, the release mechanism used to get out of the quarantine, and the time spent in the walled garden.

To evaluate cleanup, we distinguish three outcomes when users are released from the walled garden: (i) the user successfully performed cleanup and then stays clean for the rest of the study period; (ii) the user successfully performed cleanup, but the machine is reinfected at a later time in the study period, at least 30 days after the quarantine event; and (iii) the user did not successfully clean up the machine, as evidenced by seeing the offending IP address reported again for the same infection within 30 days of leaving the walled garden.

There is no clear basis for drawing the boundary between a persistent infections and a clean and reinfected machine. Even persistently-infected machines are not seen in the Shadowserver feed every day or even every few days. This depends on a variety of factors, like the malware type and whether the user even turns on the machine. He or she might be on vacation, for example. We decided to count conservatively in terms of cleanup success and use a long period (30 days) before considering the machine clean. Figure 3 shows how these metrics are calculated based on the abuse feeds and the walled garden logs.

There is no clear evidence on where to establish the cut-off point to distinguish persistently infected from clean and reinfected. Figure 4 shows the time between consecutive quarantine events. The median time between quarantine events is 4 days. Roughly 70% of the customers who are seen again after being released from quarantine, are seen within 10 days. As gaps in observations are normal for infected machines, this short interval suggests that these machine were probably not cleaned up. After 20 days, the distribution becomes more or less flat with a slow decay. Choosing a cut-off

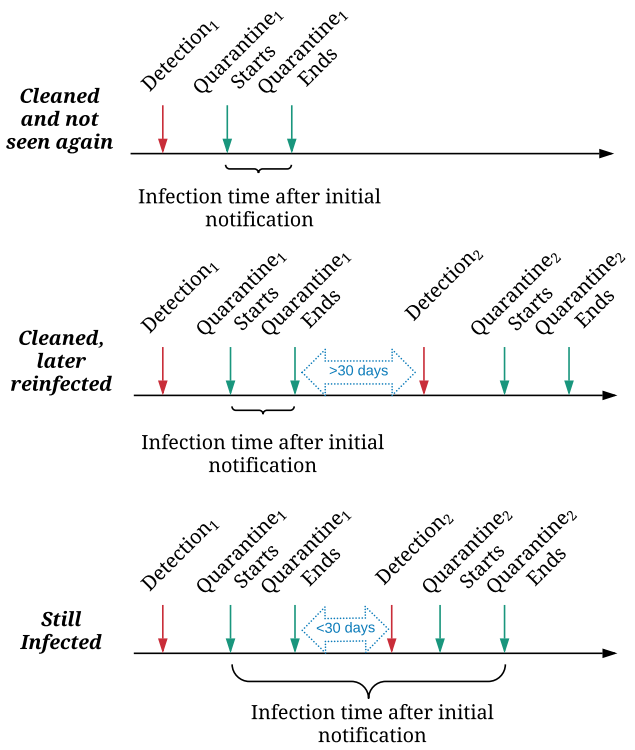


Figure 3: Definition of quarantine outcomes

beyond this point only a modest impact on the results. Reinfection rates would change from 16% (day 20 cut-off) to 13% (day 30) to 7% (day 40). As can be seen in the cumulative distribution, around 13% of the users had a gap between quarantine events of 30 days or more – in other words, these are the users we count as cleaned, but later reinfected.

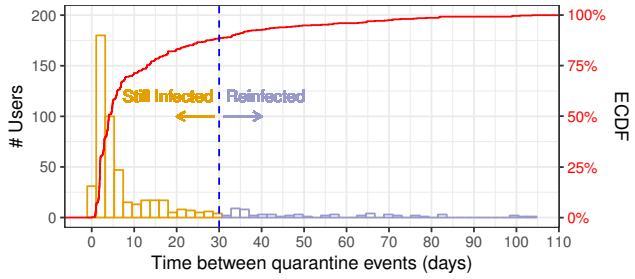


Figure 4: Time between consecutive quarantine events

5.1 Overall remediation rates

In order to understand the effectiveness of the walled garden notifications, we first observe the cleanup and infection rates of the quarantined users after the notifications. We find that 69% of the end users cleaned the infection during their first quarantine event, as shown in Table 3. Another 4% of the clean end users got reinfected with the same malware strain at a later point, more than 30 days after the quarantine event. This suggests they did not correctly address the root cause of the infection. The remaining 27% of users were not able to clean the infection.

Most, but not all, users who remained infected or suffered a rein-

Status	Number of times in quarantine						
	#1	#2	#3	#4	#5	#6	#7
Clean and not seen again	830 (69 %)	148 (49 %)	73 (52 %)	18 (35 %)	17 (65 %)	3 (50 %)	2 (67 %)
Clean and later reinfected	51 (4 %)	13 (4 %)	5 (4 %)	2 (4 %)	1 (4 %)	0	0
Still infected	327 (27 %)	142 (47 %)	61 (44 %)	31 (61 %)	8 (31 %)	3 (50 %)	1 (33 %)

Table 3: Cleanup success over number of times in quarantine

fection, end up in a second quarantine event. Around 20% of them were not quarantined again for a variety of reasons, such as being allowed to leave the quarantine environment to download antivirus solutions. While this makes the infection show up again in the Shadowserver reports, the abuse desk employees withhold the second quarantining action to see if the user is able to resolve it or not.

Of those users who ended up in quarantine for the second time, 49% of them now successfully cleaned up the infection. Again, another 4% also cleaned up, but got reinfected later. Around 47% remained infected. We observed that 139 infected end users ended up in quarantine a third time. This time 56% of them managed to remove the infection, including those who got reinfected later on.

In the tail is a group of users, around 4% of all users who ended up in the walled garden during our study period, who suffered four or more quarantine events. At the extreme end, we found three end users who were put into the walled garden seven times over the course of six months.

Next, we explored the infection time after the initial notification for all quarantined end users. Figure 5 shows the Kaplan-Meier survival curve of the users' infection and the number of remaining infected users every other day. We find that more than 40% of the infected end users cleaned the infection within a day after initial walled garden notification, 70% within 5 days and only 22% remained after a week. After a month time, only 7% of the users remained infected.

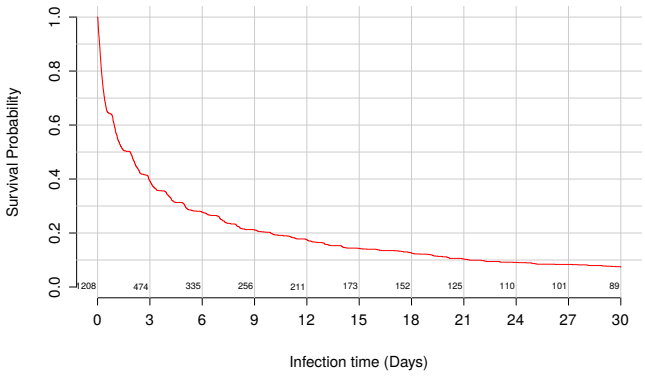


Figure 5: Survival curve of the users' infections

5.2 Malware type

We saw that most of the users in quarantine manage to clean up the infection. Does the complexity of an infection influences their success rate and time it takes them to perform the cleanup? Some malware infections might be harder to resolve than others and the white-listed cleanup tools might not always succeed. To understand the influence of the infection type on the cleanup rates, we use the infection names mentioned in the quarantine event logs. The events were triggered by 38 unique infection types. Table 4 shows the number of users and quarantined events for the top 10 most frequent infection types, which cover 89% of all the users in our dataset.

Infection	# Users	# Quarantine events
Ramnit	444	675
Mirai	275	410
Nymaim	145	159
Downadup	44	65
ZeroAccess	38	51
Rovnix	34	53
Sality-p2p	34	63
Gozi	21	30
Fobber	20	31
Zeus	20	22

Table 4: Number of users and quarantine events per malware

Figure 6 plots the survival curves for these infection types during a 30 days period. We can see significant differences in terms of infection duration for the different infection types (Gehan-Wilcoxon test, $\chi^2 = 58.6$ with $p - value = 2.5e-09$). For instance, end users infected with "Gozi" managed to cleanup all their infections during a 30 days period. On the contrary, cleanup of the more recent "Fobber" and "Rovnix" malware families was slower than the others. One possible explanation is that the more recent malware is more resistant to the standard cleanup tools linked to in the ISP notification [5].

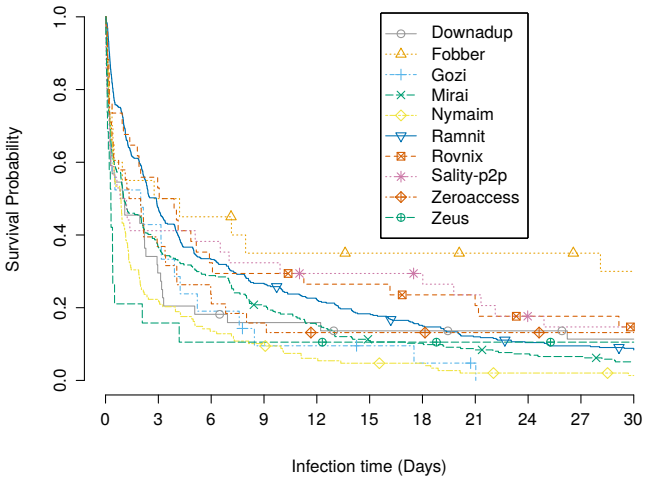


Figure 6: Survival probabilities top 10 infection types during 30 days period

5.3 Release mechanisms

As we mentioned in Section 3, the walled garden contains three mechanisms to release users from the quarantine environment: self-release, assisted release performed by the abuse staff, and quarantine expiry release. Self-release can be used only twice in one

Status	1st Quarantine Event				2nd Quarantine Event				3rd Quarantine Event			
	Total # users	Cleaned, not seen again	Cleaned, later reinfected	Still Infected	Total # users	Cleaned, not seen again	Cleaned, later reinfected	Still Infected	Total # users	Cleaned, not seen again	Cleaned, later reinfected	Still Infected
Self release	805 (67 %)	539 (67 %)	36 (4 %)	230 (29 %)	195 (64 %)	84 (43 %)	9 (5 %)	102 (52 %)	17 (12 %)	5 (29 %)	2 (12 %)	10 (59 %)
Assisted	361 (30 %)	259 (72 %)	11 (3 %)	91 (25 %)	102 (34 %)	61 (60 %)	3 (3 %)	38 (37 %)	114 (82 %)	62 (54 %)	2 (2 %)	50 (44 %)
Expired	42 (3 %)	32 (76 %)	4 (10 %)	6 (2 %)	6 (1 %)	3 (50 %)	1 (17 %)	2 (33 %)	8 (6 %)	6 (75 %)	1 (13 %)	1 (13 %)
Total	1208 (100 %)	830 (69 %)	51 (4 %)	327 (27 %)	303 (25 %)	148 (49 %)	13 (4 %)	142 (47 %)	139 (12 %)	73 (53 %)	5 (4 %)	61 (44 %)

Table 5: Quarantine outcomes per release mechanism

month. If this option is disabled, end users can contact help desk employees or abuse desk employees to get out of the quarantine or to ask for more help. However, before releasing the connection back to normal, employees might require evidence of the cleanup action, such as log files of the antivirus software that was used to remove the infection that triggered the notification.

Is there a relationship between the release mechanism and cleanup success? Since self-release is the fastest and easiest option, one might expect poorer cleanup rates. In the worst case, users simply release themselves without doing anything. To analyze the influence of the release mechanism, we compared the cleanup rates across the first three quarantine actions for all users. As shown in Table 5, the first quarantine action ended with 805 users self-releasing, 361 users following assisted release by abuse staff and 42 users were released when the quarantine period expired after 30 days. Of the 805 self-releasing end users, 67% managed to clean the infection. Another 4% also got cleaned, but was later reinfected. In other words, around 71% of all users managed to perform cleanup. Compare this to the cleanup rate of the users who were released by abuse staff after providing evidence of successful cleanup: 75%. These cleanup rates are very close together. Remarkably, self-release does not invite lax security behavior.

Another surprising finding relates to the 3% of users who remained in quarantine until it expired. They had an even higher success rate: around 86%. We do not have an explanation for this. Perhaps these users were fine with only using the white-listed webmail services and, while remaining in quarantine, automated cleanup tools – e.g., Microsoft's Malicious Software Removal Tool, which is downloaded as part of Windows updates – kicked in at some point.

Users who experienced a second quarantine event chose the self-release option in almost the same proportion (64% versus 67% in the first quarantine event). That being said, cleanup rates are not as high as during the first quarantine. In the self-release group, 48% cleaned up successfully (though 5% later got reinfected). In the provider-assisted release, the cleanup rate is 63%.

During the third walled garden notification period, 82% of the remaining end users ask ISP employees to get them out of the quarantine environment. At this stage, most users no longer get the self-release option, because they were quarantined twice already in one month. Of the users going through assisted release, 54% managed to clean up.

The drop in cleanup rates over successive quarantine events is not large, but might still suggest that perhaps users become habituated and try to get out faster, potentially spending less effort on cleaning and more on getting released. An alternative, and arguable more likely, explanation is that this is caused by selection bias. The users who end up in a second and third quarantine event are likely to be more at risk and perhaps less technically competent. This fits with the fact that with successive quarantine events, the cleanup effectiveness of the assisted-release users become slightly higher compared to the self-release group.

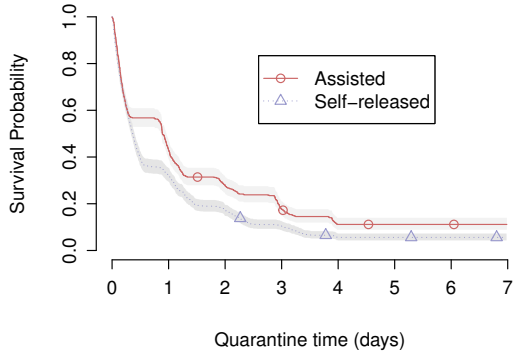


Figure 7: Survival probabilities per release mechanism

Figure 7 shows the duration of all infections per release mechanism in the form of Kaplan-Meier survival curves. As expected, users that needed assistance to cleanup their infections left the walled garden at a slower rate than the users that self-released. Looking at the speed at which they got removed from the quarantine, we can observe significant differences between these two groups (Gehan-Wilcoxon test, $\chi^2 = 23.1$ with $p - value = 1.5e-06$). For instance, within the first 2 days in quarantine, 84% of the users that self-released left the walled garden while only 71% of users that needed assistance did so.

5.4 Time spent in the walled garden

We now take a closer look at the time users spend in quarantine. Figure 8 displays the distribution of the duration of the quarantine events. The majority of quarantine events lasted less than one day and only 25% of them lasted more than 3 days. A small fraction (57 events) last until they automatically expire.

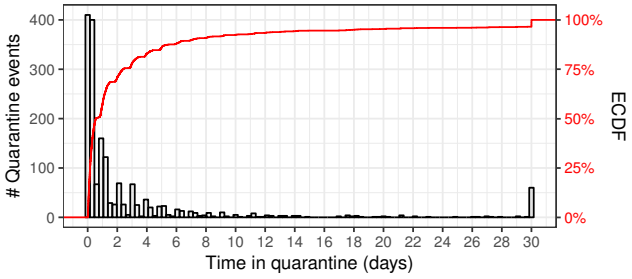


Figure 8: Histogram and cumulative density function of the quarantine period

Figure 9 displays the survival probability curves of users in terms of time spent in the quarantine environment for the first three quarantine events and the rest. As demonstrated in Figure 9, end users spent more time in quarantine during their first time than the second time. This might be due to being unfamiliar with the environment or with the process to clean up the infection.

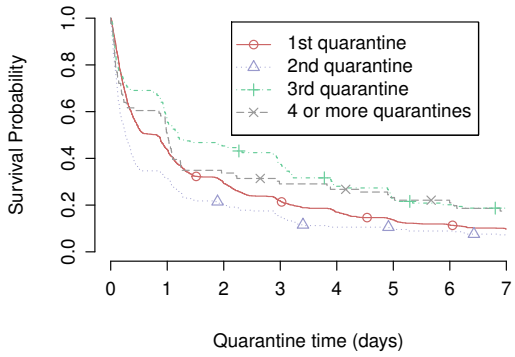


Figure 9: Survival probabilities over different quarantine events

To further investigate, Table 6 shows the median time spent in quarantine during the first three quarantine events. We compare them across the different release mechanisms and cleanup outcomes. End users that managed to remove the infection, stayed longer in the walled garden than those who remained infected, regardless of which release mechanism was used. Take a look at the median time of the assisted end users in the first quarantine event, for example. Those who managed to clean up spent 24 hours in quarantine, while users who remained infected took just around 7 hours. In the self-release group, successful cleanup also took longer, though for the first quarantine event, the difference is surprisingly small with the group that remains infected or got re-infected at a later stage (roughly 11 versus 10 hours).

During the second and third quarantine event, the differences become more pronounced. Longer time spent in quarantine is now clearly related to cleanup success. Users who remain infected spend about half as long in quarantine as the other two groups. It seems a certain group of users is becoming habituated to the walled garden notification and environment. They self-release very quickly and it seems unlikely that they made a serious attempt to perform cleanup.

It is important to note, though, that the self-releasing users that do succeed in cleaning up also leave the walled garden faster over successive quarantine events. The median time drops from 11 hours during the first quarantine to 8 hours (second quarantine) and then to just 3 hours (third quarantine). In other words, it seems there is not just habituation going on, but also actual positive learning effects in terms of how to perform cleanup and navigate the release from the walled garden.

6. END USER REACTIONS

To get a better sense of the actual experience of the end users, we qualitatively analyzed the communication of the quarantined users with the abuse and support staff at the ISP. Each communication channel was used for different of reasons. Generally, emails were sent to inform abuse desk employees about the cleanup efforts and possible causes of the infection. Interaction with the support staff, on the other hand, were more often asking for more information about the quarantine and how to resolve the situation. The content of the submitted walled garden forms often contained more specific information on the cleanup actions taken by the quarantined users. For instance, some users pasted the output of the antivirus scans in these forms to prove that the infection was no longer present.

First, we manually analyzed a sample of 200 walled garden forms, 200 help desk logs and 50 emails to the abuse desk. We saw five recurring themes that speak to the user experiences of the walled garden: (i) asking for additional help to resolve the infection and leave the walled garden; (ii) requesting a paid technician to visit the user; (iii) expressing distrust of the walled garden notification; (iv) complaining about the disruption of service; and (v) threatening to terminate the contract with the ISP. To get a sense of how many users were associated with these types of communications, we collected keywords from the manual analysis of the sample and then searched the full communication data for their presence. Table 7 shows the number of unique users associated with each topic. For 51% of the users who communicated with the ISP, their messages did not fit any of these topics and we categorized them as 'Miscellaneous'.

6.1 Requesting additional help

Almost 27% of the users at some point contacted the ISP to ask for additional help to cleanup the infection. The users wanted to solve the problem, but they were unable to understand the notification or to follow the steps towards quarantine release. The type of help that is requested varies widely. Some of this is driven by differences in the type of infection and the operating system of the user. Cleanup software and materials provided in the notification content would not work on all OS types, OS versions and patch levels. Some customers in our study downloaded the requested software to remove the infection, only to find out that it would not install correctly. Some users could not download the software at all from the links provided by the ISP. In those cases, they requested to be released from the quarantine environment so that they could download additional software.

One of the malware families was Mirai – the infamous botnet made up of Internet-of-Things devices. Not surprisingly, users with these infections asked for help in identifying which of their many devices was the problem and how to then secure it from future infections. Not to put too fine a point on it, but from a usability perspective the cleanup of compromised IoT is a world of pain for which we have very little practical guidance. In these cases, ISP staff would ask users additional questions about what devices they had connected to their home network. Based on the replies, staff would try to identify the offending device and more specific cleanup actions. In one case, after contacting the ISP, a user disconnected his IP camera from the network so as to prevent future infections and quarantine events, while the actual problem later turned out to be a DVR. The user ended up getting infected and quarantined again.

6.2 Requesting a paid technician

About 7% of the users in our study were not capable of removing the infection by themselves and requested the ISP to send a paid technician to their home. In a handful of cases, end users mentioned taking their computer to technicians at local computer repair shops. The ISP's technicians are typically people who also have a background in abuse handling. Some of the communications we analyzed were from these technicians themselves who contacted their colleagues at the ISP abuse department from the customer premises and provided detailed information about their cleanup actions. This way, the abuse desk employees got the required proof of cleanup and could release the connection from the walled garden. Interestingly, in a few rare cases, we found that the paid technician could not actually find the infection. They then referred the end users back to abuse desk employees to communicate the occurrence of a false positive. Unfortunately, as a result of this process, users remained in the walled garden environment longer.

Release mechanism	1st Quarantine event				2nd Quarantine event				3rd Quarantine event			
	Total # users	Median quarantined time (hours)			Total # users	Median quarantined time (hours)			Total # users	Median quarantined time (hours)		
		Cleaned, not seen again	Cleaned, later reinfected	Still Infected		Cleaned, not seen again	Cleaned, later reinfected	Still Infected		Cleaned, not seen again	Cleaned, later reinfected	Still Infected
Self release	805	11.16	10.69	10.24	195	8.07	7.04	3.74	17	3.15	11.42	3.29
Assisted	361	24.25	4.90	6.82	102	24.87	25.28	3.30	114	69.72	49.60	22.51

Table 6: Summary statistics on the time to cleanup for self released and ISP assisted released mechanisms

6.3 Distrust of the notification

Around 2% of the users contacted ISP employees to confirm the veracity of the email and walled garden notifications. They did not expect that their ISP would notify them about an infection and were worried that this could be a phishing attack to install ransomware or steal personal information. Users mainly contacted help desk employees to confirm the veracity of the notifications. One user replied directly to the notification email, i.e., using the very channel that he did not trust, and voiced his concerns this way to the recipient at the abuse department.

6.4 Complaints over disruption of service

Placing a customer in a walled garden environment is a strong incentive for end users to clean up, but also an intrusive measure. During in our study period, around 10% of the users complained in some shape or form. Some reported that their business was disrupted due to having no Internet to work with. Usually, these turned out to be users that run small businesses over their consumer broadband connection: shops, restaurants and even a small medical clinic. They claimed that they could not provide services to their customers and, as a consequence, lost customers. Some mentioned, for example, that the payment terminals did not work and so their customers could not complete their purchases. In two cases, the owner of the shop stated he had to close the shop until the problem was fixed. Several of these users provided a calculation of the monetary loss they suffered and demanded a reimbursement from the ISP.

6.5 Threats to terminate the contract

Around 3% of the users were so unhappy about their connection getting quarantined that they threatened to terminate their subscription and move to one of the ISP's competitors. Some of the users pointed to the losses they had incurred, others to the fact that they had to pay for the subscription even though they no longer were provided with Internet access. Also several users threaten to leave the ISP because the user could not, even with their best effort, identify and remove the infection. These users were quarantined multiple times and they spent quite a bit of time in the walled garden environment.

Topics	# of users
Request additional help	323 (27 %)
Request paid technician	80 (7 %)
Distrust of the notification	19 (2 %)
Complain over disruption of service	126 (10 %)
Threaten to terminate the contract	39 (3 %)
Miscellaneous	621 (51 %)

Table 7: User issues raised in communication with ISP

7. ETHICAL CONSIDERATIONS

Access to data about the user's experience upon abuse notifications is extremely limited and cooperation with an ISP is essential to enable otherwise impossible research. For this study we leverage secondary data that was originally collected by an ISP for business purposes. This data was pre-processed by a coauthor of this paper while working for this ISP and with the consent of the ISP's abuse desk manager. Moreover, the data was processed on the ISP premise and within the ISP privacy policies.

Unavoidably, the processed dataset was not fully anonymized as the high dimensionality of the data did not allow for a robust anonymization, i.e., the anonymization would have led to an unacceptable level of data loss. To ensure confidentiality, the raw dataset was stored in a secure server to which only authorized users could access. Moreover, the data was analyzed while preserving the privacy of the ISP's customers and ensuring that it is not possible to identify them from any of our results. Both the processed and anonymized data were removed after the publication. The original data remains in the ISP systems, allowing for replication if needed.

8. LIMITATIONS

We underline four limitations relevant to the findings of our study. First, we based our study on a single ISP with a relatively strict implementation of the walled garden notification system. The generalizability of our results to other implementations and ISPs is a matter for further studies. Second, our study uses data collected as part of the operational process of the ISP. As such, the study lacks an experimental design and a control group. This means we cannot compare the effectiveness of the walled garden notification to the cleanup rate of a mere email or no notification whatsoever. Third, our dataset on infections is limited to what has been reported in the Shadowserver feeds. As a result of this, we lack visibility into notifications triggered by other feeds and infections that are not reported by Shadowserver. This makes our coverage of malware infections biased towards those that are sinkholed and reported by Shadowserver. Malware that has escaped these defender efforts might also be harder to clean. Fourth, the cleanup outcomes are also based on the Shadowserver feeds. It is possible that an infection might not show up in the Shadowserver feeds right away. This is partly driven by user behavior, such as temporarily turning off the infected device or disconnecting it from the Internet, and partly by other factors, such as the properties of the malware families. Some are less aggressive in terms of scanning for victims or contacting the command-and-control server for commands. This absence in the feed may cause us to overestimate the cleanup rate. For this reason we chose a conservative time frame. We only counted a machine as cleaned up if we did not see it for 30 days after release from the walled garden.

9. CONCLUSION

In this study, we explored the effectiveness of walled garden notifications and quarantining in terms of helping users in residential networks to perform malware cleanup. Based on data on 1,736 quarantining actions involving 1,208 unique users, collected from April 2017-October 2017 by a medium-sized European ISP, we found that roughly half to three quarters of the quarantined users had managed to clean their machine. There is no clear point of reference for this success rate. When we look at prior work on abuse and vulnerability notifications, it seems to be quite high. Most of

those studies find rates well below 50%. That being said, comparison is difficult as the typical recipient of those notifications is a server admin or webmaster, not a home user.

Most users are quarantined only once, so the effort of cleanup kept them clean for months, if not longer. Perhaps the quarantine experience made users adapt their online behavior or improve their system's security defaults, like automatic patching and the installation of antivirus tools. This suggests there may also be long-term benefits to quarantining, beyond mitigating the immediate threat posed by the infection.

Users could self-release easily and quickly for the first two quarantine events in a month. Remarkably, this easy way out does not incite lax security behavior. Cleanup rates are either as high, or just a bit lower, than users who have to submit proof of cleanup to the provider and wait for the abuse staff to release them. We see a bit of evidence for habituation among a small group of users who learn how to release themselves from quarantine, rather than clean the infection. We also saw evidence, however, of a positive learning effect: successful cleanup also became faster for users going through successive quarantining events.

All in all, we found substantial support for the effectiveness of this best practice for ISPs in the fight against botnets. Since effectiveness of the other recommended best practices has been questioned, this suggests more ISPs should be considering to adopt a walled-garden solution. In light of the rising problem with IoT malware, this might become a critical line of defense. That being said, IoT malware remediation methods will differ from traditional cleanup strategies and, thus, walled garden implementations will have to be revisited to accommodate the cleanup requirements for IoT malware.

On the downside: setting up and maintaining a walled garden environment is a significant investment for an ISP. Furthermore, providing support to users in their attempts to clean up also imposes a significant cost. Around one out of four quarantined users posed a question for help to a staff member. These costs could perhaps be reduced by allowing self-release more broadly, since it seems to be more or less equally effective as the more labor-intensive form of provider-assisted release. Some of this assistance might provide a business opportunity, as we found that around 7% of the quarantined users asked for a paid technician.

A fraction of the users, around 10% of them, voiced complaints over the disruption. Around 3% even threatened to terminate the contract. We do not know how many users actually terminated their subscription, but the threat alone might, unfortunately, be enough to scare off some ISPs from investing in a walled garden. In competitive broadband markets with high penetration rates, customer acquisition is very expensive. In these situations, a prisoner dilemma might appear as not having a walled garden might be a competitive advantage. This could push ISPs to not deploy it, even though it is effective. On the other hand, if all ISPs adopted it simultaneously, it would generate collective benefits, though these would not necessarily flow back to the ISP, except through lower customer churn rates.

We did notice that the group which seemed the most negative about the quarantining actions were small businesses operating on a consumer broadband connection. ISPs could prevent them from being affected in the future by providing an easy transition to a comparatively-priced business subscription, which would take them out of the consumer market – and thus keep them away from the walled garden. This would reduce the pushback over time and allow the walled garden to do what it does best: protecting home users from further damage caused by their infection, and protecting the rest of the Internet from the infected home user.

10. ACKNOWLEDGMENT

This publication was supported by a grant from the Netherlands Organisation for Scientific Research (NWO), under project number 628.001.022. Also, we would like to thank the anonymous reviewers, Folkert Visser and Dennis van Beusekom for their helpful comments.

11. REFERENCES

[1] Do malware reports expedite cleanup? an experimental study. In *Presented as part of the 5th Workshop on Cyber Security Experimentation and Test*, Bellevue, WA, 2012. USENIX.

[2] H. Almuhimedi, A. P. Felt, R. W. Reeder, and S. Consolvo. Your reputation precedes you: History, reputation, and the chrome malware warning. In *Symposium on Usable Privacy and Security (SOUPS)*, volume 4, page 2, 2014.

[3] H. Asghari, M. Ciere, and M. J. Van Eeten. Post-mortem of a zombie: conficker cleanup after six years. In *USENIX Security Symposium*, pages 1–16. USENIX Association, 2015.

[4] H. Asghari, M. J. van Eeten, and J. M. Bauer. Economics of fighting botnets: Lessons from a decade of mitigation. *IEEE Security & Privacy*, 13(5):16–23, 2015.

[5] P. Black, I. Gondal, and R. Layton. A survey of similarities in banking malware behaviours. *Computers & Security*, 2017.

[6] C. Bravo-Lillo, L. Cranor, S. Komanduri, S. Schechter, and M. Sleeper. Harder to ignore. *Revisiting pop-up fatigue and approaches to prevent it, USENIX Association*, pages 105–111, 2014.

[7] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs, and S. Schechter. Your attention please: designing security-decision uis to make genuine risks harder to ignore. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, page 6. ACM, 2013.

[8] D. Canali, D. Balzarotti, and A. Francillon. The role of web hosting providers in detecting compromised websites. In *Proceedings of the 22nd international conference on World Wide Web*, pages 177–188. International World Wide Web Conferences Steering Committee, 2013.

[9] O. Cetin, C. Ganán, M. Korczynski, and M. van Eeten. Make notifications great again: learning how to notify in the age of large-scale vulnerability scanning. In *16th Workshop on the Economics of Information Security (WEIS 2017)*, 2017.

[10] O. Cetin, M. Hanif Jhaveri, C. Gañán, M. van Eeten, and T. Moore. Understanding the role of sender reputation in abuse reporting and cleanup. *Journal of Cybersecurity*, 2(1):83–98, 2016.

[11] ECO Internet industry association. Botfree. `https://www.botfree.eu/en/aboutus/information.html`, 2013.

[12] European Network and Information Security Agency (ENISA). Involving Intermediaries in Cyber-security Awareness Raising. `https://www.enisa.europa.eu/publications/involving-intermediaries-in-cyber-security-awareness-raising`, 2012.

[13] M. Fagan and M. M. H. 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.

[14] A. Forget, S. Pearman, J. Thomas, A. Acquisti, N. Christin, L. F. Cranor, S. Egelman, M. Harbach, and R. 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.

[15] International Telecommunication Union (ITU). ITU Botnet Mitigation Toolkit. http://www.itu.int/ITU-D/cyb/cybersecurity/projects/botnet.html, 2007.

[16] International Telecommunication Union (ITU). ITU Botnet Mitigation Toolkit. https://www.itu.int/ITU-D/cyb/cybersecurity/projects/botnet.html, 2018.

[17] Jilani, Umair. The ACMA and Internet providers working together to combat malware. https://www.acma.gov.au/theACMA/engage-blogs/engage-blogs/Cybersecurity/The-ACMA-and-internet-providers-working-together-to-combat-malware, 2015.

[18] S. Kim and M. S. Wogalter. Habituation, dishabituation, and recovery effects in visual warnings. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 53, pages 1612–1616. Sage Publications Sage CA: Los Angeles, CA, 2009.

[19] K. Krol, M. Moroz, and M. A. Sasse. Don't work. Can't work? Why it's time to rethink security warnings. In *Risk and Security of Internet and Systems (CRiSIS), 2012 7th International conference on*, pages 1–8. IEEE, 2012.

[20] M. Kührer, T. Hupperich, C. Rossow, and T. Holz. Exit from Hell? Reducing the Impact of Amplification DDoS Attacks. In *USENIX Security Symposium*, 2014.

[21] F. Li, G. Ho, E. Kuan, Y. Niu, L. Ballard, K. Thomas, E. Bursztein, and V. Paxson. Remedying Web Hijacking: Notification Effectiveness and Webmaster Comprehension. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1009–1019. International World Wide Web Conferences Steering Committee, 2016.

[22] J. Livingood, N. Mody, and M. O'Reirdan. Recommendations for the Remediation of Bots in ISP Networks (RFC 6561). *Internet Eng. Task Force*, 2012.

[23] A. Mathur, J. Engel, S. Sobti, V. Chang, and M. Chetty. They Keep Coming Back Like Zombies: Improving Software Updating Interfaces. In *SOUPS*, pages 43–58, 2016.

[24] Messaging Anti-Abuse Working Group and others. Abuse Desk Common Practices. https://www.m3aawg.org/sites/default/files/document/MAAWG_Abuse_Desk_Common_Practices.pdf, 2007.

[25] Messaging Anti-Abuse Working Group and others. M3AAWG best practices for the use of a walled garden. https://www.m3aawg.org/documents/en/m3aawg-best-common-practices-use-walled-garden-version-20, 2015.

[26] National Institute of Standards and Technology. Models To Advance Voluntary Corporate Notification to Consumers Regarding the Illicit Use of Computer Equipment by Botnets and Related Malware. https://www.nist.gov/itl/upload/SANS_BotNet-FRN-Comment-11-4-11.pdf, 2011.

[27] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. The emperor's new security indicators. In *Security and Privacy, 2007. SP'07. IEEE Symposium on*, pages 51–65. IEEE, 2007.

[28] B. Stock, G. Pellegrino, F. Li, M. Backes, and C. Rossow. Didn't You Hear Me?–Towards More Successful Web Vulnerability Notifications. In *The Network and Distributed System Security Symposium (NDSS)*, 2018.

[29] B. Stock, G. Pellegrino, C. Rossow, M. Johns, and M. Backes. Hey, you have a problem: On the feasibility of large-scale web vulnerability notification. In *USENIX Security Symposium (Aug. 2016)*, 2016.

[30] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor. Crying wolf: An empirical study of ssl warning effectiveness. In *USENIX security symposium*, pages 399–416, 2009.

[31] R. Wash, E. Rader, K. Vaniea, and M. Rizor. Out of the loop: How automated software updates cause unintended security consequences. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 89–104, 2014.

APPENDIX
A. WALLED GARDEN LANDING PAGE

Secure environment

A safe Internet is in everyone's interest. We strongly care about protecting your (confidential) information.

We have received information from one of our partners that a security issue has been detected on your Internet connection. You probably have not noticed anything yet.

Don't worry. To protect you against the security risks we have placed your Internet connection in our secure environment. In this environment you can safely solve the security issues. We are willing to help you to do so.

What is the problem and how can you solve it?
One or more computers connected to your Internet connection are infected with a virus.

We kindly ask you to follow the steps to remove viruses on all computers/laptops as described on:

https://address.com

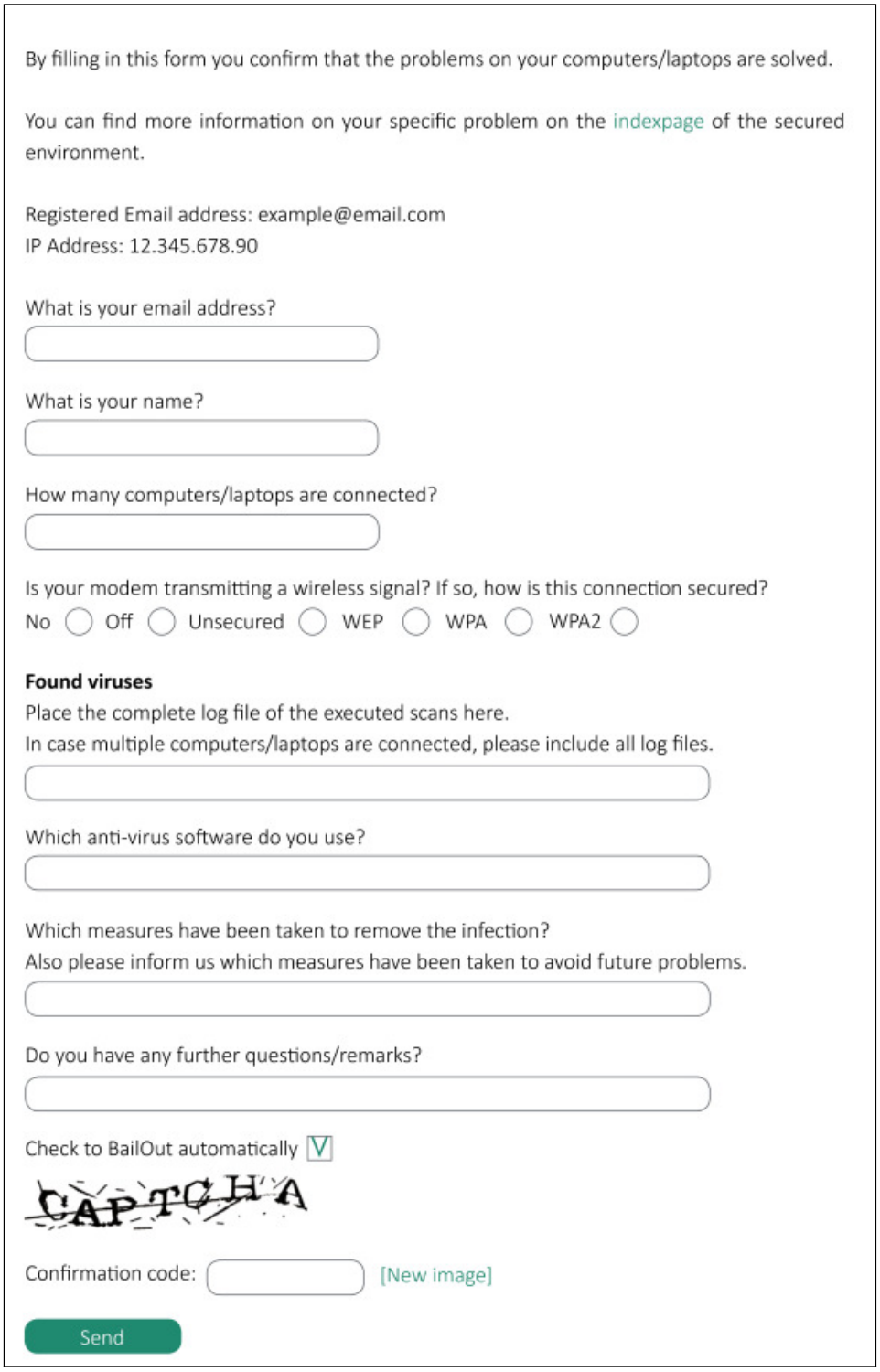
When the scan has finished the program will create a log file with the scan results. Please send us the content of this log file(s).

We would like to be informed what measures have been taken to make sure this abuse will not take place again.

Necessary steps
1. Take the measures stated above
2. Fill in our form (and restore your Internet connection)

General security tips
* Use an up-to-date virus scanner to keep out potential hazards
* Keep computer software, like your operating system, up to date
* Do not open messages and unknown files that you do not expect or trust
* Secure your wireless connection with a unique and strong password

B. WALLED GARDEN RELEASE FORM



By filling in this form you confirm that the problems on your computers/laptops are solved.

You can find more information on your specific problem on the indexpage of the secured environment.

Registered Email address: example@email.com
IP Address: 12.345.678.90

What is your email address?

What is your name?

How many computers/laptops are connected?

Is your modem transmitting a wireless signal? If so, how is this connection secured?
No ◯ Off ◯ Unsecured ◯ WEP ◯ WPA ◯ WPA2 ◯

Found viruses
Place the complete log file of the executed scans here.
In case multiple computers/laptops are connected, please include all log files.

Which anti-virus software do you use?

Which measures have been taken to remove the infection?
Also please inform us which measures have been taken to avoid future problems.

Do you have any further questions/remarks?

Check to BailOut automatically ☑

CAPTCHA

Confirmation code: [New image]

Send

Developers Deserve Security Warnings, Too

On the Effect of Integrated Security Advice on Cryptographic API Misuse

Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke*,
Christian Stransky*, Sebastian Moeller†, Yasemin Acar*, Sascha Fahl**
Cologne University of Applied Sciences, *Leibniz University Hannover,
†Quality and Usability Lab, Technical University Berlin, **Ruhr-University Bochum

ABSTRACT

Cryptographic API misuse is responsible for a large number of software vulnerabilities. In many cases developers are overburdened by the complex set of programming choices and their security implications. Past studies have identified significant challenges when using cryptographic APIs that lack a certain set of usability features (e. g. easy-to-use documentation or meaningful warning and error messages) leading to an especially high likelihood of writing functionally correct but insecure code.

To support software developers in writing more secure code, this work investigates a novel approach aimed at these hard-to-use cryptographic APIs. In a controlled online experiment with 53 participants, we study the effectiveness of API-integrated security advice which informs about an API misuse and places secure programming hints as guidance close to the developer. This allows us to address insecure cryptographic choices including encryption algorithms, key sizes, modes of operation and hashing algorithms with helpful documentation in the guise of warnings. Whenever possible, the security advice proposes code changes to fix the responsible security issues. We find that our approach significantly improves code security. 73% of the participants who received the security advice fixed their insecure code.

We evaluate the opportunities and challenges of adopting API-integrated security advice and illustrate the potential to reduce the negative implications of cryptographic API misuse and help developers write more secure code.

1 Introduction

A large number of software vulnerabilities are caused by developers who misuse security APIs [12,19,36]. Previous work identified multiple trouble spots including secure network connections [15], the use of permissions in mobile apps [17] and the use of cryptographic APIs [1,32]. Some of the most serious data breaches in recent history were caused by not properly using TLS to secure data in transit or not securely storing data in rest [12,19]. Such incidents affect millions or even billions of users worldwide and jeopardize their security and privacy.

In this work we focus on the challenges of using cryptographic APIs securely. Using cryptographic APIs correctly in many cases requires detailed knowledge and overburdens non-security expert developers on a regular basis. Acar et al. conducted several studies and investigated the usability of cryptographic APIs and the impact of information resources developers use to solve programming questions on code security [1, 2]. They find that the design of cryptographic APIs and the quality of available developer documentation amongst other factors have a significant impact on code security. In particular, the availability of easy-to-understand documentation and ready-to-use and functional code snippets helped participants in their studies to produce more secure results. Motivated by their findings and results of measurement studies of real world software repositories [2], we design and implement a novel approach to help software developers write more secure cryptographic code.

While there is previous work that tries to improve code security by enhancing API simplicity [23] or by providing IDE plugins [29, 34], we propose a different and novel approach that allows providers of existing and future cryptographic APIs to improve code security. Therefore, they do not have to change their programming interfaces, rely on the development and integration of plugins for integrated development environments (IDEs) or hope that security of unsafe information sources such as Stack Overflow becomes better. Instead, we propose the integration of effective security advice directly into cryptographic APIs. We develop an API-integrated security advice concept that provides context sensitive help and offers ready-to-use and secure code snippets to fix security issues. We implement our approach for Python and the PyCrypto cryptographic API and conduct a between-subjects online study with 53 experienced Python developers. In the course of this study we try to answer the following research questions:

RQ1: *Does API-integrated security advice have a significant effect on code security?* With this research question we try to assess the ability of our approach to improve code security. We analyze all changes made to the code after security advice has been shown. We find that our approach had a significant positive impact on 73% of our participants who left their code insecure at the first place: They upgraded bad cryptographic choices to secure ones.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

RQ2: *Does API-integrated security advice have a significant impact on perceived API usability?* We were interested in whether providing context sensitive security advice affects the perceived usability of the PyCrypto API. We find that while security significantly improved, the security advice had no statistical significant impact on the perceived usability.

RQ3: *How does our approach compare to other approaches?* Previous work by Acar et al. [1, 2] found that interfaces and supported use cases of cryptographic APIs and the types of information resources developers use have a significant impact on code security. Nguyen et al. [34] tested their Android studio plugin and found that their approach has a significant contribution to code security.

We find that similar to high quality developer documentation, good API design and helpful IDE plugins, our approach has a significantly positive effect on code security, but allows API providers to improve code security for existing cryptographic APIs in a low-level approach without having to change API design or relying on third party tools.

Our work makes the following contributions:

1. We design a security advice concept that is directly integrated into an API, drawing on guidelines, suggestions and research on human factors on security APIs and warning messages.
2. We implement our concept for the PyCrypto API.
3. We conduct a between-subjects online controlled experiment with experienced Python developers to test the effectiveness of our approach.
4. We assess the real world applicability, limitations and potential of API-integrated security advice, and conclude with lessons learned from our experiment.

2 Related Work

We discuss related work in two key areas: research on human factors on security APIs and tools for developers; research on security warning messages.

Research on Security APIs and Tools: Researchers have investigated challenges developers have when interacting with security APIs and tools.

Both Wurster and Oorschot [41] and Green and Smith [22] analyze the developers' roles in writing secure software and come to the conclusion that security is often only of secondary or tertiary concern for developers and that (security) APIs and libraries need to be designed with usability in mind. Lo Iacono and Gorski [30] present a classification of security APIs according to their abstraction level. They evaluate their approach by investigating a set of popular software development kits and conducting an online study with developers. They find that developers prefer APIs that provide comfortable abstractions and enable them to take full control as required by the specific programming task. Gorski and Lo Iacono propose a set of eleven characteristics to evaluate security API usability as they find that security API usability goes beyond general API usability [21].

Nadi et al. manually examined the top 100 Java cryptography posts on Stack Overflow and found that a majority of problems were related to API complexity rather than a lack of domain knowledge [32]. Relatedly, Acar et al. investigated how the use of different documentation resources affects developers' security decisions, including decisions about certificate validation. They report that good usability and the availability of ready-to-use and functional code snippets as part of documentation significantly impacts code security [2]. Barik et al. [6] conducted an eye-tracking study to investigate the use of Java compiler error messages finding that the difficulty of reading error messages is comparable to reading source code. Acar et al. conducted a controlled experiment online and compared the usability of different cryptographic libraries for Python [1]. They found that in addition to safe defaults, the number of supported use cases and the availability of good documentation have a significant impact on code security. Naiakshina et al. conducted a qualitative developer study and investigated how computer science students implemented secure password storage [33]. We develop and test a novel approach that supports developers using a security warning that presents context-sensitive documentation and code snippets as part of an API.

Nguyen et al. present a plugin for the Android Studio IDE called FixDroid which helps developers write more secure code by highlighting insecure code and providing quick fixes. In a user study they find that FixDroid users write significantly more secure code than participants without FixDroid [34]. Similarly, Krueger et al. present the CogniCrypt tool which is an Eclipse IDE plugin for the Java programming language that helps developers to securely use cryptographic APIs by auto-generating secure code for common tasks [29]. Xie et al. present and evaluate an Eclipse IDE called ASIDE that interactively reminds programmers of secure programming practices [42]. Johnson et al. conducted a user study and investigated why developers do not use static analysis tools to find bugs and report that too many false positives and complicated errors messages were significant hurdles for their participants [25]. We propose an IDE-agnostic approach that allows API and library providers to improve code security without having to rely on third parties such as IDE plugins or static code analysis tools.

To the best of our knowledge, in contrast to previous work our paper is the first to introduce and study security advice as part of an API.

Research on Security Warnings: Researchers have investigated challenges in designing usable security warnings. Due to a lack of related work for software developers we limit the following presentation to previous work for warnings for end-users.

Sunshine et al. conducted multiple studies to investigate the effectiveness of SSL warnings and found while they could improve warning message effectiveness still many participants clicked-through a warning. In addition to further improve warnings, they recommend to reduce their occurrence [38]. Felt et al. experimented with SSL warnings for Google Chrome and found that while they could not improve the rate of comprehension of the warnings' text significantly, opinionated design drastically improved the warnings' adherence rate [4, 16, 18].

Weinberger and Felt run a field study to investigate how long the Chrome browser should store users' decisions for SSL warnings to minimize the effect of habituation [40]. Sim-

ilarly, Vance et al. conduct an fMRI experiment to study warning message habituation [39]. Both studies conclude that the risk of habituation decreases after one week.

Almuhimedi et al. investigate factors that contribute to why Chrome users click-through their malware warnings and find that familiarity with a website had significant impact on users' click-through behavior [5]. Egelman et al. investigated the difference between passive and active warnings against phishing attacks and found that active warnings were more successful [13]. Bravo-Lillo et al. designed and tested multiple attractors for security warnings [8].

Bauer et al. present and discuss a set of design guidelines for warning messages [7]. Our approach follows their guidelines and includes lessons learned from other related work presented above.

In contrast to end-users, our work is the first to investigate a novel security warning concept targeted at software developers.

3 API Level Advantages

Making security advice part of the API has multiple advantages over other approaches:

Environment Agnostic: Integrating security advice into an API instead of providing extensions or plugins for integrated development environments (IDEs) or editors (e. g. [29, 34]) makes the security warnings agnostic to developers' programming environments. Instead of having to provide multiple extensions or plugins for different programming environments only one implementation for a particular API is needed. API integration is not just agnostic to the IDE or editor used but also to the way developers use programming language interpreters or compilers. Security warnings that are part of an API can provide helpful information in terminal as well as in IDE environments.

Immediate Feedback: Making security advice part of an API can provide context sensitive and secure information (e. g. secure code snippets or targeted information) as immediate feedback. Such an approach has the potential to prevent developers from falling back on insecure information resources online such as Stack Overflow [2].

A Bottom Up Approach: An integrated feedback mechanism gives API providers the power to provide very specific and context sensitive security advice and make it available through the regular distribution channels of an API. API users immediately benefit from feedback integration after using an updated API version. Instead of having to install or update external third party tools such as plugins or extensions, relying on the regular update channels of an API has the potential to speed up distribution of feedback mechanisms.

4 Security Advice Design

Below we discuss design decisions for our security warning.

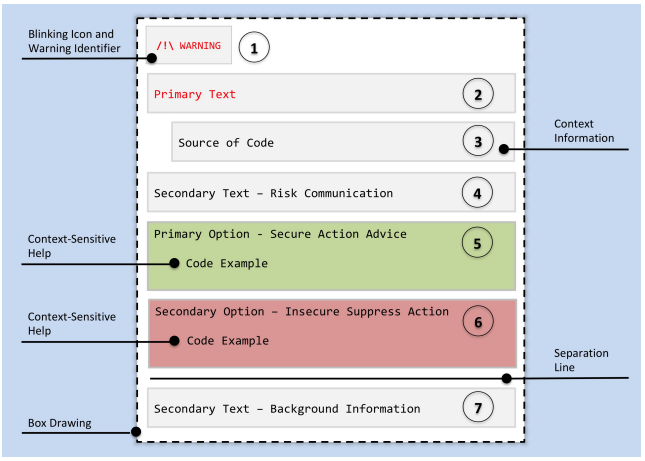


Figure 1: Design concept of our security feedback mechanism.

4.1 Design Decisions

The main goal of our approach is to help users of security APIs to avoid insecure programming choices whenever possible. However, due to the complex nature of security decisions and the fact that information security is not a top priority [30] for many developers, we expected this to be challenging.

There is no previous work on designing security warnings with a specific focus on making secure programming choices. Therefore, we based our work on previous research and lessons learned from warning message design for end-users. Especially, the warning design guidelines by Bauer et al. [7] provide a comprehensive list of principles for designing security warnings with a focus on end-users. We rely on their general and abstract principles for designing developer centered security warnings. Thus we adapted design goals from their guidelines and applied them to an API-integrated security advice concept. Additionally, we considered lessons learned from previous secure programming studies with software developers [1–3, 6].

Figure 1 illustrates the design concept of the security feedback mechanism we present below. Although we contribute a design concept based on existing principles and lessons learned, we do not comparatively evaluate different design approaches in our work. We expect this to be future work.

Goal 1: *Follow a Consistent Layout.* In contrast to security warnings for graphical user interfaces, an API-integrated security feedback mechanism that relies on terminal and console output only allows for limited interactions with users. Interface control elements such as buttons are hardly available in such an environment.

However, we still aimed to provide a consistent look and feel and layout concept for security warnings in different scenarios. Figure 1 illustrates all seven sections relevant for our security warning. The upper left corner (1) is used to indicate a dangerous situation. We use section (2) to give a brief description of the security warning's root cause and section (3) to point the developer to the file and line number that triggered the warning. Section (4) is used to communicate

consequences of using the insecure API calls that were responsible for the security warning. We use sections (5) and (6) to provide context sensitive and actionable advice for the developer. Section (5) provides actionable advice to improve code security while section (6) shows information that allows developers to turn off future security warnings. Section (7) provides links to further background information.

Goal 2: *Describe the Risk Comprehensively.* We aim to clearly and comprehensively communicate the underlying risk to the developer. In contrast to TLS warnings [15] or Android permission dialogs [17], our security warning does not have to deal with false positives. Even in cases when developers made insecure choices intentionally, e.g., for backward compatibility requirements of legacy systems, a security warning is still a true positive.

We rely on sections (1), (2) and (4) to communicate the respective risk to the developer. Section (1) uses a red flashing text icon "/!\", indicating a warning sign. Additionally, we integrated "WARNING" text in capital letters and red color in section (1). Section (2) uses red colored text explaining the root causes of the warning, e. g.: "You are using the weak encryption algorithm RC4 (aka ARC4 or ARCFOUR)". Additional details to communicate the existing risk and its potential consequences are provided in section (4). In case of an RC4 warning, e.g., "The use of ARC4 puts the processed data's confidentiality at risk and may lead to data disclosure."

Goal 3: *Present Relevant Contextual Information.* We aim to present relevant contextual information including the specific location in the source code that triggered the security advice. This helps developers to identify the insecure API use that needs to be fixed. In addition to the filename and line number, section (3) includes a snippet of the source code that triggered the warning.

Goal 4: *Offer Meaningful Options.* The most crucial aspect of a security warning is to offer meaningful options to get out of the situation that triggered the warning. In our case we expect the developer to modify code to either fix a security issue or suppress the warning message for future runs (i. e. click through the warning). In section (5) we provide a secure code snippet to turn the insecure code into secure code and offer an insecure option in section (6) which disables this specific security warning in future runs. Additionally, we provide links to more background information such as OWASP or NIST guidelines for secure programming.

Goal 5: *Be Concise and Accurate.* The guideline of Bauer et al. [7] focused on the design of end-user warnings and recommends to refrain from technical jargon. However, since we target software developers, we do not adopt this recommendation. Technical jargon from the software development domain such as specific names, locations and values of source codes are common elements for developers. Thus, we made such terms part of the warning message. Terms, concepts, technologies and standards from the cryptography domain, however, can not be expected to be general knowledge of a developer [1]. Hence, we omitted cryptographic jargon as much as possible.

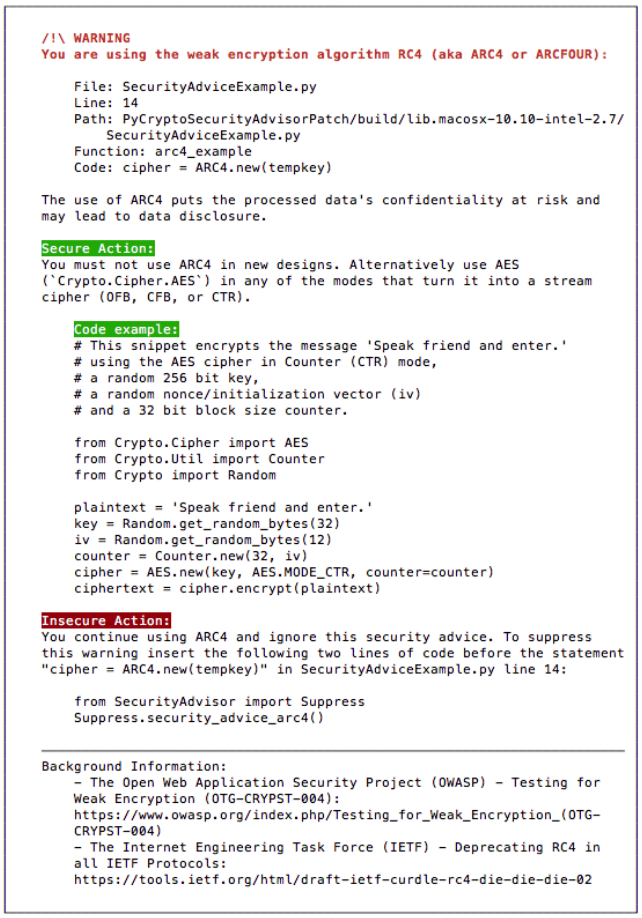


Figure 2: Security advice design of the patched version of PyCrypto triggered by an RC4 usage and displayed in a terminal running python code.

5 Implementation

We implemented the security warning concept from above for a subset of the PyCrypto API for the Python programming language. Figure 2 shows a sample security warning for the insecure RC4 algorithm for symmetric encryption. To assess API call security, we followed the classification provided by Acar et al. [1].

Selecting Python and PyCrypto: We chose to use Python for our experiment because it is very popular, used across many communities and supports many different fields of application. Since Python is easy to read and write and has a large user base [20] we reasoned that recruiting Python developers for our study would be straightforward.

The Python cryptographic PyCrypto [35] API is widely used amongst Python developers. The API provides low level interfaces for cryptographic functionalities, features symmetric as well as asymmetric encryption and supports multiple hashing algorithms as well as some utility features.

PyCrypto comes with primarily auto-generated documentation that includes minimal code examples. The documentation recommends the Advanced Encryption Standard (AES) and provides an example, but also describes the weaker

Data Encryption Standard (DES) as cryptographically secure.[1] The documentation warns against weak exclusive-or (XOR) encryption. However, the documentation does not warn against using the default Electronic Code Book (ECB) mode, or the default empty IV, neither of which is secure. [11, 31]

We chose this API as Acar et al. [1] had identified that developers using this API are likely to produce functionally correct but insecure code. This indicates in general a high potential for improvement. Furthermore, more than 30% of 307 participants in another study [3] preferred PyCrypto over other cryptographic APIs for Python.

5.1 How our Patch works

The PyCrypto patch hooks specific API calls that create instances of weak cryptographic objects such as the call to `Crypto.Cipher.ARC2.new()` which creates a new cipher object that uses the insecure ARC2 [27] algorithm. Whenever an insecure cryptographic object is created, our patch calls an advice method that uses contextual information to show a security warning. To fetch contextual information, the advice method relies on Python's `inspect` module and accesses the cryptographic object's stack frame. The stack frame is used to add information about the responsible file, the line number in that file and the name of the method that triggered a new security warning. Using the respective stack frame information and information about the cryptographic object instantiation that called the security advice method is then used to compile a context specific security warning (cf. Section 4).

5.2 Covered API Calls

For the security warnings, we focused on aspects that we wanted to test in a developer study later on (cf. Section 6). Table 1 gives an overview of both the API calls the security advice does cover and the API calls for which we did not implement security warnings.

In particular, we addressed weak symmetric encryption algorithms (cf. Table 1) and recommended the use of the Advanced Encrytion Standard (AES) as a secure alternative. This is in line with the recommendation of the official developer documentation of PyCrypto. The security warning also recommended an upgrade from the insecure Electronic Code Book (ECB) mode of operation to the secure counter mode (CTR) streaming cipher. In general, we recommended the counter mode (CTR) as a secure mode of operation in all symmetric security warnings. CTR is considered a secure mode of operation and is recommended by the official PyCrypto documentation.

In addition to security warnings for insecure symmetric encryption algorithms, we triggered security warnings for weak hash algorithms (cf. Table 1) and recommended the use of the SHA-512 hash function as a secure alternative.

In general, all security warnings we provided adhered to the documentation to not confuse participants in case they looked up programming questions.

[1]This might be due to the fact that the library has last been updated on 20 Jun 2014.

5.3 Not Implemented

We did not implement a security warning for every insecure cryptographic choice PyCrypto users can make. While we implemented all features that affected the programming tasks in our developer study (cf. Section 5.2), our patch does not cover the PyCrypto API calls below.[2]

We did not implement security warnings for any of the public key and digital signature schemes provided by PyCrypto (cf. Table 1).

6 Developer Study

We used an online, between-subjects study to compare how effectively developers could write correct, secure code using either PyCrypto as a control, or our patched version of PyCrypto with the security intervention. We recruited developers with demonstrated Python experience (on GitHub) for an online study; we also recruited via mailing lists and developer forums.

Participants were assigned to complete a short set of programming tasks; they were randomly assigned either the PyCrypto control condition, or the PyCrypto patch condition, where we tested our security warning.

Within each condition, task order was randomized. All participants were given a symmetric encryption task and a symmetric key generation and storage task.

After finishing the tasks, participants completed a brief exit survey about the experience. We examined participants' submitted code for functional correctness and security.

Ethics and Pre-testing: Due to the location of our universities, there was no formal IRB process. We did, however, model our study material and procedures after an IRB-approved study and adhered to the strict German data and privacy protection laws.

We conducted expert reviews for the design and implementation of our security advice. Therefore, we asked experienced human computer interaction researchers to walk through the warnings and give us feedback. Additionally, we pre-tested the functionality of our PyCrypto patch extensively with participants we excluded from the study later on.

6.1 Study Design

Our study has two conditions; it is modeled closely after the Acar et al. 2017 study on cryptographic Python APIs, which compared the usability of five cryptographic APIs for Python, namely PyCrypto, cryptography.io, M2Crypto, Keyczar and PyNaCl [1], in a between-subjects study for symmetric and asymmetric encryption via three symmetric or four asymmetric programming tasks: (a) a key generation and storage task, (b) an encryption and decryption task, (c) key derivation (symmetric condition only), (d) certificate validation (asymmetric condition only). They find that usability varies wildly across libraries and tasks, with poor usability contributing to insecure code.

In our study, we compare the PyCrypto library to our patched version of PyCrypto. The PyCrypto condition

[2]However, extending the patch to cover a more comprehensive list of features is possible.

Triggers a Security Warning	Security Advice
Crypto.Cipher	
↳ AES.new(k, AES.MODE_ECB, iv)	→ AES.new(k, AES.MODE_CTR, iv)
↳ CAST.new(k, CAST.MODE_ECB, iv)	→ CAST.new(k, CAST.MODE_CTR, iv)
↳ CAST.new(length(k) < 128 bit, mode, iv)	→ AES.new(k, AES.MODE_CTR, iv)
↳ ARC2.new(k, mode, iv)	
↳ DES.new(k, mode, iv)	
↳ DES3.new(k, mode, iv)	
↳ Blowfish.new(k, mode, iv)	
↳ XOR.new(k, mode, iv)	
↳ ARC4.new(k, mode, iv)	
Crypto.Hash	
↳ MD2.new()	→ SHA512.new()
↳ MD4.new()	
↳ MD5.new()	
↳ RIPEMD.new()	
↳ SHA.new()	

Not Implemented
Crypto.Cipher
↳ PKCS1_OAEP
↳ PKCS1_v1_5
Crypto.Protocol
↳ AllOrNothing.isInt
Crypto.PublicKey
↳ DSA
↳ ElGamal
↳ RSA
Crypto.Signature
↳ PKCS1_PSS
↳ PKCS1_v1_5
Crypto.Util
↳ RFC1751
↳ strxor

Table 1: All implemented/not implemented PyCrypto API calls. Implemented calls trigger security warnings; `k` is the key parameter; `iv` is the initialization vector parameter. Not implemented calls did not affect our study results, as they were not useful for our selection of study tasks.

serves as a control. In this study, we focus on symmetric encryption only, and only assign two tasks in random order: In an online Python coding environment [37], our participants were asked to solve two randomly ordered tasks – a symmetric encryption and key generation and storage task – after which they were asked to complete an exit survey that asked usability questions about the library they used in their condition, familiarity with programming in general and Python in particular, and demographic information. In the PyCrypto patch condition, participants were asked to solve both programming tasks with the pre-installed patched PyCrypto version that showed our security advice when triggered according to Section 5.2. Participants could then take the advice for their final solution; ignoring or bypassing the security advice was also possible. In the PyCrypto condition, participants were asked to solve the same two tasks without the support of security advice. Our control condition replicates the 2017 PyCrypto condition for a subset of two out of three of the original tasks [1]. We can therefore not only compare our results across our conditions, but also to the 2017 study.

6.2 Recruitment and Framing

Our study reuses most of the infrastructure of the publicly available Developer Observatory [1,37], and our recruitment strategy closely resembles that described in past studies with the same framework. To gain meaningful, ecologically valid results, we aimed to recruit developers familiar with Python.

We sent a total of 38,533 email invites to randomly sampled contributors from 100,000 publicly available Python repositories. We additionally posted invitations in Python forums and sent emails to our personal network.

In our invitation, we asked Python developers to participate in a Python study via an online code editor. Our invitations did not mention a security or cryptography context to avoid biasing potential participants. The invitation email included links to learn more about the study and to blacklist the recipient email from any further communication related to our research, a request which we honored. The participation link contained a unique pseudonymous identifier (ID), which allowed us to assign study results and GitHub statistics to the invited email addresses.

Recipients who clicked the link to participate in the study were sent to a landing page containing a consent form. Once they confirmed their legal age, consented to the study and were comfortable with participating in the study in English, they were introduced to the study framing previously used by Acar et al. [1]. We asked participants to imagine they were developing code for an app called CitizenMeasure, "a new global monitoring system that will allow citizen-scientists to travel to remote locations and make measurements about such issues as water pollution, deforestation, child labor, and human trafficking. Please keep in mind that our citizen-scientists may be operating in locations that are potentially dangerous, collecting information that powerful interests want kept secret. Our citizen scientists may have their devices confiscated and hacked." We hoped that this framing would both engage participants' interest and nudge them to attempt to write secure code. We also gave instructions for the study infrastructure, which we describe next.

6.3 Experiment Infrastructure

Our online developer study uses our publicly available framework (cf. [37]). The framework allows participants to write and test cryptographic code in their browser, is based on a Jupyter Notebook environment [26] and was hosted on our server. This allowed us to control the development environment including available libraries (PyCrypto in this study) and to retrieve written code and corresponding metadata (e.g., copy&paste events).

As our security advice implementation (cf. Section 5) uses ANSI ESCAPE sequences [24] to colorize text in diverse terminals on various platforms, we had to update Jupyter Notebook to the latest version 4.4.0 in order to be able to display our warning appropriately (ANSI colors were not processed correctly by Jupyter until version 4.1.0). Due to API changes we had to adjust some parts of the Developer Observatories implementation. Depending on conditions, the original version of PyCrypto or the patched version of PyCrypto were used by a participant. Because both share

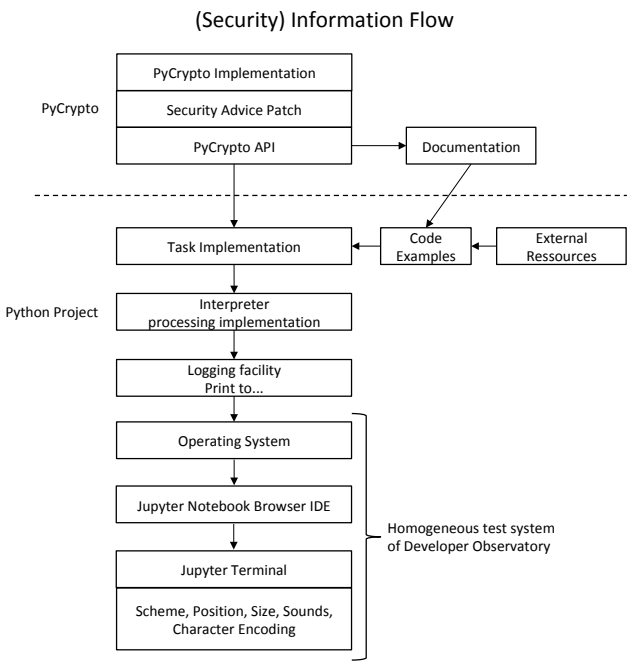


Figure 3: Security Information Flows in development environments through the example of the cryptographic Python API PyCrypto and Developer Observatory [37]

exactly the same name space and the identical API, we installed each library in a virtual Python 2.7.12 environment of which only one was used as kernel in Jupyter.

Figure 3 illustrates how the information of our security advice is technically transferred from the patched PyCrypto API via the Python intepreter and Python logging facility to the participants homogeneous test environment of Developer Observatory.

To prevent interference between participants, each participant was assigned to a Notebook running on a separate Amazon Web Service (AWS) instance. We maintained a pool of prepared instances so that each new participant could begin without waiting for an instance to boot. Instances were shut down when each participant finished, to avoid between-subjects contamination.

Tasks were shown one at a time, with a progress indicator showing that the participant had completed, e. g., 1 of 2 tasks. For each task, participants were given buttons to "Run and test" their code, and to move on using "Solved, next task" or "Not solved, but next task." After each button press, we stored the participant's current code, along with metadata like timing, in a remote database.

Allowing participants to write and execute Python code presents serious security concerns. To mitigate this, we removed all unnecessary software packages from the AWS image. We used the AWS firewall to restrict incoming traffic to port 80 and prevent outgoing traffic other than to our study database, which was password protected and restricted to sanitized insert commands. All instances were shut down within 4 hours of the last observed participant activity.

6.4 Task Design

To be able to compare our results not only to our own control, but also to past results both in functionality outcome, security outcome and usability, we re-used a subset of tasks from the Acar et al. study on the usability of cryptographic APIs [1]. These tasks had previously been chosen to be "short enough so that the uncompensated participants would be likely to complete them before losing interest, but still complex enough to be interesting and allow for some mistakes" and designed to "model real world problems that Python developers could reasonably be expected to encounter in their professional career." We chose two symmetric encryption tasks: generating an encryption key and storing it securely in a password-protected file, and using the key to encrypt some plain text.

For both tasks, participants were provided with stub code and some commented instructions. These stubs were designed to make the task clear and ensure the results could be easily evaluated. We also provided a main method pre-filled with code to test the provided stubs. This helped orient participants and saved time, but it did prevent us from learning how participants might have designed their own tests.

We also asked participants to please use only the PyCrypto documentation, if at all possible, and to report (in comments) any additional documentation resources they consulted. Task order was randomized between participants.

Replication: In the control group, participants were asked to solve the tasks using PyCrypto as-is. Except for a change in task design (i.e., removing the decryption task), this condition is identical with the Acar et al. study PyCrypto condition for their set of symmetric tasks.

Security Advice Condition: Participants in the PyCrypto patch condition were asked to solve the same set of tasks using PyCrypto; they were not alerted that they were using a patched version of PyCrypto. If they successfully executed functional code that was insecure according to the classification in Table 1, and the insecure programming choice was covered by the patched version of PyCrypto, the respective warning message was shown.

6.5 Exit Survey

Once both tasks had been completed or abandoned, the participants were directed to a short exit survey. We asked for their opinions about the tasks they had completed and the PyCrypto API, including the Acar et al. usability questionnaire for security APIs [1]. We also collected their demographics and programming experience. The participant's code for each task was displayed (imported from our database) for their reference with each question about that task. We were also interested in whether participants perceived the security warning at all, if it was helpful and if participants could recall the security warning's content. The Exit survey can be found in the Appendix D.

6.6 Evaluating Solutions

We based our analysis on the code submitted for each task by our participants. Submitted solutions were evaluated both for functional correctness and security. We evaluated each

task independently with two coders based on a subset of the codebook provided by [1]. Disagreements between the two coders were adjudicated by a third coder allowing us to solve all conflicts.

Functionality: For each programming task, we assigned a participant a functionality score of 1 if the code ran without errors, passed the tests and completed the assigned task, or 0 if not.

Security: We assigned security scores only to those solutions which were graded as functional. To determine a security score, we considered several different security parameters. Our scoring followed the relevant parts of the security scoring in [1]. Still we give a brief summary of the security scoring we applied.

For key generation, we checked key size and randomness. For key storage we checked if encryption keys were actually encrypted and if a proper encryption key was derived from the password we provided. For key derivation, we scored use of a static or empty salt, HMAC-SHA1 or below as the pseudorandom function, and less than 10,000 iterations as insecure. For the symmetric encryption task, participants had to select encryption parameters. Therefore, we scored the security of the chosen encryption algorithm, mode of operation, and initialization vector. We scored ARC2, ARC4, Blowfish, (3)DES, and XOR as insecure, and AES as secure. We scored the ECB as an insecure mode of operation and scored CBC, CTR and CFB as secure. Static, zero or empty initialization vectors were scored insecure.

We calculated Krippendorff' alpha [28] for the initial coding by two coders across all security codes; $\alpha = 0.764$, which is within reasonable bounds for agreement [14]. Conflicts were resolved afterwards.

Participant Stories: In addition to our assessment of code functionality and security, we analyzed participants' code in detail, qualitatively, based on the recorded code and console output that we automatically stored for each test run of code. We recreated the sequence of task solutions that each participant executed, the *participant story*, where we could see whether they were shown our security advice and which version was shown, whether or not they subsequently adapted their code to incorporate our suggestions, and whether or not this reaction lead to a secure version of their solution. We additionally see whether they reported having seen a warning in the exit survey, and whether or not they perceived it as useful. We use these participant stories to give insight into four questions: (1) did the developers see the warning?, (2) did they react by modifying their code? (3) did they use our examples in their code? and (4) did this consideration lead to improved code security?

7 Data Analysis

In our data analysis, we use the non-parametric Mann-Whitney-U test (MWU) to compare two groups with continuous responses, compare categorical responses with Person's chi-squared test (χ^2) or instead with Fisher's exact test where applicable, and fit regression models to our results.

For each regression analysis, we consider a set of candidate models and select the model with the lowest Akaike Information Criterion (AIC) [9]. In cases when we consider results on a per-task rather than a per-participant basis, we use a mixed model that adds a random intercept to account for multiple tasks from the same participant. We consider candidate models consisting of the required factors "Task" and "Warning displayed", as well as (where applicable) the participant random intercept, plus every possible combination of the optional variables. Required factors, optional factors, and corresponding baseline values are described in Table 2.

We present the outcome 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. For logistic regressions, the odds ratio (O.R.) measures change in likelihood of the targeted outcome in relation to the baseline factor O.R. of one. 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.

8 Results

We present the results for our study based on 53 valid participants. Participants were generally successful in functionally solving the tasks, while security results varied across conditions, the patched condition being an improvement over PyCrypto where applicable. This improvement was pronounced: participants who wrote code that triggered a warning message were 15× as likely to convert it to a secure condition as opposed to participants who wrote similar insecure code in the PyCrypto condition. However, the effectiveness of our PyCrypto patch was negatively impacted by the limited applicability of the warnings.

8.1 Participants

We recruited participants for our study by sending email invitations to GitHub developers (cf. Figure 4) and by advertising the study in developer forums. Of 38,533 sent invitation emails, 3,422 (8.9%) bounced and 65 (0.2%) recipients requested to be removed from our mailing list.

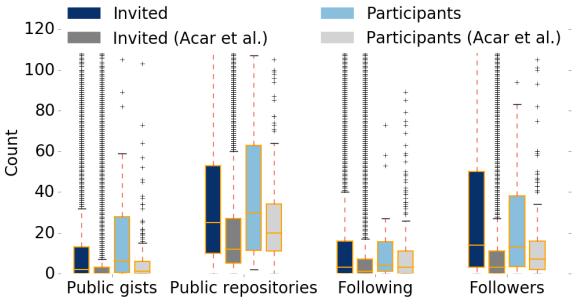


Figure 4: Boxplots comparing invited participants with valid participants and participants from Acar et al. [1]. The center line indicates the median; the boxes indicate the first and third quartiles. The whiskers extend to ±1.5 times the interquartile range. Outliers greater than 150 were truncated for space.

Factor	Description	Baseline
Required		
Task	Performed task (Storage or Encryption).	Encryption
Warning	True or False, whether a warning was displayed.	False
Participant	Random effect accounting for repeated measures.	n/a
Optional		
Used documentation	True or False, used API documentation, self-reported.	False
Development experience	Development experience in years, self-reported.	n/a
Python experience	Python programming experience in years, self-reported.	n/a
PyCrypto experience	Previous experience with the PyCrypto library (used, seen, none), self-reported.	None
Security task experience	Previous experience in solving security tasks (written, seen, none), self-reported.	None

Table 2: Factors used in regression models. Model candidates were defined using all possible combinations of optional factors, with the required factors included in every candidate. Final models were selected by minimum AIC. Categorical factors are individually compared to the baseline.

We received no reports of technical errors with the survey infrastructure; one participant refused to participate in the study, because our Amazon AWS instances were not accessible via HTTPS. One participant refused to participate because he perceived our invitation email to be dubious.

272 people agreed to the consent form and 177 started working on the tasks. Of those, 70 finished the tasks and 68 completed the exit survey. We excluded 15 participants since results indicated a lack of serious answers (4) or were the result of curious clicking-through (3). Unless stated otherwise, we report results for the remaining 53 valid participants which finished the tasks and completed our exit survey.

The majority of our 53 valid participants reported being male (49, 92.5%) while the remaining participants reported being female (1), other (1), or preferred to not answer (2). The reported age was between 20 and 60 (Mean 34.9, SD 8.1). 44 of our participants received invitation emails as GitHub developers, while the remaining 9 were recruited on developer forums. Our participants reported a mean developer experience of 15.8 years (SD 8.2, prefer not to answer: 3) and a mean Python experience of 8.44 years (SD 4.7, prefer not to answer: 3). 48 reported their occupation as being professionals and 3 reported being students (Both: 1, prefer not to answer: 1).

8.2 Dropouts

95 did not continue to the study while 177 started the first programming tasks by clicking the begin button. 57 participants stopped in the key storage task, additional 44 in the content encryption task and 4 in the final test routine before finishing the online programming part of the study. 29 had written code to solve a task in contrast to 76 who did not modify any text in the Jupyter notebook. 5 dropped out of our PyCrypto patch condition after having triggered security advice. 70 proceeded to the exit survey. 68 participants finished the exit survey of which we had to exclude 15 persons due to non serious participation and technical issues in our infrastructure.

We saw that out of 115 participants in the PyCrypto patch condition, 90 participants dropped out of whom 5 were shown a warning. However, the 26 who finished the study were shown 11 warnings, so we assume that seeing a warning was not a strong reason to drop out of the study. We compare this to 34 dropouts out of 62 who started the PyCrypto condition. The increased count of starting participants was due to an effort to counterbalance for the limited applicability of the warnings.

8.3 Results for Functionality

Generally, participants were well able to solve tasks: 87.8% of attempted tasks were functional (89.7% functional in the PyCrypto condition, 85.9% in the PyCrypto patch condition).

We were unable to observe a significant impact, positive or negative, of our warning messages on results, as shown in Table 3. Since the warning message was only presented after functionally correct code was executed, this is to be expected. However, the interruption caused by the warning message did not cause developers to break their code.

Factor	O.R.	C.I.	p-value
Storage Task	0.00	[0, ∞]	0.972
Warning displayed	0.22	[0.03, 1.9]	0.169
Development experience	0.95	[0.84, 1.07]	0.369
Python experience	1.19	[0.93, 1.52]	0.169

Table 3: Results of the final logistic regression model examining whether displayed warnings affect task functionality. Odds ratio (O.R.) indicates relative likelihood of a task being functional. Some trends are observable but no results are statistically significant. See Table 2 for further details.

8.4 Results for Security

For security, we observed 26.9% secure solutions in the PyCrypto condition; compared with 50.7% in the PyCrypto patch condition. We were not able to obtain a meaningful regression model (cf. Appendix B), caused by the small number of tasks that triggered and ended up with insecure code in the PyCrypto patch condition (11), as well as the small number of tasks that would have triggered a warning but were not modified to be secure in the PyCrypto condition (22). We followed this inconclusive model up with Fisher's exact test (cf. Table 4, which was significant ($p<0.01$), with an odds ratio of 56. The warning messages were noticed by participants who saw them, which was clear both from self-reported memory of them as well as changes in their code:

		Secure	
		F	T
Warning	F	21	1
	T	3	8

Table 4: Contingency table for secure task solutions and triggered warnings used in our Fisher's exact test.

Factor	Coef.	C.I.	*p*-value
Warning displayed	0.00	[0, 112.51]	0.271
Development experience	0.67	[0.35, 1.27]	0.229
Python experience	0.73	[0.23, 2.3]	0.595

Table 5: Linear regression model examining usability perceived by participants. See Table 2 for further details.

the warning message lead to a change from initial insecure code to a secure solution in most cases (8 out of 11). Generally, the applicability of the warning message was limited; it applied to 24 of 44 insecure solutions across conditions, and was shown in 11 of 22 insecure cases in the PyCrypto patch condition.

Impact of Intervention on Perceived Usability: API usability as interpreted by answers to questionnaire by Acar et al. [1] based on the Cognitive Dimensions framework [10] did not change for better or worse with the warning (cf. Table 5). This is to be expected, as only one of our 10 questions that are calculated into the usability score focus on meaningful warning/error messages. We investigate in detail the answers to the following questions:

W1 The security warnings displayed in the console helped to solve this task.

W2 When I made a mistake, I got a meaningful error message/exception.

W3 Using the information from the error message/exception, it was easy to fix my mistake.

We transform agreement on a 5-point likert-scale as follows: neutral is represented by 0, while strong disagreement is represented by -2 and strong agreement is represented by $+2$. The mean agreement to W1 was 1 (median = 1) in the PyCrypto condition compared with 0.76 (median = 1) in the PyCrypto patch condition (MWU-test; U=32.5; p=0.4205). Participants gave a mean agreement of 0.593 (median = 1) to W2 in the PyCrypto condition compared with 0.833 (median = 1) in the PyCrypto patch condition (MWU-test; U=384; p=0.2167), and a mean agreement of of 0.846 (median = 1) to W3 in the PyCrypto condition compared with 0.917 (median = 1) in the PyCrypto patch condition (MWU-test; U=296; p=0.7484). We interpret this as a generally positive impression of our warning, despite our preliminary fear of annoying or overwhelming developers. However, even in these specific cases, perceptions were not significantly more positive or negative than in the control condition.

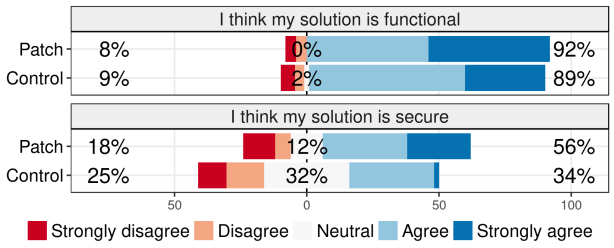


Figure 5: Likert-plot showing our participants' perceptions regarding functionality and security of their solutions. "I don't know" answers were omitted.

8.5 Detailed Task Analysis

Participants were asked to rate their functional and security success after completing the tasks (cf. Figure 5). Interestingly, we found that for the encryption tasks, all participants who saw the warning message were correct in assessing their solution's security. We compare this to the control condition, where, for the encryption task, only 66% task security ratings were correct in cases where the warning would have applied. In the key storage task in the patched condition, 73% of assessments were correct, while all assessments were correct in the control group.

Participant Stories: From the collected participants' stories we derived further qualitative results. When focusing on the content encryption task, 7 participants were shown a security warning. All of them saw and remembered it, as they reported in the exit survey. 2 of the 7 participants did not choose to use our guidance to improve their code. One tried to suppress the advice, another one ignored it. The remaining 5 participants accepted the advice and modified their code: 2 of them adopted the example code provided by the advice; they later stated their satisfaction: "*The warning helpfully directed me towards an improved solution, and provided example code*" and "*The warning explained clearly that DES was considered as insecure, and provided an example to use AES instead. This helped me solving this task in a more secure manner*". The remaining 3 participants partially followed the advice: they did adapt their code in response to the warning, but chose a different mode of operation than was suggested in the warning. The proposed solution recommended the use of standard encryption algorithm AES in counter (CTR) mode. The 3 participant instantiated AES in cipher block chaining (CBC) mode instead. A closer look at their code revealed that 2 of them appeared to have problems in transferring the suggested code snippet into running code. While this points to a usability problem with the warning/advice, we were able to observe that 4 out of 5 participants who modified their code in reaction to our warning at least attempted to adhere to our suggestion. Altogether, 5 out of 7 participants who saw the warning for the encryption task modified their code into a secure solution.

We could observe similar behavior for the key generation and storage task. Here, 4 participants were shown security advice; all of them noticed the warning. One ignored the warning; the remaining 3 modified their code. One adopted the suggested code snippet as-is; the other 2 chose CBC mode instead of CTR mode.

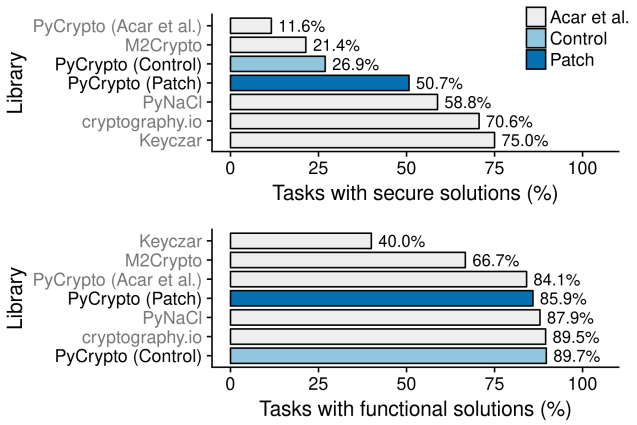


Figure 6: Comparison between secure solution percentages of libraries from Acar et al. [1] and our PyCrypto Control/Patch data. To match our tasks, only symmetric encryption tasks are considered for libraries from Acar et al.

Limited applicability of warning message: The programming tasks in our study were designed in a way participants had to only use symmetric cryptography. Thus we did not cover insecure asymmetric API features (cf. Table 1). However, 4 participants in the patch condition solved tasks by using asymmetric methods. They implemented key derivation for asymmetric RSA keys or applied RSA to encrypt keys and messages. Our security advice implementation were not able to help these participants using symmetric cryptography since it is not possible to give task sensitive advice at this position. For this reason we had to exclude these tasks from the detailed analysis.

8.6 Replication Results

Our study is based on the study that compared the usability of different cryptographic APIs conducted by Acar et al. [1]. This section discusses the aspects we replicated and the replication results.

Our participants created a similar level of functional tasks as compared to the 2017 study (cf. Figure 6). However, our control group achieved better security results than the original study.

Participants in the PyCrypto patch condition of our study achieved a higher level of security than our own control group, which places the PyCrypto patch condition among the better-performing of libraries. While the effect of more experience applies here, too, it is interesting to see that this result was achieved without changes to the API abstraction level, learnability, documentation, Stack Overflow or the study design. Additional details about common errors of our participants compared to PyCrypto library users from Acar et al. [1] can be found in Appendix C.

9 Limitations

We address multiple limitations below:

Security Advice Design: The design of our security advice is based on heuristics defined by previous research from warning message design for end-users. Additionally, we considered lessons learned from previous work on secure programming studies. After manual pre-testing and expert reviews, we opted for a solution shown in Figure 1. However, there might be more effective designs we did not consider (e. g. following an opinionated design approach might provide better results). Although this is a limitation of our current approach, results for our solution show a significant positive impact on code security. Hence, we leave changes to the design and comparing different versions to future work.

Security Advice Implementation: The implementation of our security advice does not cover all possible insecure choices PyCrypto users can make, e.g. we did not implement security warnings for PyCrypto's asymmetric API (cf. Section 5), however, these were not included in our study design. We address participants using APIs not covered by our security advice, as well as cases where we failed to show security advice (e.g., non-random IVs for symmetric tasks) in our data analysis (cf. Section 7).

User Study: We decided to conduct an online study over a laboratory study because it is difficult to recruit software developers (rather than students) at a reasonable cost. This design decision allowed us less control over the study environment. On the other side, we were able to recruit a geographically diverse set of participants. Sadly, we could not simply recruit participants from an online service such as Amazon Mechanical Turk for end-user focused studies. Since it is difficult to manage participants compensations outside such infrastructures, we did not offer our participants compensation. Due to the combination of unsolicited email invites and no compensation we expected a strong self-selection bias and are aware of the fact that our results might not necessarily be representative for all developers but in particular for those who are interested and motivated enough to participate. Our participants seem to be more active than average GitHub users (cf. Appendix A). However, these limitations apply across both conditions. In any online study, some participants may not provide full effort, or may answer haphazardly. We attempted to remove any obviously low-quality data before analysis, but cannot discriminate perfectly. Additionally, we tested a simple and limited scenario, which may have limited applicability to complex real world code.

Real-world applicability: Critically, a real-world roll-out of our advice is contingent on buy-in from library developers. While this requirement severely limits employment across all libraries, several cryptographic library developers have reached out after the 2017 study and showed commitment to improve their libraries' usability. We therefore hope that our study is not only of academic relevance, but can and will be applied to libraries with a large userbase.

10 Discussion

Overall, we found that our API-integrated security advice had a significantly positive effect on code security. However, we only tested a first implementation of our approach. Changing parameters such as text or advice design might result in even more secure code. We leave this to future

	Functionality	Security	Usability
Information Source [2]	✓	✓	—
Cryptographic Library [1]	✓	✓	✓
FixDroid [34]	✗	✓	—
Security Advice	✗	✓	✗

Table 6: Comparison of the impact of our security warning compared to previous investigations of the impact of other factors on code security.

work. The majority of the participants who were shown a security warning, fixed their code. Interestingly, showing participants a security warning had no effect on the functionality of participant solutions. Also, the perceived usability of PyCrypto as a cryptographic API was not affected by the security warning. Only one participant who received the advice, suppressed security warnings for future runs and two participants copied secure code snippets from a warning into their code.

Other Approaches: Comparing our security advice approach to previous work yields interesting results. Similar to high quality developer documentation, simple programming interfaces or IDE plugins our approach has a positive impact on code security. However, in contrast we could not find a positive effect on functionality (cf. Table 6). Also, in contrast to API design, our warning did not have a positive impact on perceived API usability.

However, our approach has multiple advantages in terms of deployability (cf. Section 4) and allows API providers to improve code security for existing cryptographic APIs in a bottom-up approach without having to change API design or relying on third party tools.

Lessons Learned: Most importantly we learned that API-integrated security advice can have a significant impact on code security. The majority of the participants who received security advice turned insecure code into secure code. Also, the adherence rate to our security advice (73%) was similar to adherence rates for browser warnings reported in previous work [15]. However, additionally we learned that designing and implementing effective security advice is challenging and has its limitations. Providing context sensitive information and secure and ready-to-use code snippets is complex and requires future work.

Future Work: Our work leaves room for future work in multiple directions.

While we evaluated API-integrated security advice for Python's PyCrypto API and reported a significantly positive effect on code security, it is unclear to which extent our concept can be applied to other security APIs. Hence, we aim to implement and test similar security warning concepts for a number of other security APIs such as for secure networking (e. g. TLS and HTTPS) or authentication (e. g. OAuth) as suggested by [30].

We followed security warning design guidelines by Bauer et al. [7] and considered lessons learned from related work on developer usable security research (cf. [1, 2, 34]) to design a first attempt at security advice. However, we only chose one specific design to test, and did not conduct any testing against other designs. Likely, the concrete design, content and presentation of the security advice can be improved. Future work could investigate the effect of an opinionated design approach or other security indicators. Warning message research for end-users showed significant impact of such factors on security (cf. [15]). Also, the integration of our approach in an integrated development environment (IDE) needs to be considered.

We conducted a between-subjects first contact study. In future work we plan to conduct a large scale in-situ field experiment to investigate the impact of habituation and fatigue on our approach.

11 Conclusion

In this paper, we evaluate the first API-integrated security advice for cryptographic APIs. We follow design guidelines by Bauer et al. [7] and consider lessons learned from previous work on human factors research for software developers. We implement a first design approach for Python's PyCrypto API and use the Developer Observatory framework [37] to conduct a between-subjects online controlled experiment. We evaluate the impact of our security advice on code security and perceived API usability and put our results in perspective of other approaches that try to support developers to write more secure cryptographic code.

Overall, we find that our security advice had a significantly positive impact on code security (**RQ1**) and did not affect the perceived API usability of our participants (**RQ2**). Similar to other approaches in previous work, the presented security advice helps to improve code security. Differently from other work, our approach allows API providers themselves to fix security and usability shortcomings of their interfaces without having to change programming interfaces or relying on resources outside their sphere of influence, such as third party information resources, IDE plugins or static code analysis tools (**RQ3**).

12 Acknowledgments

The authors would like to thank Joe Calandrino and the anonymous reviewers for providing feedback; and all participants of this study for their voluntary participation. This work has been partially funded by the German Federal Ministry of Education and Research within the funding program "Forschung an Fachhochschulen" (contract no. 13FH016IX6).

13 References

[1] Y. Acar, M. Backes, S. Fahl, S. Garfinkel, D. Kim, M. L. Mazurek, and C. Stransky. Comparing the usability of cryptographic APIs. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 154–171, 2017.

[2] Y. Acar, M. Backes, S. Fahl, D. Kim, M. L. Mazurek, and C. 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, May 2016.

[3] Y. Acar, C. Stransky, D. Wermke, M. L. Mazurek, and S. 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.

[4] D. Akhawe and A. P. Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *Proceedings of the 22Nd USENIX Conference on Security*, SEC'13, pages 257–272, Berkeley, CA, USA, 2013. USENIX Association.

[5] H. Almuhimedi, A. P. Felt, R. W. Reeder, and S. Consolvo. Your reputation precedes you: History, reputation, and the chrome malware warning. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 113–128, Menlo Park, CA, 2014. USENIX Association.

[6] T. Barik, J. Smith, K. Lubick, E. Holmes, J. Feng, E. Murphy-Hill, and C. Parnin. Do developers read compiler error messages? In *39th IEEE/ACM International Conference on Software Engineering (ICSE)*, pages 575–585. IEEE, 2017.

[7] L. Bauer, C. Bravo-Lillo, L. Cranor, and E. Fragkaki. Warning design guidelines. Technical Report CMU-CyLab-13-002, CyLab, Carnegie Mellon University, 2013.

[8] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs, and S. Schechter. Your attention please: Designing security-decision uis to make genuine risks harder to ignore. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, pages 6:1–6:12, New York, NY, USA, 2013. ACM.

[9] K. P. Burnham. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304, 2004.

[10] S. Clarke. How usable are your APIs? In A. Oram and G. Wilson, editors, *Making software: what really works, and why we believe it*, Theory in practice, pages 545 – 565. O'Reilly, Beijing, 1 edition, 2010.

[11] Cwe-329: Not using a random iv with cbc mode. [Online]. Available: http://cwe.mitre.org/data/definitions/329.html, 2018.

[12] M. Egele, D. Brumley, Y. Fratantonio, and C. Kruegel. An empirical study of cryptographic misuse in android applications. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS '13, pages 73–84, New York, NY, USA, 2013. ACM.

[13] S. Egelman, L. F. Cranor, and J. Hong. You've been warned: An empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1065–1074, New York, NY, USA, 2008. ACM.

[14] P. J. Fahy. Addressing some common problems in transcript analysis. *The International Review of Research in Open and Distributed Learning*, 1(2), 2001.

[15] A. P. Felt, A. Ainslie, R. W. Reeder, S. Consolvo, S. Thyagaraja, A. Bettes, H. Harris, and J. Grimes. Improving ssl warnings: Comprehension and adherence. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2893–2902, New York, NY, USA, 2015. ACM.

[16] A. P. Felt, A. Ainslie, R. W. Reeder, S. Consolvo, S. Thyagaraja, A. Bettes, H. Harris, and J. Grimes. Improving SSL warnings: Comprehension and adherence. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2893–2902. ACM, 2015.

[17] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, pages 3:1–3:14. ACM, 2012.

[18] A. P. Felt, R. W. Reeder, H. Almuhimedi, and S. Consolvo. Experimenting at scale with google chrome's ssl warning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2667–2670, New York, NY, USA, 2014. ACM.

[19] M. Georgiev, S. Iyengar, S. Jana, R. Anubhai, D. Boneh, and V. Shmatikov. The most dangerous code in the world: Validating ssl certificates in non-browser software. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, pages 38–49, New York, NY, USA, 2012. ACM.

[20] Githut - programming languages and github. [Online]. Available: http://githut.info/, 2018.

[21] P. L. Gorski and L. Lo Iacono. Towards the Usability Evaluation of Security APIs. In *10th International Symposium on Human Aspects of Information Security and Assurance (HAISA)*, 2016.

[22] M. Green and M. Smith. Developers are not the enemy!: The need for usable security apis. *IEEE Security & Privacy*, 14(5):40–46, 2016.

[23] S. Indela, M. Kulkarni, K. Nayak, and T. Dumitraş. Helping johnny encrypt: Toward semantic interfaces for cryptographic frameworks. In *Proceedings of the 2016 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, Onward! 2016, pages 180–196, New York, NY, USA, 2016. ACM.

[24] ISO/IEC 6429. Information technology – control functions for coded character sets. [Online]. Available: https://www.iso.org/standard/12782.html, 1992. Third edition.

[25] B. Johnson, Y. Song, E. Murphy-Hill, and R. 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.

[26] Jupyter notebook. [Online]. Available: http://jupyter.org/, 2018.

[27] J. Kelsey, B. Schneier, and D. Wagner. Related-key cryptanalysis of 3-way, biham-des,cast, des-x, newdes, rc2, and tea. In Y. Han, T. Okamoto, and S. Qing, editors, *Information and Communications Security*, pages 233–246, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg.

[28] K. Krippendorff. *Content Analysis: An Introduction to*

Its Methodology (2nd ed.). SAGE Publications, 2004.

[29] S. Krüger, S. Nadi, M. Reif, K. Ali, M. Mezini, E. Bodden, F. Göpfert, F. Günther, C. Weinert, D. Demmler, and R. Kamath. Cognicrypt: Supporting developers in using cryptography. In *Proceedings of the 32Nd IEEE/ACM International Conference on Automated Software Engineering*, ASE 2017. IEEE Press, 2017.

[30] L. Lo Iacono and P. L. Gorski. I Do and I Understand. Not Yet True for Security APIs. So Sad. In *The 2nd European Workshop on Usable Security*, EuroUSEC '17, 2017. doi: 10.14722/eurousec.2017.23015.

[31] A. J. Menezes, P. C. Van Oorschot, and S. A. Vanstone. *Handbook of applied cryptography*. CRC press, 1996.

[32] S. Nadi, S. Krüger, M. Mezini, and E. Bodden. "Jumping Through Hoops": Why do Java Developers Struggle With Cryptography APIs? In *Proceedings of the 37th International Conference on Software Engineering (ICSE 2016)*, 2016.

[33] A. Naiakshina, A. Danilova, C. Tiefenau, M. Herzog, S. Dechand, and M. 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. ACM, 2017.

[34] D. C. Nguyen, D. Wermke, Y. Acar, M. Backes, C. Weir, and S. 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*, CCS '17. ACM, 2017.

[35] Pycrypto - the python cryptography toolkit. [Online]. Available: https://www.dlitz.net/software/pycrypto/, 2018.

[36] B. Reaves, N. Scaife, A. Bates, P. Traynor, and K. R. Butler. Mo(bile) money, mo(bile) problems: Analysis of branchless banking applications in the developing world. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 17–32, Washington, D.C., 2015. USENIX Association.

[37] C. Stransky, Y. Acar, D. C. Nguyen, D. Wermke, D. Kim, E. M. Redmiles, M. Backes, S. Garfinkel, M. L. Mazurek, and S. 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)*. USENIX Association, 2017.

[38] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor. Crying wolf: An empirical study of ssl warning effectiveness. In *Proceedings of the 18th Conference on USENIX Security Symposium*, SSYM'09. USENIX Association, 2009.

[39] A. Vance, B. Kirwan, D. Bjornn, J. Jenkins, and B. B. 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 the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2215–2227, New York, NY, USA, 2017. ACM.

[40] J. Weinberger and A. P. Felt. A week to remember: The impact of browser warning storage policies. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 15–25, Denver, CO, 2016. USENIX Association.

[41] G. Wurster and P. C. van Oorschot. The developer is the enemy. In *Proceedings of the 2008 New Security Paradigms Workshop*, NSPW '08, pages 89–97, New York, NY, USA, 2008. ACM.

[42] J. Xie, B. Chu, H. R. Lipford, and J. T. Melton. ASIDE: IDE support for web application security. In *Twenty-Seventh Annual Computer Security Applications Conference, ACSAC 2011, Orlando, FL, USA, 5-9 December 2011*, pages 267–276, 2011.

APPENDIX

A Participants

Age	Youngest, Oldest	20, 60
	Prefer not to answer	3
	Mean years (SD)	34.9 (8.1)
Sex	Male	49
	Female	1
	Other	1
	Prefer not to answer	2
Recruitment	GitHub	44
	Other	9
Experience	Mean development years (SD)	15.8 (8.2)
	Mean Python years (SD)	8.44 (4.7)
	Prefer not to answer	3
Occupation	Pro	48
	Student	3
	Both	1
	Prefer not to answer	1

Demographic	**Invited**	**Valid**
Hireable	20.7%	13.0%
Company listed	41.4%	30.4%
URL to Blog	49.4%	47.8%
Biography added	19.1%	21.7%
Location provided	63.9%	65.2%
Public gists (median)	2.0	6.0
Public repositories (median)	25.0	30.0
Following (users, median)	3.0	4.0
Followers (users, median)	14.0	13.0
GitHub profile creation (days ago, median)	2431.0	2589.0
GitHub profile last update (days ago, median)	30.0	30.0

Table 7: GitHub-related demographics for invited users and valid GitHub participants.

Error	Our Study	Acar et al.
No Encryption	0 (0.00%)	0 (0.00%)
Weak Algorithm	10 (35.71%)	17 (41.46%)
Weak Mode	9 (32.14%)	23 (56.10%)
Static IV	11 (39.29%)	29 (70.73%)
Participants	53 (100%)	41 (100%)

Table 10: Common errors in the encryption task of our participants compared to PyCrypto library users from Acar et al. [1].

B Regression Model

Factor	O.R.	C.I.	p-value
Warning displayed	4.70	[0.04, 492.14]	0.515
Storage Task	0.13	[0.01, 1.25]	0.078
Development experience	1.11	[0.88, 1.39]	0.378

Table 8: Results of the final logistic regression model examining whether displayed warnings improve task security in cases where a warning would have been triggered. Odds ratio (O.R.) indicates relative likelihood of a task being secure. Some trends are observable but not results are statistically significant. See Table 2 for further details.

C Replication

Error	Our Study	Acar et al.
Key In Plain	1 (3.57%)	4 (9.76%)
Weak Cipher	9 (32.14%)	11 (26.83%)
Weak Mode	7 (25.00%)	14 (34.15%)
Static IV	9 (32.14%)	3 (7.31%)
No KDF	16 (57.14%)	15 (36.59%)
Custom KDF	16 (57.15%)	11 (26.83%)
KDF Salt	1 (3.57%)	1 (2.44%)
KDF Algorithm	3 (10.71%)	1 (2.44%)
KDF Iterations	1 (3.57%)	2 (4.88%)
Participants	53 (100%)	41 (100%)

Table 9: Common errors in the key file task of our participants compared to PyCrypto library users from Acar et al. [1].

D Exit Survey Questions

D.1 Task-specific questions: Asked about each task

Please rate your agreement to the following statements:

I think I solved this task correctly.

- strongly agree
- agree
- neutral
- disagree
- strongly disagree
- I don't know

I think I solved this task securely.

- strongly agree
- agree
- neutral
- disagree
- strongly disagree
- I don't know

Did you use the PyCrypto API documentation to solve this task?

- Yes
- No

If Yes: Please rate your agreement to the following statements:

The documentation was helpful in solving this task.

- strongly agree
- agree
- neutral
- disagree
- strongly disagree
- I don't know

Which parts of the documentation did you use?

Did you see any security warnings while working on this task?

- Yes
- No

If Yes: Please rate your agreement to the following statements:

The security warnings displayed in the console helped to solve this task.

- strongly agree
- agree
- neutral
- disagree
- strongly disagree
- I don't know

Please explain why the security warnings were helpful or rather unhelpful.

- freetext answer

D.2 General questions about previous experience

Have you used the PyCrypto library before? For example, maybe you worked on a project that used PyCrypto, but someone else wrote that portion of the code.

- I have used PyCrypto before
- I have seen PyCrypto used but have not used it myself
- No, neither
- I don't know

Have you used or seen code for tasks similar to the tasks given in the study before? For example, maybe you worked on a project that included a similar task, but someone else wrote that portion of the code.

- I have written similar code
- I have seen similar code but have not written it myself
- No, neither
- I don't know

D.3 Usability perception

Please rate your agreement to the following questions on a scale from 'strongly agree' to 'strongly disagree.' (strongly agree; agree; neutral; disagree; strongly disagree; does not apply)

- I had to understand how most of the assigned library works in order to complete the tasks.
- It would be easy and require only small changes to change parameters or configuration later without breaking my code.
- After doing these tasks, I think I have a good understanding of the assigned library overall.
- I only had to read a little of the documentation for the assigned library to understand the concepts that I needed for these tasks.
- The names of classes and methods in the assigned library corresponded well to the functions they provided.
- It was straightforward and easy to implement the given tasks using the assigned library.
- When I accessed the assigned library documentation, it was easy to find useful help.
- In the documentation, I found helpful explanations
- In the documentation, I found helpful code examples.
- When I made a mistake, I got a meaningful error message/exception.
- Using the information from the error message/exception, it was easy to fix my mistake.

D.4 Message design assessment

Please rate your agreement to the following statements concerning this console warning:

[Example security advice figure]

How helpful would you rate... (not helpful at all; somewhat unhelpful; neutral; somewhat helpful; very helpful; I don't know)

- ...the risk explanation?
- ...the recommendation for secure action?
- ...the given code example?
- ...the described option for insecure action?
- ...the given background information?
- ...the structure of this security advice?
- ...the amount of information in the message?
- ...the appearance of this kind of messages when using the PyCrypto Library?

What aspects of the warning could be improved, in your opinion?

- free text

D.5 Development Environment

Please tell us some details about your usual Python software development tool chain.

Which console do you use?

- free text

Which text editor do you use?

- free text

What IDE do you use?

- free text

Do you use other tools for software development?

- free text

D.6 Demographic Questions

What type(s) of software do you develop?

- Web Applications
- Mobile Applications
- Desktop Applications
- Embedded Applications
- Enterprise Applications
- Other:

How many years of development experience do you have?

- Number field 0-100
- Prefer not to answer

How many years have you been programming in Python?

- Number field 0-100
- Prefer not to answer

What is your current occupation?

- Freelance developer
- Industrial developer
- Industrial researcher
- Academic researcher
- Graduate student
- Undergraduate student
- Prefer not to answer
- Other:

What is your gender?

- Female
- Male
- Prefer not to answer
- Other:

What country do you live in?

- Please choose... (Dropdown)

How old are you?

- Free text for number of years
- Prefer not to answer

Security in the Software Development Lifecycle

Hala Assal
School of Computer Science
Carleton University
Ottawa, ON Canada
HalaAssal@scs.carleton.ca

Sonia Chiasson
School of Computer Science
Carleton University
Ottawa, ON Canada
Chiasson@scs.carleton.ca

ABSTRACT

We interviewed developers currently employed in industry to explore real-life software security practices during each stage of the development lifecycle. This paper explores steps taken by teams to ensure the security of their applications, how developers' security knowledge influences the process, and how security fits in (and sometimes conflicts with) the development workflow. We found a wide range of approaches to software security, if it was addressed at all. Furthermore, real-life security practices vary considerably from best practices identified in the literature. Best practices often ignore factors affecting teams' operational strategies. *Division of labour* is one example, whereby complying with best practices would require some teams to restructure and re-assign tasks—an effort typically viewed as unreasonable. Other influential factors include company culture, security knowledge, external pressure, and experiencing a security incident.

1. INTRODUCTION

Software security focuses on the resistance of applications to malicious attacks resulting from the exploitation of vulnerabilities. This is different from security functions, which can be expressed as functional requirements, such as authentication [60]. With increasing connectivity and progress towards the Internet of Things (IoT), threats have changed [30]. In addition to vulnerabilities in traditional computing systems (*e.g.*, Heartbleed [21]), vulnerabilities are found in devices and applications that are not necessarily considered security sensitive, such as cars [28], and medical devices [43]. Moreover, the threat is no longer limited to large enterprises; Small and Medium Enterprises (SMEs) are increasingly becoming targets of cyberattacks [50].

With increasing threats, addressing security in the Software Development Lifecycle (SDLC) is critical [25, 54]. Despite initiatives for implementing a *secure* SDLC and available literature proposing tools and methodologies to assist in the process of detecting and eliminating vulnerabilities (*e.g.* [16, 18, 20, 48]), vulnerabilities persist. Developers are often viewed as "the weakest link in the chain" and are blamed for security vulnerabilities [27, 58]. However, simply expecting developers to keep investing more efforts in security is unrealistic and unlikely to be fruitful [14].

Usable security research focusing on developers and the human factors of software security–a new area that has not been sufficiently investigated–has the potential for a widespread positive influence on security [14, 27]. Towards guiding research in this area, Acar *et al.* [14] proposed a research agenda for usable security for developers where they highlight important research questions.

Our work is a step towards addressing one of the prominent research areas outlined by Acar *et al.*'s research agenda [14]. This paper explores steps that teams are taking to ensure the security of their applications, how developers' security knowledge influences the process, and how security fits in (and sometimes conflicts with) the development workflow. We interviewed 13 developers who described their tasks, their priorities, as well as tools they use. During the data analysis we recognized that our participants' practices and attitudes towards security formed two groups, each with trends distinguishable from the other group. On comparing real-life security practices to best practices, we also found significant deviations.

This paper makes the following contributions.

- We present a qualitative study looking at real-life practices employed towards software security.
- We amalgamate software security best practices extracted from the literature into a concise list to assist further research in this area.
- We reflect on how well current security practices follow best practices, identify significant pitfalls, and explore why these occur.
- Finally, we discuss opportunities for future research.

2. RELATED WORK

Green and Smith [27] discussed how research addressing the human factors of software security is generally lacking, and that developers are often viewed as "the weakest link"—mirroring the early attitude towards end-users before usable security research gained prominence. While developers are more technically experienced than typical end-users, they should not be mistaken for security experts [14, 27]. They need support when dealing with security tasks, *e.g.*, through developer-friendly security tools [58] or programming languages that prevent security errors [27]. To this end, Acar *et al.* [14] outlined a research agenda towards understanding developers' attitudes and security knowledge,

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

exploring the usability of available security development tools, and proposing tools and methodologies to support developers in building secure applications. We now discuss relevant research addressing such human aspects of software security.

Generally, studies in this area face challenges in recruiting developers and ensuring ecologically validity. Developers are busy and must often comply with organizational restrictions on what can be shared publicly. To partially address these issues, Stransky *et al.* [51] designed a platform to facilitate distributed online programming studies with developers.

Oliveira *et al.* [22] showed that security vulnerabilities are "blind spots" in developers' decision-making processes; developers mainly focus on functionality and performance. To improve code security, Wurster and van Oorschot [58] recommend taking developers out of the development loop through the use of Application Programming Interfaces (APIs). Towards this goal, Acar *et al.* [12] evaluated five cryptographic APIs and found usability issues that sometimes led to insecure code. However, they found that documentation that provided working examples was significantly better at guiding developers to write secure code. Focusing on software security resources in general, Acar *et al.* [15] found that some available security advice is outdated and most resources lack concrete examples. In addition, they identified some underrepresented topics, including program analysis tools.

Focusing on security analysis, Smith *et al.* [48] showed that tools should better support developers' information needs. On exploring developers' interpretation of Static-code Analysis Tool (SAT) warnings, they found that participants frequently sought additional information about the software ecosystem and resources. To help developers to focus on the overall security of their code, Assal *et al.* [16] proposed a visual analysis environment that supports collaboration while maintaining the codebase hierarchy. This allows developers to build on their existing knowledge of the codebase during code analysis. Perl *et al.* [41] used machine learning techniques to develop a code analysis tool. Their tool has significantly fewer false-positives compared to similar ones. Nguyen *et al.* [40] developed a plugin to help Android application developers adhere to, and learn about, security best practices without distributing their workflow.

Despite their benefits [17], SATs are generally underused [31]. Witschey *et al.* [56] investigated factors influencing the adoption of security tools, such as tool qualities, and developers' personalities and experiences. They found that more experienced developers are more likely to adopt security tools, whereas tool complexity was a deterring factor. Additionally, Xiao *et al.* [59] found that the company culture, the application's domain, and the company's standards and policies were among the main determinants for the developers' adoption of security tools. To encourage developers to use security tools, Wurster and van Oorschot [58] suggest mandating their use and rewarding developers who code securely.

As evidenced, several research gaps remain in addressing the human aspects of software security. Our study takes a holistic perspective to explore real-life security practices, an important step in improving the status-quo.

3. STUDY DESIGN AND METHODOLOGY

We designed a semi-structured interview study and received IRB clearance. The interviews targeted 5 main topics: general development activities, attitude towards security, security knowledge, security processes, and software testing activities (see Appendix A for interview script). To recruit participants, we posted on development forums and relevant social media groups, and announced the study to professional acquaintances. We recruited 13 participants; each received a $20 Amazon gift card for participation. Before the one-on-one interview, participants filled out a demographics questionnaire. Each interview lasted approximately 1 hour, was audio recorded, and later transcribed for analysis. Interviews were conducted in person ($n = 3$) or through VOIP/video-conferencing ($n = 10$). Data collection was done in 3 waves, each followed by preliminary analysis and preliminary conclusions [26]. We followed Glaser and Strauss's [26] recommendation by concluding recruitment on saturation (*i.e.*, when new data collection does not add new themes or insights to the analysis).

Teams and participants. A project team consist of teams of developers, testers, and others involved in the SDLC. Smaller companies may have only one project team, while bigger companies may have different project teams for different projects. We refer to participants with respect to their project teams; team i is referred to as Ti and P-Ti is the participant from this team. We did not have multiple volunteers from the same company. Our data contains information from 15 teams in 15 different companies all based in North America; one participant discussed work in his current (T7) and previous (T8) teams, another discussed his current work in T10 and his previous work in T11. In our dataset, seven project teams build web applications and services, such as e-finance, online productivity, online booking, website content management, and social networking. Eight teams deliver other types of software,*e.g.*, embedded software, kernels, design and engineering software, support utilities, and information management and support systems. This classification is based on participants' self-identified role and products with which they are involved, and using Forward and Lethbridge's [24] software taxonomy. Categorizing the companies to which our teams belong by number of employees [19], 7 teams belong to SMEs (T4, T7, T10–T14) and 8 teams belong to large enterprises (T1–T3, T5, T6, T8, T9, T15). All participants hold university degrees which included courses in software programming, and are currently employed in development with an average of 9.35 years experience ($Md = 8$). We did not recruit for specific software development methodologies. Some participants indicated following a waterfall model or variations of Agile. See Table 3 in Appendix B for participant demographics.

Analysis. Data was analyzed using the Qualitative Content Analysis methodology [9,23]. It can be deductive, inductive, or a combination thereof. For the deductive approach, the researcher uses her knowledge of the subject to build an analysis matrix and codes data using this matrix [9]. The inductive method, used when there is no existing knowledge of the topic, includes open coding, identifying categories, and abstraction [9].

We employed both the deductive and inductive methods of content analysis. The deductive method was used to structure our analysis according to the different development stages. We built an initial analysis matrix of the main SDLC stages [49]. After a preliminary stage of categoriz-

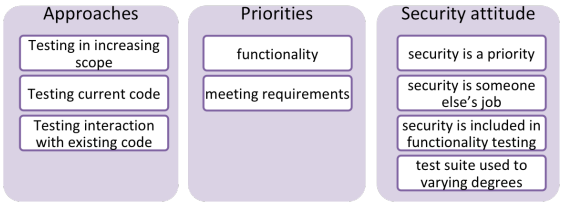


Figure 1: Security adopters: developer testing abstraction

ing interview data and discussions between the researchers, the matrix was refined. The final analysis matrix defines the stages of development as follows. *Design* is the stage where the implementation is conceptualized and design decisions are taken; *Implementation* is where coding takes place; *Developer testing* is where testing is performed by the developer; *Code analysis* is where code is analyzed using automated tools, such as SATs; *Code review* is where code is examined by an entity other than the developer; *Post-development testing* is where testing and analysis processes taking place after the developer has committed their code.

We coded interview data with their corresponding category from the final analysis matrix, resulting in 264 unique excerpts. Participants talked about specific tasks that we could map to the matrix stages, despite the variance in development methodologies. We then followed an inductive analysis method to explore practices and behaviours within each category (development stage) as recommended by the content analysis methodology. We performed open coding of the excerpts where we looked for interesting themes and common patterns in the data. This resulted in 96 codes. Next, data and concepts that belonged together were grouped, forming sub-categories. Further abstraction of the data was performed by grouping sub-categories into generic categories, and those into main categories. The abstraction process was repeated for each stage of development. As mentioned earlier, during our analysis we found distinct differences in attitudes and behaviours that were easily distinguishable into two groups, we call them *the security adopters* and *the security inattentive*. We thus present the emerging themes and our analysis of the two groups independently. Figure 1 shows an example of the abstraction process for developer testing data for the security adopters. While all coding was done by a single researcher, two researchers met regularly to thoroughly and collaboratively review and edit codes, and group and interpret the data. To verify the reliability of our coding, we followed best practices by inviting a researcher who has not been involved with the project to act as a second coder, individually coding 30% of the data. We calculated Krippendorff's alpha [33] to assess inter-rater reliability, and $\alpha = 0.89$ (percentage of agreement = 91%). According to Krippendorff [34], *alpha* ≥ 0.80 indicates that coding is highly reliable and that data is "similarly interpretable by researchers". In case of disagreements, we had a discussion and came to an agreement on the codes.

Limitations: Our study included a relatively small sample size, thus generalizations cannot be made. However, our sample size followed the concept of saturation [26]; participant recruitment continued until no new themes were emerging. Additionally, recruiting participants through personal contacts could result in biasing the results. While we cannot guarantee representativeness of a larger population, the interviewer previously knew only 3/13 participants. The remaining ten participants were previously unknown to the researcher and each represented a different company. While interviews allowed us to explore topics in depth, they presented one perspective on the team. Our data may thus be influenced by participants' personal attitudes and perspectives, and may not necessarily reflect the whole team's opinions. However, we found that participants mainly described practices as encouraged by their teams.

Table 1: The degree of security in the SDLC. ● : secure, ○ : somewhat secure, × : not secure, ⊗ : not performed, ? : no data

(a) The Security Adopters

	Design	Implementation	Developer testing	Code analysis	Code review	Post-dev testing
T1	×	●	×	●	●	●
T3	?	●	?	●	●	●
T5	●	●	○	●	●	●
T11	?	●	○	●	●	?
T12	×	●	○	●	●	●
T14	×	●	●	⊗	●	●

(b) The Security Inattentive

	Design	Implementation	Developer testing	Code analysis	Code review	Post-dev testing
T2	×	●	×	○	○	●
T4	○	○	○	⊗	○	○
T6	×	○	×	×	○	○
T7	×	×	×	⊗	×	○
T8	×	×	×	⊗	×	●
T9	○	●	○	⊗	○	○
T10	○	○	×	⊗	○	⊗
T13	×	●	×	⊗	○	●
T15	×	×	×	×	×	×

4. RESULTS: SECURITY IN PRACTICE

We assess the degree of security integration in each stage of the SDLC as defined by our final analysis matrix. As mentioned earlier, we found differences in participants' attitudes and behaviours towards security that naturally fell into two distinct groups. We call the first group the *security adopters*: those who consider security in the majority of development stages (at least four stages out of six[1]). The second group who barely considered security or did not consider it at all form the *security inattentive*. We chose the term *inattentive*, as it encompasses different scenarios that led up to poor security approaches. These could be that security was considered and dismissed or it was not considered at all, whether deliberately or erroneously. Table 1 presents two heat maps, one for each group identified in our dataset (see Appendix C for more information). We classified practices during a development stage as:

(●) *secure*: when security is actively considered, *e.g.*, when developers avoid using deprecated functions during the implementation stage.
(○) *somewhat secure*: when security is not consistently considered, *e.g.*, when threat analysis is performed only if someone raises the subject.
(×) *not secure*: when security is not considered at all, *e.g.*, when developers do not perform security testing.
(⊗) *not performed*: when a stage is not part of their SDLC (*i.e.*, considered not secure).
(?): when a participant did not discuss a stage during their interview, therefore denoting missing data.

The heat maps highlight the distinction in terms of security between practices described by participants from the security adopters and the security inattentive groups. The overwhelming red and orange heat map for the security inattentive group visually demonstrates their minimal secu-

[1]At least three stages in cases where we have information about four stages only. Note that this is just a numeric representation and the split actually emerged from the data.

rity integration in the SDLC. Particularly, comparing each stage across all teams shows that even though the security adopters are not consistently secure throughout the SDLC, they are generally more attentive to security than the other group. The worst stage for the security inattentive group is *Code analysis*, which is either not performed or lacks security, followed by the *developer testing* stage, where security consideration is virtually non-existent.

We initially suspected that the degree of security integration in the SDLC would be directly proportional to the company size. However, our data suggests that it is not necessarily an influential factor. In fact, T14, the team from the smallest company in our dataset, is performing much better than T6, the team from the largest company in the security inattentive group. Additionally, we did not find evidence that development methodology influenced security practices.

Although our dataset does not allow us to make conclusive inferences, it shows an alarming trend of low security adoption in many of our project teams. We now discuss data analysis results organized by the six SDLC stages defined in our analysis matrix. All participants discussed their teams' security policies, as experienced from their perspectives, and not their personal preferences. Results, therefore, represent the reported perspectives of the developers in each team.

4.1 Exploring practices by development stage

We found that the prioritization of security falls along a spectrum: at one end, security is a main priority, or it is completely ignored at the other extreme. For each SDLC stage, we discuss how security was prioritized, present common trends, and highlight key messages from the interviews. Next to each theme we indicate which group contributed to its emergence: `(SA)` for the security adopters, `(SI)` for the security inattentive, and `(SA/SI)` for both groups. Table 2 provides a summary of the themes.

4.1.1 Design stage

We found a large gap in security practices described by our participants in the design stage. This stage saw teams at all points on the security prioritization spectrum, however, most participants indicated that their teams did not view security as part of this stage. Our inductive analysis revealed three emerging themes reflecting security prioritization, with one theme common to both the security adopters and the security inattentive, and one exclusive to each group.

Security is not considered in the design stage. `(SA/SI)` Most participants indicated that their teams did not apply security best practices in the design stage. Although they did not give reasons, we can infer from our data (as discussed in other stages) that this may be because developers mainly focus on their functional design task and often miss security [22], or because they lack the expertise to address security. As an example of the disregard for security, practices described by one participant from the security inattentive group violates the recommendation of simple design; they intentionally introduce complexity to avoid rewriting existing code, and misuse frameworks to fit their existing codebase without worrying about introducing vulnerabilities. P-T10 explained how this behaviour resulted in a highly complex code, "*Everything is so convoluted and it's like going down rabbit holes, you see their code and you are like 'why did you write it this way?' [...] It's too much different custom code that only those guys understand.*" Such complexity increases the potential for vulnerabilities and complicates subsequent stages [47]; efforts towards evaluating code security may be hindered by poor readability and complex design choices.

Security consideration in the design stage is adhoc. `(SI)` Two developers said their teams identify security considerations within the design process. In both cases, the design is done by developers who are not necessarily formally trained in security. Security issue identification is adhoc, *e.g.*, if a developer identifies a component handling sensitive information, this triggers some form of threat modelling. In T10, this takes the form of discussion in a team meeting to consider worst case scenarios and strategies for dealing with them. In T4, the team self-organizes with the developers with most security competence taking the responsibility for designing sensitive components. P-T4 said, "*Some developers are assigned the tasks that deal with authorization and authentication, for the specific purpose that they'll do the security testing properly and they have the background to do it.*" In these two teams, security consideration in the design stage lies in the hands of the developer with security expertise; this implies that the process is not very robust. If this developer fails to identify the feature as security-sensitive, security might not be considered at all in this stage.

Security design is very important. `(SA)` Contrary to all others, one team formally considers security in this stage with a good degree of care. P-T5 indicated that his team considers the design stage as their first line of defense. Developers from his team follow software security best practices [1, 8, 47], *e.g.*, they perform formal threat modelling to generate security requirements, focus on relevant threats, and inform subsequent SDLC stages. P-T5 explains the advantages of considering security from this early stage, "*When we go to do a further security analysis, we have a lot more context in terms of what we're thinking, and people aren't running around sort of defending threats that aren't there.*"

4.1.2 Implementation stage

Most participants showed general awareness of security during this stage. However, many stated that they are not responsible for security and they are not required to secure their applications. In fact, some developers reported that their companies do not expect them to have any software security knowledge. Our inductive analysis revealed three themes regarding security prioritization in this stage.

Security is a priority during implementation. `(SA/SI)` All security adopters and two participants from the security inattentive group discussed the importance of security during the implementation stage. They discussed how the general company culture encourages following secure implementation best practices and using reliable tools. Security is considered a developer's responsibility during implementation, and participants explained they are conscious about vulnerabilities introduced by errors when writing code.

Developers' awareness of security is expected when implementing. `(SA/SI)` For those prioritizing security, the majority of security adopters and one participant from the security inattentive group are expected to stay up-to-date

on vulnerabilities, especially those reported in libraries or third-party code they use. The manner of information dissemination differs and corroborates previous research findings [59]. Some have a structured approach, such as that described by P-T1, "*We have a whole system. Whenever security vulnerability information comes from a third-party, [a special team follows] this process: they create an incident, so that whoever is using the third-party code gets alerted that, 'okay, your code has security vulnerability', and immediately you need to address it.*" Others rely on general discussions between developers, *e.g.*, when they read about a new vulnerability. Participants did not elaborate on if and how they assess the credibility and reliability of information sources. The source of information could have a considerable effect on security; previous research found that relying on informal programming forums might lead to insecure code [13]. In Xiao *et al.*'s [59] study, developers reported taking the information source's thoroughness and reputation into consideration to ensure trustworthiness.

Security is not a priority during implementation. `(SI)` On the other end of the security prioritization spectrum, developers from the security inattentive group prioritize functionality and coding standards over security. Their primary goal is to satisfy business requirements of building new applications or integrating new features into existing ones. Some developers also follow standards for code readability and efficiency. However, security is not typically considered a developer's responsibility, to the extent that there are no consequences if a developer introduces a security vulnerability in their code. P-T7 explained, "*If I write a bad code that, let's say, introduced SQL injection, I can just [say] 'well I didn't know that this one introduces SQL injection' or 'I don't even know what SQL injection is'. [...] I didn't have to actually know about this stuff [and] nobody told me that I need to focus on this stuff.*" This statement is particularly troubling given that P-T7 has security background, but feels powerless in changing the perceived state of affairs in his team.

Our analysis also revealed that some developers in the security inattentive group have incomplete mental models of security. This led to the following problematic manifestations, which could explain their poor security practices.

Developers take security for granted. `(SI)` We found, aligning with previous research [22], that developers fully trust existing frameworks with their applications' security and thus take security for granted. Our study revealed that these teams do not consider security when adopting frameworks, and it is unclear if, and how, these frameworks' security is ever tested. To partially address this issue, T4 built their own frameworks to handle common security features to relieve developers of the burden of security. This approach may improve security, however verifying frameworks' security is an important, yet missing, preliminary step.

Developers misuse frameworks. `(SI)` Despite their extreme reliance on frameworks for security, developers in T10 do not always follow their recommended practices. For example, although P-T10 tries to follow them, other developers in his team do not; they occasionally overlook or work-around framework features. P-T10 explains, "*I have expressed to [the team] why I am doing things the way I am, because it's correct, it's the right way to do it with this framework. They chose to do things a completely different way, it's completely messed up the framework and their code. They don't care, they just want something that they feel is right and you know whatever.*" Such framework misuse may result in messy code and could lead to potential vulnerabilities [47]. Although frameworks have shown security benefits [52], it is evident that the manner by which some teams are currently using and relying on them is problematic.

Developers lack security knowledge. `(SI)` Developers from the security inattentive group vary greatly in their security knowledge. Some have haphazard knowledge; they only know what they happen to hear or read about in the news. Others have formed their knowledge entirely from practical experience; they only know what they happen to come across in their work. Developers' lack of software security knowledge could explain why some teams are reluctant to rely on developers for secure implementation. P-T7 said, "*I think they kind of assume that if you're a developer, you're not necessarily responsible for the security of the system, and you [do] not necessarily have to have the knowledge to deal with it.*" On the other hand, some developers have security background, but do not apply their knowledge in practice, as it is neither considered their responsibility nor a priority. P-T7 said, "*I recently took an online course on web application security to refresh my knowledge on what were the common attack on web applications [...] So, I gained that theoretical aspect of it recently and play[ed] around with a bunch of tools, but in practice I didn't actually use those tools to test my software to see if I can find any vulnerability in my own code because it's not that much of a priority.*"

Developers perceive their security knowledge inaccurately. `(SI)` We identified a mismatch between developers' perception of their security knowledge and their actual knowledge. Some developers do not recognize their secure practices as such. When asked about secure coding methods, P-T6 said, "*[The] one where we stop [cross-site scripting]. That's the only one I remember I explicitly used. Maybe I used a couple of other things without knowing they were security stuff.*" In some instances, our participants said they are not addressing security in any way. However, after probing and asking more specific questions, we identified security practices they perform which they did not relate to security.

Furthermore, we found that some developers' mental model of security revolves mainly around security functions, such as using the proper client-server communication protocol. However, conforming with previous research [59], it does not include software security. For example, P-T9 assumes that following requirements generated from the design stage guarantees security, saying "*if you follow the requirements, the code is secure. They take those into consideration.*" However, he mentioned that requirements do not always include security. In this case, and especially by describing requirements as a definite security guarantee, the developer may be referring to security functions (*e.g.*, using passwords for authentication) that he would implement as identified by the requirements. However, the developer did not discuss vulnerabilities due to implementation mistakes that are not necessarily preventable by security requirements.

Our study also revealed the following incident which illustrates how **Vulnerability discovery can motivate secu-**

rity (SI) and improve mental models. Developers in T13 became more security conscious after discovering a vulnerability in their application. P-T13 said, "*We started making sure all of our URLs couldn't be manipulated. [..] If you change the URL and the information you are looking at, [at the] server side, we'd verify that the information belongs to the site or the account you are logged in for.*" Discovering this vulnerability was eye-opening to the team; our participant said that they started thinking about their code from a perspective they had not been considering and they became aware that their code can have undesirable security consequences. In addition, this first-hand experience led them to the knowledge of how to avoid and prevent similar threats.

4.1.3 Developer testing stage

Across the vast majority of our participants, whether adopters or inattentive, security is lacking in the developer testing stage. Functionality is developers' main objective; they are blamed if they do not properly fulfil functional requirements, but their companies do not hold them accountable if a security vulnerability is discovered. P-T7 said, "*I can get away with [introducing security bugs] but with other things like just your day-to-day developer tasks where you develop a feature and you introduce a bug, that kind of falls under your responsibility. Security doesn't.*" Thus, any security-related efforts by developers are viewed as doing something extraordinary. For example, P-T2 explained, "*If I want to be the hero of the day [and] I know there's a slight possible chance that these can be security vulnerabilities, [then] I write a test and submit it to the test team.*" We grouped participants' approaches to security during this stage into four categories.

Developers do not test for security. (SA/SI) The priority at this stage is almost exclusively functionality; it increases in scope until the developer is satisfied that their code is fulfilling functional requirements and does not break any existing code. And even then, these tests vary in quality. Some developers perform adhoc testing or simply test as a sanity check where they only verify positive test cases with valid input. Others erroneously, and at times deliberately, test only ideal-case scenarios and fail to recognize worst-case scenarios. The majority of developers do not view security as their responsibility in this stage; instead they are relying on the later SDLC stages. P-T2 said, "*I usually don't as a developer go to the extreme of testing vulnerability in my feature, that's someone else's to do. Honestly, I have to say, I don't do security testing. I do functional testing.*" The participant acknowledged the importance of security testing, however, this task was considered the testing team's responsibility as they have more knowledge in this area.

Security is a priority during developer testing. (SA) As an exception, our analysis of P-T14's interview indicates that his company culture emphasizes the importance of addressing security in this stage. His team uses both automated and manual tests to ensure that their application is secure and is behaving as expected. P-T14's explained that the reason why they prefer to incorporate security in this stage was that it is more cost efficient to address security issues early in the SDLC. He explained, "*We have a small company, so it's very hard to catch all the bugs after release.*"

Developers test for security fortuitously. (SA) In other cases, security is not completely dismissed, yet it is not an explicit priority. Some security adopters run existing test suites that may include security at varying degrees. These test suites include test cases that any application is expected to pass, however, there is not necessarily a differentiation between security and non-security tests. Some developers run these tests because they are required to, without actual knowledge of their purpose. For example, P-T3 presumes that since his company did not have security breaches, security must be incorporated in existing test suites. He explained, "*[Security] has to be there because basically, if it wasn't, then our company would have lots of problems.*"

Developers' security testing is feature-driven. (SI) In another example where security is not dismissed, yet not prioritized, one participant from the security inattentive group (out of the only two who perform security testing), considers that security is not a concern as his application is not outward facing, *i.e.*, it does not involve direct user interaction. P-T9 explained, "*Security testing [pause] I would say less than 5%. Because we're doing embedded systems, so security [is] pretty low in this kind of work.*" While this may have been true in the past, the IoT is increasingly connecting embedded systems to the Internet and attacks against these systems are increasing [28]. Moreover, classifying embedded systems as relatively low-risk is particularly interesting as it echoes what Schneier [46] described as a road towards "a security disaster". On the other hand, P-T4 explained that only features that are classified as sensitive in the design stage are tested, due to the shortage in security expertise. As the company's only developer with security background, these features are assigned to P-T4. Other developers in T4 do not have security experience, thus they do not security-test their code and they are not expected to.

4.1.4 Code analysis stage

Eight developers reported that their teams have a mandatory code analysis stage. Participants from the security adopters group mentioned that the main objectives in this stage is to verify the code's conformity to standards and in-house rules, as well as detect security issues. On the other hand, participants from the security inattentive group generally do not perform this stage, and rarely for security.

Security is a priority during code analysis. (SA) All security adopters who perform this stage reported that security is a main component of code analysis in their team. T5 mandates analysis using multiple commercial tools and in-house tools before the code is passed to the next stage. T3 has an in-house tool that automates the process of analysis to help developers with the burden of security. P-T3 explained, "*[Our tool] automatically does a lot of that for us, which is nice, it does static analysis, things like that and won't even let the code compile if there are certain requirements that are not met.*" One of the advantages of automating security analysis is that security is off-loaded to the tools; P-T3 explains that security "*sort of comes for free*".

Security is a secondary objective during code analysis. (SI) P-T2 explained that in his team, developers' main objective when using a SAT is to verify conformity to industry standards. Although they might check security warnings, other security testing methods are considered more powerful. P-T2 explained, "*[SAT name] doesn't really look at the whole picture. [...] In terms of: is it similar to a security*

vulnerability testing? No. Pen testers? No. It's very weak." In addition to the lack of trust in SATs' ability to identify security issues, and similar to previous research (*e.g.*, [31]), our participants complained about the overwhelming number false positives and irrelevant warnings.

Developers rarely perform code analysis, never for security. (SI) Code analysis is not commonly part of the development process for the security inattentive group. According to their developers, T2, T6, and T15 use SATs, but not for security. Code analysis is performed as a preliminary step to optimize code and ensure readability before the code review stage, with no consideration to security.

Reasons for underusing SATs were explored in other contexts [31]. The two main reasons in our interviews were that their use was not mandated or that developers were unaware of their existence. We found that **Developers vary in awareness of analysis tools.** (SI) In addition to those unaware, some developers use SATs without fully understanding their functionality. P-T10 does not use such tools since it is not mandated and his teammates are unlikely to do so. He said, "*I know that there's tools out there that can scan your code to see if there's any vulnerability risks [...] We are not running anything like that and I don't see these guys doing that. I don't really trust them to run any kind of source code scanners or anything like that. I know I'm certainly not going to.*" Despite his awareness of the potential benefits, he is basically saying *no one else is doing it, so why should I?* Since it is not mandatory or common practice, running and analyzing SATs reports would add to the developer's workload without recognition for his efforts.

4.1.5 Code review stage

Most security adopters say that security is a primary component in this stage. Reviewers examine the code to verify functionality and to look for potential security vulnerabilities. P-T14 explained, "*We usually look for common mistakes or bad practices that may induce attack vectors for hackers such as, not clearing buffers after they've been used. On top of that, it's also [about the] efficiency of the code.*"

Contrarily, the security inattentive discount security in this stage—security is either not considered, or is considered in an informal and adhoc way and by unqualified reviewers. Code review can be as simple as a sanity check, or a walkthrough, where developers explain their code to other developers in their team. Some reviewers are thorough, while others consider reviews a secondary task, and are more inclined to accept the code and return to their own tasks. P-T10 explained, "*Sometimes they just accept the code because maybe they are busy and they don't want to sit around and criticize or critically think through everything.*" Moreover, reviewers in T9 examine vulnerabilities to assess their impact on performance. P-T9 explained, "*[Security in code review is] minimum, I'd say less than 5%. So, yeah you might have like buffer overflow, but then for us, that's more of the stability than security issue.*" We grouped participants' descriptions of the code review stage into four distinct approaches.

Code review is a formal process that includes security. (SA) All security adopters mentioned that their teams include security in this stage. For some teams, it is a structured process informed by security activities in previous stages. For example, security-related warnings flagged during the code analysis phase are re-examined during code reviews. Reviewers can be senior developers, or an independent team. Being independent, reviewers bring in a new perspective, without being influenced by prior knowledge, such as expected user input. P-T5 said, "*We do require that all the code goes through a security code review that's disconnected from the developing team, so that they're not suffered by that burden of knowledge of 'no one will do this', uh, they will.*" Sometimes reviewers might not have adequate knowledge of the applications. In such cases, T1 requires developers to explain the requirements and their implementation to the reviewers. P-T1 said, "*You have to explain what you have done and why. [...] so that they need not invest so much time to understand what is the problem [...] Then they will do a comparative study and they will take some time to go over every line and think whether it is required or not, or can it be done in some other way.*" Although cooperation between different teams is a healthy attitude, there might be a risk of developers influencing the reviewers by their explanation. P-T13 indicated the possibility of creating a bias when reviewers are walked-through the code rather than looking at it with a fresh set of eyes. He said, "*umm, I have not really thought about [the possibility of influencing the reviewers.] [...] Maybe. Maybe there is a bit.*"

Preliminary code review is done as a checkpoint before the formal review. (SA) This is an interesting example of developers collaborating with reviewers. P-T1 mentioned that reviewers sometimes quickly inspect the code prior to the formal review process and in case of a potential issue, they provide the developer with specific testing to do before the code proceeds to the review stage. This saves reviewers time and effort during the formal code review, and it could help focus the formal process on intricate issues, rather than being overwhelmed with simple ones.

Security is not considered during code review. (SI) The majority of the security inattentive participants explained that their teams' main focus for code review is assessing code efficiency and style, and verifying how well new features fulfill functional requirements and fit within the rest of the application. In fact, some participants indicated that their teams pay no attention to security during this stage. It is either not the reviewers' responsibility, or is not an overall priority for the team. P-T7 explained that because reviewers are developers, they are not required to focus on security. In addition to not being mandated, our participants explained that most developers in their teams do not have the necessary expertise to comment on security. P-T7 said, "*Probably in the two years that I've been working, I never got feedback [on] the security of my code [...] [Developers] don't pay attention to the security aspect and they can't basically make a comment about the security of your code.*"

Security consideration in code review is minimal. (SI) According to developers from the security inattentive group, some of their teams pay little attention to security during code review only by looking for obvious vulnerabilities. Additionally, this may only be performed if the feature is security-sensitive. In either case, teams do not have a formal method or plan, and reviewers do not necessarily have the expertise to identify vulnerabilities [22]. Our participants explained that reviewers are either assigned or chosen

by the developer, based on the reviewer's qualifications and familiarity with the application. However, this can have serious implications, *e.g.*, those who have security expertise will carry the burden of security reviews in addition to their regular development tasks. P-T12 explained that this caused the individuals who had security knowledge to become "*overloaded*". Although our data does not allow us to make such explorations, it is important to investigate the effect of workload on the quality of code reviews, and whether it has an effect on developers' willingness to gain security knowledge. For example, does being the person designated to do security code reviews motivate developers to gain security knowledge? Or would they rather avoid being assigned extra reviewing workload?

4.1.6 Post-development testing stage

Security is a priority during post-development testing. (SA) Three participants from the security adopters group mentioned that their project teams have their own testers that evaluate different aspects, including security. The general expectation is that the testers would have some security knowledge. Additionally, P-T12 mentioned that his company hires external security consultants for further security testing of their applications. However, because the testing process by such experts is usually "*more expensive and more thorough,*" (P-T12), they usually postpone this step until just before releasing the application. We identified two distinct motivations for performing security testing at this stage: **Post-development testing is used to discover security vulnerabilities, or for final verification.** (SA) Unsurprisingly, the majority of security adopters rely on post-development testing as an additional opportunity to identify and discover security vulnerabilities before their applications are put out to production. T1, on the other hand, expects security post-development testing to reveal zero vulnerabilities. P-T1 explained, "*If they find a security issue, then you will be in trouble. Everybody will be at your back, and you have to fix it as soon as possible.*" Thus, this stage is used as a final verification that security practices in the previous stages were indeed successful in producing a vulnerability-free application.

Similar to the code review stage, we found evidence of collaboration between the development and the testing team, however, **Testers have the final approval.** (SA). Testers would usually discuss with developers to verify that they understand the requirements properly, since they do not have the same familiarity with the application as developers. However, P-T5 explained that although developers can challenge the testing team's analysis, they cannot dismiss their comments without justification. Addressing security issues is consistently a priority. P-T5 said, "*[The testing team will] talk to the development teams and say, 'here's what we think of this', and the development team will sometimes point out and say, 'oh, you missed this section over here' [...] but one of the things is, we don't let the development teams just say, 'oh, you can't do that because we don't want you to'. So the security teams can do whatever they want.*" Cooperation between developers and testers could help clear ambiguities or misunderstandings. In T5 testers have some privilege over developers; issues raised by testers have to be addressed by developers, either by solving them or justifying why they can be ignored. P-T5 hinted that disagreements may arise between different teams, but did not detail how they are resolved. Further exploration of this subject is needed, taking into consideration the level of security knowledge of the development team compared to the testing team.

Security is prioritized in post-development testing for all of our security adopters, where they rely on an independent team to test the application as a whole. On the other hand, although post-development testing appears to be common to all teams from the security inattentive group (with the exception of T10), it often focuses primarily on functionality, performance and quality analysis, with little to no regard for security. Our analysis revealed the following insights and approaches to post-development security testing.

Security is not considered in post-development testing. (SI) According to their developers, two teams (T10, T15) do not consider security during this stage. T10 does not perform any testing, security or otherwise. The company to which T15 belongs has its own Quality Analysis (QA) team, though they do not perform security testing. P-T15 said, "*I've never seen a bug related to security raised by QA.*" The case of T15 is particularly concerning; many teams rely on this stage to address software security, while T15 does not. According to our data, security is not part of the development lifecycle in T15. It would be interesting to further explore why some teams completely ignore software security, and what factors could encourage them to adopt a security initiative.

Post-development testing plans include a security dimension. (SI) As mentioned earlier, P-T2 relies mainly on this stage for security testing, In addition, P-T6, and P-T13 say that their teams consider security during this stage. However, there seems to be a disconnect between developers and testers in T6; developers are unaware of the testing process and consider security testing out of scope. Despite her knowledge that security is included in this stage, P-T6 mentioned, "*I don't remember any tester coming back and telling [me] there are [any] kinds of vulnerability issues.*" T13 started integrating security in their post-development testing after a newly hired tester who decided to approach the application from a different perspective discovered a serious security issue. P-T13 explained, "*No one had really been thinking about looking at the product from security standpoint and so the new tester we had hired, he really went at it from 'how can I really break this thing?' [..] and found quite a few problems with the product that way.*" The starting point of security testing in T13 was a matter of chance. When an actual security issue was discovered in their code, security was brought to the surface and post-development testing started addressing security.

Through our analysis, we found that along the security prioritization spectrum, there are cases where security in this stage is driven by different factors, as explained below.

Some participants discussed that their team relies on a single person to handle security, thus security consideration is driven by specific factors. For example, in T4, **Post-development security testing is feature-driven.** (SI). P-T4 is the only developer in his company with security expertise, thus he is responsible for security. He explained that his company has limited resources and few employees, thus they focus their security testing efforts only on security-

sensitive features (*e.g.*, authentication processes), as flagged by the developers. Thus, the question is how reliable are assessments in this case given that they are done by developers with limited security expertise? On the other hand, in T7, **Post-development security testing is adhoc.** (SI) . P-T7 explained that they rely on a single operations-level engineer who maintains the IT infrastructure and handles security testing. Thus, testing is unplanned and could happen whenever the engineer has time or "*whenever he decides.*" P-T7 erroneously [50] presumes their applications are risk-free since they are a "*small company*", and thus they are not an interesting target for cyberattacks. Company size was used by some of our participants to justify their practices in multiple instances. Although in our data we did not find evidence to support that company size affects actual security practices, it shows our participants' perception.

We also found that an external mandate to the company can be a driving factor for security consideration. For example, P-T8 reported that his company needs to comply with certain security standards, thus his team performs security testing when they are expecting an external audit "*to make sure the auditors can't find any issue during the penetration test.*" In this case, **Post-development security testing is externally-driven.** (SI) Such external pressure by an overseeing entity was described as "*the main*" driving factor to schedule security testing; P-T8 explained that if it were not for these audits, his team would not have bothered with security tests. Mandating security thus proved to be effective in encouraging security practices in a team that was not proactively considering it.

As evidenced by our data, the security inattentive group's security practices, if existent, are generally informal, unstructured, and not necessarily performed by those qualified. The main focus is delivering features to customers; security is not necessarily a priority unless triggered, *e.g.*, by experiencing a security breach or expecting an external audit.

4.2 The adopters vs. the inattentive

In general, security practices appear to be encouraged in teams to which the security adopters belong. In contrast, as explained by participants from the security inattentive group, their teams' main priority is functionality; security is an afterthought. Contrary to a trend towards labelling developers as "the weakest link" [27], our analysis highlights that poor security practices is a rather complex problem that extends beyond the developer. Just as we have identified instances where developers lack security knowledge or lack motivation to address security, we have also identified instances where security was ignored or dismissed by developers' supervisors, despite the developer's expertise and interest. It is especially concerning when security is dismissed by those high in the company hierarchy. As an extreme case, P-T15 reported zero security practices in their SDLC; she explained "*To be honest, I don't think anybody cares about [security]. I've never heard or seen people talk about security at work [...] I did ask about this to my managers, but they just said 'well, that's how the company is. Security is not something we focus on right now.'*"

It was interesting to find that all our participants who identified themselves as developers of web applications and services, *i.e.*, in their current daily duties, (namely, P-T4, P-T6, P-T7, P-T8, P-T10, P-T13, P-T15) fall in the security inattentive group. Specific reasons for this are unclear. It may be because web-development is generally less mature and has a quick pace [44], and teams are eager to roll-out functionality to beat their competitors. In such cases, functional requirements may be prioritized and security may be viewed as something that can be addressed as an update, essentially gambling that attackers will miss any vulnerabilities in the intervening time. Teams who have not yet become victims may view this as a reasonable strategy, especially since patching generally does not requires end-user involvement (*e.g.*, web server fixes do not require users to update their software), making it a less complicated process. However, since participants building other types of software also fall in the security inattentive group, it is hard to draw a generic conclusion that web-development is particularly insecure.

Table 2 summarizes the themes that emerged from our analysis. As expected, we found conflicting themes between the security adopters and the security inattentive group, where the more secure themes consistently belongs to the security adopters. However, our analysis also revealed common themes (see Table 2), some of which are promising while others are problematic for security. On the positive side, participants from both groups discussed developers' role in security during implementation. On the other hand, participants from both groups also indicated a lack of attention to security in the design stage. Reasons leading to these common themes sometimes vary. Consider the theme *Developers do not test for security*; the security inattentive group ignored security testing because developers often lack the knowledge necessary to perform this task. Whereas for the security adopters, the reason is that security testing is not included in developers' tasks even if they have the required knowledge. In Section 6.2 we discuss factors that we identified as influential to security practices.

5. INITIATIVES AND BEST PRACTICES

After exploring real life security practices, how do these compare to security best practices? To answer, we offer background on popular sources of best practices. We then amalgamate them into a concise list of the most common recommendations. In Section 6, we discuss the relationship between practices found in our study and best practices.

5.1 Secure SDLC initiatives

This section gives a brief background on prominent processes and recommendations for secure software development.

Security Development Lifecycle (SDL). Microsoft SDL [8] is the first initiative to encourage the integration of security in the SDLC from the early stages. It consists of 16 security practices and can be employed regardless of the platform.

Building Security In Maturity Model (BSIMM). Currently maintained by Cigital [2], the BSIMM [6] recommends 12 main security practices. It provides high-level insights to help companies plan their secure SDLC initiative and assess their security practices compared to other organizations.

Open Web Application Security Project (OWASP) initiatives. OWASP's Software Assurance Maturity Model (SAMM) [3] recognizes 4 main classes of SDLC activities and provides 3 security best practices for each. Additionally, the Developer Guide [1] provides best practices for architects

Table 2: Summary of themes emerging from the security adopters and the security inattentive, and common themes between the two groups. Although common themes exist, driving factors for these themes may differ. See Section 4.2 for more details.

Security Adopters Themes	Common Themes	Security Inattentive Themes
	Design	
· Security design is very important	· Security is not considered in the design stage	· Security consideration in the design stage is adhoc
	Implementation	
	· Security is a priority during implementation · Developers' awareness of security is expected when implementing	· Security is not a priority during implementation · Developers take security for granted · Developers misuse frameworks · Developers lack security knowledge · Developers perceive their security knowledge inaccurately · Vulnerability discovery can motivate security
	Developer Testing	
· Developers test for security fortuitously · Security is a priority during developer testing	· Developers do not test for security	· Developers' security testing is feature-driven
	Code Analysis	
· Security is a priority during code analysis		· Security is a secondary objective during code analysis · Developers rarely perform code analysis, never for security · Developers vary in awareness of analysis tools
	Code Review	
· Code review is a formal process that includes security · Preliminary code review is done as a checkpoint before the formal review		· Security is not considered during code review · Security consideration in code review is minimal
	Post-development Testing	
· Security is a priority during post-development testing · Post-development testing is used to discover security vulnerabilities, or for final verification · Testers have the final approval		· Security is not considered in post-development testing · Post-development testing plans include a security dimension · Post-development security testing is feature-driven · Post-development security testing is adhoc · Post-development security testing is externally-driven

and developers, whereas the Testing Guide [4] focuses on best practices for testing and evaluating security activities.

Others. Additional resources for security best practices include: NASA's *Software Assurance Guidebook* [39], NIST's *Special Publication 800-64* [32], US-CERT's *Top 10 Secure Coding Practices* [47], as well as various articles emphasizing the importance of secure development [7, 36, 37, 57].

5.2 Security Best Practices

Available resources for security best practices vary in their organization and their presentation style, *e.g.*, they vary in technical details. Practitioners may find difficulty deciding on best practices to follow and establishing processes within their organizations [38, 42, 54]. To help frame security practices we identified, we collected recommendations from the sources discussed in Section 5.1 to compose a concise set of best practices. This resulted in an initial set of 57 unorganized recommendations varying in format and technical details. We then grouped related recommendations, organized them in high-level themes, and iterated this process to finally produce the following 12 best practices. Other amalgamations may be possible, but we found this list helpful to interpret our study results. The list could be of independent interest to complementary research in this area.

B1 **Identify security requirements.** Identify security requirements for your application during the initial planning stages. The security of the application throughout its different stages should be evaluated based on its compliance with security requirements.

B2 **Design for security.** Aim for simple designs because the likelihood of implementation errors increases with design complexity. Architect and design your software to implement security policies and comply with security principles such as: secure defaults, default deny, fail safe, and the principle of least privilege.

B3 **Perform threat modelling.** Use threat modelling to analyze potential threats to your application. The result of threat modelling should inform security practices in the different SDLC stages, *e.g.*, for creating test plans.

B4 **Perform secure implementation.** Adopt secure coding standards for the programming language you use, *e.g.*, validate input and sanitize data sent to other systems, and avoid using unsafe or deprecated functions.

B5 **Use approved tools and analyze third-party tools' security.** Only use approved tools, APIs, and frameworks or those evaluated for security and effectiveness.

B6 **Include security in testing.** Integrate security testing in functional test plans to reduce redundancy.

B7 **Perform code analysis.** Leverage automated tools such as SATs to detect vulnerabilities like buffer overflows and improper user input validation.

B8 **Perform code review for security.** Include security in code reviews and look for common programming errors that can lead to security vulnerabilities.

B9 **Perform post-development testing.** Identify security issues further by using a combination of methods, *e.g.*, dynamic analysis, penetration testing, or hiring external security reviewers to bring in a new perspective.

B10 **Apply defense in depth.** Build security in all stages of the SDLC, so that if a vulnerability is missed in one stage, there is a chance to eliminate it through practices implemented in the remaining stages.

B11 **Recognize that defense is a shared responsibility.** Address software security as a collective responsibility of all SDLC entities, *e.g.*, developers, testers, and designers.

B12 **Apply security to all applications.** Secure low risk applications and high risk ones. The suggested effort spent on security can be derived from assessing the value of assets and the risks, however, security should not be ignored in even the lowest risk applications.

6. INTERPRETATION OF RESULTS

In this section, we compare security practices from our study to best practices, present factors influencing those practices, and discuss future research directions. We comment on teams' practices as described by their developers (our participants), recognizing that we have only one perspective per team. Compliance (or lack thereof) to all best practices is not proof of a secure (or insecure) SDLC. However, this list of widely agreed upon best practices allows us to make preliminary deduction on the software security status quo.

6.1 Current practices versus best practices

Our analysis showed different approaches to security and varying degrees of compliance with best practices. The best practice with most compliance is B9; almost all participants reported that their team performs security post-development testing (to varying degrees). Contrarily, most do not *apply defense in depth* (B10); the security adopters do not consistently integrate security throughout the SDLC and the security inattentive group relies mainly on specific stages to verify security (*e.g.*, post-development testing). In addition, security is generally not a part of the company culture for the security inattentive group and they commonly delegate a specific person or team to be solely responsible for security. This leads to adhoc processes and violates B11: *recognize that defense is a shared responsibility*. Moreover, the security inattentive group violate B12 by ignoring security in applications considered low-risk without evidence that they performed proper risk analysis.

Deviations from best practices are apparent even from the design stage. The majority of participants indicate that their teams do not address security during design, contradicting B1–B3. Some developers may even deliberately violate the *Design for security* best practice (B2) to achieve their business goals and avoid extra work. On the other hand, the two participants who discussed formal consideration of security in design claim the advantages of having more informed development processes, identifying all relevant threats and vulnerabilities, and not getting distracted by irrelevant ones [47].

The implementation stage is particularly interesting; it shows the contradictions between the security adopters and the security inattentive. Participants from both groups *perform secure implementation* (B4), yet this only applied to three security inattentive participants. For most of the security inattentive group, *security is not a priority* and *developers take security for granted*, assuming that frameworks will handle security. While frameworks have security benefits [52], each has its own secure usage recommendations (*e.g.*, [5]), often buried in their documentations, and it is unclear if developers follow them. In fact, our study suggests that *developers misuse frameworks* by circumventing correct usage to more easily achieve their functional goals, another violation of B4. Moreover, despite their reliance on frameworks, participants report that security is not factored in their teams' framework choices (violating B5).

We found non-compliance with best practices in other development stages as well. For example, some teams do not include security in their functional testing plans, violating B6, and some teams do not perform code analysis, violating B7. Ignoring code analysis is a missed opportunity for automatic code quality analysis and detection of common programming errors [17]. Participants who said their teams use security code analysis tools, do so to focus subsequent development stages on the more unusual security issues. Others do not review their code for security (violating B8); rather code review is mainly functionality-focused. In some cases, participants said that reviewers do not have the expertise to conduct security reviews, in others they maybe overloaded with tasks, and sometimes code review plans simply do not include security.

6.2 Factors affecting security practices

Through close inspection of our results and being immersed in participants' reported experiences, we recognized factors that appear to shape their practices and that may not be adequately considered by best practices. We present each factor and its conflict with best practices, if applicable.

Division of labour. Best practices conflict with some of our teams' division of labour styles. Participants explained that some teams violate the *Apply defense in depth* (B10) best practice because applying security in each SDLC stage conflicts with their team members' roles and responsibilities. In some teams, developers are responsible for the functional aspect (*i.e.*, implementation and functional testing) and testers handle security testing. These teams are also violating B6, because integrating security in functional testing plans would conflict with the developers' assigned tasks. Complying with these best practices likely means they need to change the team's structure and re-distribute the assigned responsibilities. Teams may be reluctant to make such changes [42] that may conflict with their software development methodologies [35], especially since security is not their primary objective [27].

Security knowledge. We found that the expectation of security knowledge (or lack thereof) directly affects the degree of security integration in developers' tasks. When security knowledge was expected, participants said that developers were assigned security tasks (*e.g.*, performing security testing). On the other hand, we found that developers' (expected) lack of security knowledge resulted in lax security practices (*Security is not considered in the design stage*, *Security is not a priority during implementation*, *Developers do not test for security*, and *Security is not considered during code review*). While these violate best practices (*e.g.*, B1, B4 B6, B8), it is unrealistic to rely on developers to perform security tasks while lacking the expertise. From teams' perspective, they are relieving developers from the security burden. This may be a reasonable approach, loosely following recommendations of taking the developer out of the security loop when possible [14, 27]. Another obvious, yet complicated, answer would be to educate developers [8]. However, companies may lack the resources to offer security training, and there is evidence that developers remain focused mainly

on their primary functional task and not security [22].

Company culture. Another influential factor indicated by participants is the teams' cognizance of security and whether it is part of the company culture. In teams where security was reportedly advocated, developers spoke of security as a shared responsibility (conforming with B11). In instances where security was dismissed, participants said that developers did not consider security, and even those with security knowledge were reluctant to apply it. For successful adoption of security, initiatives should emerge from upper management and security should be rooted in the company's policies and culture. Developers are more likely to follow security practices if mandated by their company and its policies [59]. Integrating and rewarding security in the company culture can significantly motivate security practices [58, 59], compared to instances where security is being viewed as something that only "heroes" do if there is time.

Resource availability. Some participants said their team decides their security practices based on the available budget and/or employees who can perform security tasks. As reported, some teams violate B10 as they do not have enough employees who can perform all the recommended security tasks in addition to their original workload. Also, others reportedly violate B9, because they neither have the budget to hire external penetration testers, nor do their members have the expertise to perform such post-development tests. For such companies, the price for conforming with these best practice is too steep for little perceived gain. In other cases, participants said their team strains their resources in ways that can be detrimental. For example, the one developer with the most security knowledge is handed responsibility to identify security-sensitive features and to verify the security of the team's code. This is a significant burden, yet with little support or guidance. Besides the obvious security risks of such an approach, it may also lead to employee fatigue and ultimately to the loss of valuable team members.

External pressure. Monitoring by an overseeing entity can drive teams to adopt security practices to ensure they comply with its standards. Encouraging security practices through external mandates is not new, *e.g.*, the UK government mandated that applications for the central government should be tested using the National Technical Authority for Information Assurance CHECK scheme [11]. As a result of this initiative, companies have improved their management and response to cyber threats [10]. It would be interesting to explore how to mandate security practices in companies, and how governments and not-for-profit agencies could support teams, particularly those from the security inattentive group, to become more secure.

Experiencing a security incident. Participants reported that discovering a vulnerability or experiencing a security breach first-hand is another factor that encouraged security practices and awareness in their teams. Despite extensive publicity around security vulnerabilities, awareness of and commitment to security remains low [45]. Our analysis shows that direct vulnerability discovery influenced security practices more than hearing news-coverage of high-profile vulnerabilities (*e.g.*, [21, 53]). This can be explained by the optimistic bias [55]: the belief that "misfortune will not strike me" [45]. Rhee *et al.* [45] found that the optimistic bias strongly influences perception of security risks in Information Technology (IT). It is even greater when the misfortune seems distant, without a close comparison target. Thus, to overcome such bias, security training and awareness has to reach all levels–from upper management to those directly involved in the development process. Similar to Harbach and Smith's [29] personalized privacy warnings which led users to make more privacy-aware decisions, software security training should be personalized and provide concrete examples of the consequences of these threats to the company. We recommend that training should also not focus exclusively on threats; it should provide concrete proactive steps with expected outcomes. Additionally, it should include case studies and first-hand accounts of security incidents, and approaches to overcome them. Hence, security training moves from the theoretical world to the real world, aiding in avoiding the optimism bias.

6.3 Future research directions

Security best practices advocate for integrating security starting from the early SDLC stages. However, with limited resources and expertise, if a team can only address security in post-development testing, is this team insecure? Or might this testing be sufficient? Is the security inattentive group in our dataset really guilty of being insecure? Or did they just find the cost of following security best practices too steep? Available best practices fail to discuss the baseline for ensuring security, or how to choose which best practices to follow based on limited resources and expertise. It was also interesting to find that most security best practices are from industry sources and are not necessarily empirically verified.

For future research, we suggest devising a lightweight version of security best practices and evaluating its benefit for teams that do not have enough resources to implement security throughout the SDLC, or when implementing traditional security practices would be too disruptive to their workflow. Additionally, teams that succeeded at building a security-oriented culture should be further explored to better understand how others can adopt their approach. Further exploration of how to incorporate security in the company culture and evaluating its benefits can be a starting point for more coherent security processes, since developers are more likely to follow security practices if mandated by their company and its policy [59]. Particularly, what lessons can be carried from the security adopters over to the security inattentive group? Our work explores some of the issues surrounding secure development practices. Surveys with a larger sample of companies and more stakeholders would be an interesting next step.

7. CONCLUSION

Through interviews with developers, we investigated SDLC practices relating to software security. Our analysis showed that real-life security practices differ markedly from best practices identified in the literature. Best practices are often ignored, simply since compliance would increase the burden on the team; in their view, teams are making a reasonable cost-benefit trade-off. Rather than blaming developers, our analysis shows that the problem extends up in company hierarchies. Our results highlight the need for new, lightweight best practices that take into account the realities and pressures of development. This may include additional automation or rethinking of secure programming practices to ease the burden on humans without sacrificing security.

8. ACKNOWLEDGMENTS

We thank our participants for their time. H. Assal acknowledges her NSERC Postgraduate Scholarship (PGS-D). S. Chiasson acknowledges funding from NSERC for her Canada Research Chair and Discovery grants.

9. REFERENCES

[1] https://www.owasp.org/index.php/Category:OWASP_Guide_Project.

[2] https://www.cigital.com.

[3] www.owasp.org/index.php/OWASP_SAMM_Project.

[4] www.owasp.org/index.php/OWASP_Testing_Project.

[5] AngularJS Developer Guide. https://docs.angularjs.org/guide/security.

[6] BSIMM. https://www.bsimm.com.

[7] Cybersecurity Engineering. https://www.cert.org/cybersecurity-engineering/.

[8] Security Development Lifecycle. https://www.microsoft.com/en-us/sdl.

[9] *Content analysis for the social sciences and humanities*. Addison-Wesley Publishing Co., 1969.

[10] Cyber security boost for UK firms. https://www.gov.uk/government/news/cyber-security-boost-for-uk-firms, 2015.

[11] IT Health Check (ITHC): supporting guidance. https://www.gov.uk/government/publications/it-health-check-ithc-supporting-guidance/it-health-check-ithc-supporting-guidance, 2015.

[12] Y. Acar, M. Backes, S. Fahl, S. Garfinkel, D. Kim, M. L. Mazurek, and C. Stransky. Comparing the usability of cryptographic apis. In *Proceedings of the 38th IEEE Symposium on Security and Privacy*, 2017.

[13] Y. Acar, M. Backes, S. Fahl, D. Kim, M. L. Mazurek, and C. Stransky. You get where you're looking for: The impact of information sources on code security. In *IEEE Symp. on Security and Privacy*, 2016.

[14] Y. Acar, S. Fahl, and M. L. Mazurek. You are not your developer, either: A research agenda for usable security and privacy research beyond end users. In *2016 IEEE Cybersecurity Development (SecDev)*, pages 3–8, Nov 2016.

[15] Y. Acar, C. Stransky, D. Wermke, C. Weir, M. L. Mazurek, and S. Fahl. Developers need support, too: A survey of security advice for software developers. In *Cybersecurity Development (SecDev), 2017 IEEE*, pages 22–26. IEEE, 2017.

[16] H. Assal, S. Chiasson, and R. Biddle. Cesar: Visual representation of source code vulnerabilities. In *VizSec'16*, pages 1–8, Oct.

[17] N. Ayewah, D. Hovemeyer, J. D. Morgenthaler, J. Penix, and W. Pugh. Using static analysis to find bugs. *IEEE Software*, 25(5):22–29, Sept 2008.

[18] M. Backes, K. Rieck, M. Skoruppa, B. Stock, and F. Yamaguchi. Efficient and flexible discovery of php application vulnerabilities. In *2017 IEEE European Symposium on Security and Privacy (EuroS P)*, pages 334–349, April 2017.

[19] G. Berisha and J. Shiroka Pula. Defining small and medium enterprises: a critical review. *Academic Journal of Business, Administration, Law and Social Sciences*, 1, 2015.

[20] B. Chess and G. McGraw. Static Analysis for Security. *IEEE Security & Privacy*, 2(6):76–79, 2004.

[21] Codenomicon. The Heartbleed Bug. http://heartbleed.com.

[22] D. Oliveira *et al.* It's the Psychology Stupid: How Heuristics Explain Software Vulnerabilities and How Priming Can Illuminate Developer's Blind Spots. In *ACSAC '14*, pages 296–305. ACM, 2014.

[23] S. Elo and H. Kyngäs. The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1):107–115, 2008.

[24] A. Forward and T. C. Lethbridge. A taxonomy of software types to facilitate search and evidence-based software engineering. In *Proceedings of the 2008 Conference of the Center for Advanced Studies on Collaborative Research: Meeting of Minds*, CASCON '08, pages 14:179–14:191, New York, NY, USA, 2008. ACM.

[25] D. Geer. Are Companies Actually Using Secure Development Life Cycles? *Computer*, 43(6):12–16, June 2010.

[26] B. G. Glaser and A. L. Strauss. *The discovery of grounded theory: strategies for qualitative research*. Aldine, 1967.

[27] M. Green and M. Smith. Developers are not the enemy!: The need for usable security apis. *IEEE Security Privacy*, 14(5):40–46, Sept 2016.

[28] A. Greenberg. Hackers Remotely Kill a Jeep on the Highway—With Me in It. https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/, 2015.

[29] M. Harbach, M. Hettig, S. Weber, and M. Smith. Using personal examples to improve risk communication for security & privacy decisions. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2647–2656, New York, NY, USA, 2014. ACM.

[30] M. Howard and S. Lipner. *The security development lifecycle: SDL, a process for developing demonstrably more secure software*. Microsoft Press, Redmond, Wash, 2006.

[31] B. Johnson, Y. Song, E. Murphy-Hill, and R. Bowdidge. Why don't software developers use static analysis tools to find bugs? In *ICSE*, 2013.

[32] R. Kissel, K. Stine, M. Scholl, H. Rossman, J. Fahlsing, and J. Gulick. *Security considerations in the system development life cycle*. 2008.

[33] K. Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70, 1970.

[34] K. Krippendorff. Testing the reliability of content analysis data. *The content analysis reader*, pages 350–357, 2009.

[35] R. C. Martin. *Agile software development: principles, patterns, and practices*. Prentice Hall, 2002.

[36] G. McGraw. *Software security: building security in*. Addison-Wesley, Upper Saddle River, NJ, 2006.

[37] G. McGraw. Seven myths of software security best practices. http://searchsecurity.techtarget.com/opinion/McGraw-Seven-myths-of-software-security-best-practices, 2015.

[38] P. Morrison. A Security Practices Evaluation Framework. In *Proceedings of the 37th International*

Conference on Software Engineering, ICSE '15, pages 935–938, Piscataway, NJ, USA, 2015. IEEE Press.

[39] NASA. Software Assurance Guidebook, NASA-GB-A201. https://www.hq.nasa.gov/office/codeq/doctree/nasa_gb_a201.pdf, 2002.

[40] D. C. Nguyen, D. Wermke, Y. Acar, M. Backes, C. Weir, and S. 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.

[41] H. Perl, S. Dechand, M. Smith, D. Arp, F. Yamaguchi, K. Rieck, S. Fahl, and Y. Acar. VCCFinder: Finding Potential Vulnerabilities in Open-Source Projects to Assist Code Audits. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 426–437, New York, NY, USA, 2015. ACM.

[42] A. Poller, L. Kocksch, S. Türpe, F. A. Epp, and K. Kinder-Kurlanda. Can security become a routine?: A study of organizational change in an agile software development group. In *ACM CSCW*, 2017.

[43] J. Radcliffe. Hacking Medical Devices for Fun and Insulin: Breaking the Human SCADA System. https://media.blackhat.com/bh-us-11/Radcliffe/BH_US_11_Radcliffe_Hacking_Medical_Devices_WP.pdf.

[44] J. Ratner. Human factors and Web development, 2003.

[45] H.-S. Rhee, Y. U. Ryu, and C.-T. Kim. Unrealistic optimism on information security management. *Computers & Security*, 31(2):221 – 232, 2012.

[46] B. Schneier. Security Risks of Embedded Systems. https://www.schneier.com/blog/archives/2014/01/security_risks_9.html.

[47] R. Seacord. Top 10 secure coding practices. https://www.securecoding.cert.org/confluence/display/seccode/Top+10+Secure+Coding+Practices, 2011.

[48] J. Smith, B. Johnson, E. Murphy-Hill, B. Chu, and H. R. Lipford. Questions developers ask while diagnosing potential security vulnerabilities with static analysis. In *ESEC/FSE 2015*, pages 248–259. ACM, 2015.

[49] I. Sommerville. *Software engineering*. Pearson, Boston, 2011.

[50] J. Sophy. 43 Percent of Cyber Attacks Target Small Business. https://smallbiztrends.com/2016/04/cyber-attacks-target-small-business.html, 2016.

[51] C. Stransky, Y. Acar, D. C. Nguyen, D. Wermke, D. Kim, E. M. Redmiles, M. Backes, S. Garfinkel, M. L. Mazurek, and S. 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)*, Vancouver, BC, 2017. USENIX Association.

[52] S. Streichsbier. Improve Web Application Security with Frameworks: A case study. http://www.vantagepoint.sg/blog/18-improve-web-application-security-with-frameworks-a-case-study.

[53] Symantec Security Response. ShellShock: All you need to know about the Bash Bug vulnerability. http://www.symantec.com/connect/blogs/shellshock-all-you-need-know-about-bash-bug-vulnerability, 2014.

[54] I. A. Tondel, M. G. Jaatun, and P. H. Meland. Security Requirements for the Rest of Us: A Survey. *IEEE Software*, 25(1):20–27, Jan 2008.

[55] N. D. Weinstein and W. M. Klein. Unrealistic optimism: Present and future. *Journal of Social and Clinical Psychology*, 15(1):1–8, 2017/08/12 1996.

[56] J. Witschey, S. Xiao, and E. Murphy-Hill. Technical and personal factors influencing developers' adoption of security tools. In *ACM SIW*, 2014.

[57] C. Woody. Strengthening Ties Between Process and Security. https://www.us-cert.gov/bsi/articles/knowledge/sdlc-process/strengthening-ties-between-process-and-security#touch, 2013.

[58] G. Wurster and P. C. van Oorschot. The developer is the enemy. In *Proceedings of the 2008 New Security Paradigms Workshop*, NSPW '08, pages 89–97, New York, NY, USA, 2008. ACM.

[59] S. Xiao, J. Witschey, and E. Murphy-Hill. Social influences on secure development tool adoption: Why security tools spread. In *ACM CSCW*, 2014.

[60] J. Xie, H. R. Lipford, and B. Chu. Why do programmers make security errors? In *VL/HCC*, pages 161–164, Sept 2011.

APPENDIX

A. INTERVIEW SCRIPT

The following questions represent the main themes discussed during the interviews. We may have probed for more details depending on participants' responses.

- What type of development do you do?
- What are your main priorities when doing development? (In order of priority)
- Do your priorities change when a deadline approaches?
- What about security? Is it something you worry about?
- How does security fit in your priorities?
- Which software security best practices are you familiar with?
- Are there any obligations by your supervisor/employer for performing security testing?
- What methods do you use to try to ensure the security of applications?
- Do you perform testing on your (or someone else's) applications/code?
- Do you perform code reviews?

B. PARTICIPANTS DEMOGRAPHICS

Table 3: Participants demographics

	Participant					Company and team	
Participant ID	Gender	Age	Years	Title	SK	Company size	Team size[1]
P-T1	F	30	1	Software engineer	4	Large enterprise	20
P-T2	M	34	15	Software engineer	5	Large enterprise	12
P-T3	M	33	10	Software engineer	4	Large enterprise	10
P-T4	M	38	21	Software developer	4	SME	7
P-T5	M	34	12	Product manager	5	Large enterprise	7
P-T6	F	26	3	Software engineering analyst	3	Large enterprise	12
P-T7, P-T8*	M	33	4	Senior web engineer	4	SME – n/a*	3
P-T9	M	34	5	Software developer	3	Large enterprise	20
P-T10, P-T11*	M	33	8	Software engineer	2	SME – SME*	5
P-T12	M	37	20	Principal software engineer	5	SME	10
P-T13	M	38	15	Senior software developer	2	SME	8
P-T14	M	26	3	Software developer	2	SME	4
P-T15	F	27	5	Junior software developer	4	Large enterprise	7

Years: years of experience in development
SK: self-rating of security knowledge 1(no knowledge) - 5(expert)
$\star$: indicates participant's previous company
SME: Small-Medium Enterprise
[1] Team size for the current company

C. DEGREE OF SECURITY IN THE SDLC

Table 4: Extending Table 1 to show the degree of security in the SDLC and the application type. ● : secure, ○ : somewhat secure, × : not secure, ⊗ : not performed, ? : no data

(a) The Security Adopters

Application		Design	Implementation	Developer testing	Code analysis	Code review	Post-dev testing
embedded software	T1	×	●	×	●	●	●
design and engineering software	T3	?	●	?	●	●	●
design and engineering software	T5	●	●	○	●	●	●
info. management & decision support	T11	?	●	○	●	●	?
support utilities	T12	×	●	○	●	●	●
support utilities	T14	×	●	●	⊗	●	●

(b) The Security Inattentive

Application		Design	Implementation	Developer testing	Code analysis	Code review	Post-dev testing
kernels	T2	×	●	×	○	○	●
website content management	T4	○	○	○	⊗	○	○
e-finance	T6	×	○	×	×	○	○
online productivity	T7	×	×	×	⊗	×	○
social networking	T8	×	×	×	⊗	×	●
embedded software	T9	●	●	○	⊗	○	○
online booking	T10	○	○	×	⊗	○	⊗
online productivity	T13	×	●	×	⊗	○	●
online productivity	T15	×	×	×	×	×	×

Deception Task Design in Developer Password Studies: Exploring a Student Sample

Alena Naiakshina
University of Bonn
naiakshi@cs.uni-bonn.de

Anastasia Danilova
University of Bonn
danilova@cs.uni-bonn.de

Christian Tiefenau
University of Bonn
tiefenau@cs.uni-bonn.de

Matthew Smith
University of Bonn
smith@cs.uni-bonn.de

ABSTRACT

Studying developer behavior is a hot topic for usable security researchers. While the usable security community has ample experience and best-practice knowledge concerning the design of end-user studies, such knowledge is still lacking for developer studies. We know from end-user studies that task design and framing can have significant effects on the outcome of the study. To offer initial insights into these effects for developer research, we extended our previous password storage study [42]. We did so to examine the effects of deception studies with regard to developers. Our results show that there is a huge effect - only 2 out of the 20 non-primed participants even attempted a secure solution, as compared to the 14 out of 20 for the primed participants. In this paper, we will discuss the duration of the task and contrast qualitative vs. quantitative research methods for future developer studies. In addition to these methodological contributions, we also provide further insights into why developers store passwords insecurely.

1. INTRODUCTION

Applying the philosophy and methods of usable security and privacy research to developers [31] is still a fairly new field of research. As such, the community does not yet have the body of experience concerning study design that it does for end-user studies. Many factors need to be considered when designing experiments. In what setting should they be conducted: a laboratory, online, or in the field? Who should the participants be: computer science students, or professional administrators and developers? Is a longitudinal study needed, or is a first contact study sufficient? Should a qualitative or quantitative approach be taken? How many participants are needed and can realistically be recruited? Is deception necessary to elicit unbiased behavior? How big do tasks need to be? And so forth. All these factors have an influence on the ecological validity of studies with developers. Thus, research is needed to analyze the effects of these design variables.

In this paper, we present a study exploring two of these design choices. First, we examine the effect of deception/priming on computer science students in a developer study.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

To do so, we extended a developer study on password storage (primary study) using different study designs (meta-study) to evaluate the effects of the design.

In end-user studies, deception is a divisive topic. For instance, Haque et al. [32] argue that deception is necessary for password studies: "We did not want to give the participants any clue about our experimental motive because we expected the participants to spontaneously construct new passwords, exactly in the same way as they do in real life." However, Forget et al. [28] explicitly told their participants that they were studying passwords and asked participants to create them as they would in real life, in the hope of getting more realistic passwords. In an experiment to determine whether stating that the study is about passwords has an effect (i.e., priming the participants), Fahl et al. [20] found that there was no significant effect in an end-user study. Thus, there is evidence that deception is not needed for end-user studies. This is particularly relevant in terms of ethical considerations, since deception studies should only be used if absolutely necessary and the potential harm to participants must be weighed carefully.

We face similar questions when designing developer studies. For example, should we inform participants that we are studying the security of their password storage code and thus prime them, or do we need to use deception to gain insights into their "natural" behavior?

Second, we share our insights on the differences between our quantitative study and a qualitative exploration of password storage. One of the big challenges of developer studies is recruiting enough participants to conduct quantitative research. To examine this, we extended our qualitative password storage study [42] to implement a quantitative analysis and contrast the insights gained with both methods.

The rest of the paper is structured as follows. In section 2, we discuss related work. In section 3, we introduce our study methodology and explain how the study was extended. Section 4 discusses the limitations of our study and section 5 the ethical considerations. Section 6 contains the main hypotheses of our study. Section 7 presents the results and section 8 discusses the methodological contributions. Finally, section 9 summarizes the take-aways and section 10 concludes the paper.

2. RELATED WORK

This paper contributes to two distinct areas of research. The main contribution concerns the effect of priming/deception in usable security studies for developers. The related work on this topic is discussed in section 2.1. We also extend the body of knowledge on

developer studies specifically concerning password storage. Here, the related work is divided into multiple sections. Section 2.2 discusses other developer studies in general, section 2.3 focuses on developer studies concerning passwords, section 2.4 is about technical studies of password storage, and finally, section 2.5 discusses work on the usability of application program interfaces (APIs).

2.1 Priming in study design

In their research on website authentication protection mechanisms, Schechter et al. [47] explored the practice of deception in the form of a *priming effect*. They conducted a study with three groups. The participants of the first two groups were asked to role-play working with authentication data in a bank setting. One of these groups was thereby primed by receiving security-focused instructions. Fahl et al. [20] conducted a between-groups study with two variables (lab vs. online study, priming vs. non-priming). They specifically compared real-world password choices with passwords chosen by end-users in a study environment, considering priming and non-priming conditions. While the primed group was asked to behave as in real life when creating and managing passwords, the term "password" was not mentioned at all in the introductory text for the non-primed group. Neither Schechter et al. nor Fahl et al. found a significant effect for either the priming or non-priming conditions. However, both studies were conducted with end-users. Past research has shown that experts such as developers differ from end-users with regard to their mental models, behavior etc. [50, 34, 8]. Research on significant effects for developers concerning the priming and non-priming conditions does not yet exist.

In [42], we conducted a qualitative study (from which the present study acquired part of its data) with 20 computer science students, in order to investigate why developers fail with regard to password storage. Participants were asked to implement a registration task for a web application in 8 hours. We explored four scenarios: (1) priming (telling participants to consider password storage security) vs. (2) non-priming (deceiving participants by telling them the study was about API usability); and (3) web application framework with password storage support vs. (4) web application framework without password storage support. Our results indicated that frameworks offering only opt-in support for password storage and participants having strong background knowledge in software security practices were not sufficient for the production of secure software. Developers need to be told about when and how to use such security mechanisms. While the study in [42] was of a qualitative nature, we aimed at a more extensive study and invited 20 more participants for a quantitative analysis.

With consideration for the results and findings of previous studies, we offer initial insights into how developer security studies should be designed. Furthermore, we compare qualitative vs. quantitative studies and analyses, discuss time conditions for studies and examine participants' search behavior when working on a security-related task.

2.2 Developer studies

Acar et al. [9] conducted an online study with Python developers, recruited from GitHub (www.github.com). They were asked to use one of five cryptographic libraries to implement a set of security-related tasks. The main finding of this study was that simple APIs help developers to produce secure code; however, good documentation with a wide range of examples is still essential. For most of the tested libraries, security success was under 80%. Furthermore, 20% of the functional solutions were incorrectly rated by participants as being secure.

Acar et al. [10] conducted a between-subjects study to examine the impact of different documentation resources on the security of code. Fifty-four developers were given a skeleton Android app, which they had to extend in four security-related tasks. For assistance, they had access either to (1) Stack Overflow (www.stackoverflow.com), (2) books, (3) the official Android documentation, or (4) could freely choose which source to use. Programmers assigned to Stack Overflow produced less secure code. Furthermore, Fischer et al. [26] analyzed 1.3 million Android apps and found that 15.4% of them contained Stack Overflow source code. Of the analyzed source code, 97.9% contained at least one insecure code part.

Fahl et al. [23] interviewed software developers who implemented vulnerable applications regarding Secure Sockets Layer (SSL) issues. As a result, a framework was designed that prevented developers from producing insecure software in terms of SSL.

Stylos and Myers [48] investigated the relationship between secure code and the method placement of crypto APIs. They created two different versions of three APIs and asked programmers to solve three small tasks. For the same task, developers tended to use the same starting class. This resulted in faster task solution when using APIs with starting classes that referenced the class they needed.

Prechelt [44] investigated whether diverse programming languages (Java EE, Perl, PHP) or differences between the programmer teams are reflected in the security of the resulting code. For each programming language, they asked three programmer teams, comprising three professional developers each, to implement a web application in 30 h. The outcome was analyzed in terms of usability, functionality, reliability, structure, and security. They found the smallest within-platform variations for PHP.

2.3 Passwords - Developer studies

As it is often difficult to recruit professional developers for studies, Acar et al. [11] wanted to find out whether active GitHub users could be of interest for usable security studies. They conducted an online experiment with 307 GitHub users, who had to implement security-related tasks. One of these tasks considered credential storage. Neither the self-reported status as student or professional developer nor the participants' security background correlated with the functionality or security of their solutions. However, they found a significant effect for Python experience on functionality and security of program code.

Bau et al. [13] examined the vulnerability rate of web applications and programming language as well as developers in terms of career (start-up, freelancer) and their security background knowledge. For the start-up group, existing programs were analyzed. For the freelance group, eight compensated developers were invited to participate in a developer study. As compared to the start-up group, the freelancers were primed for security in the task description. With regard to secure password storage, it was found that there is a huge gap between the freelancers' knowledge and their actual implementation. Furthermore, the use of PHP and freelancers increased the software vulnerability rate.

Nadi et al. [41] studied the kinds of problems developers struggle with when using APIs. They analyzed the top 100 cryptographic questions on Stack Overflow as well as 100 randomly selected GitHub repositories that used Java crypto APIs. Within the analyzed projects, they found passwords being encrypted. This is a discouraged practice, which should be replaced by hashing. Afterwards, they conducted a study with 11 developers and a survey with 48 developers. Code templates, tools to catch common mistakes and

better documentation that includes examples were suggested for solving problems.

2.4 Passwords - Technical analysis

Bonneau and Preibusch [15] analyzed 150 websites and found they all lacked secure implementation choices. They did not use encryption, stored end-user passwords in plain text, or offered only little or no protection against brute-force attacks. This was particularly true for websites with few security incentives, such as newspapers.

Finifter and Wagner [24] analyzed nine implementations of the same web application, written in three different programming languages (Java, Perl, and PHP) in order to find correlations between the number of vulnerabilities and the programming language as well as the framework support for various aspects of security. They found no correlation between the security of web applications and the programming language. Automatic features offered by frameworks were an effective way of preventing vulnerabilities in general; however, this did not apply for secure password storage.

Egele et al. [19] analyzed more than 11 000 Android apps, with a focus on previously formulated security rules as well as password storage security. According to their findings, 88% violated at least one of those rules.

2.5 Usability of crypto APIs

While APIs are crucial for implementing secure applications that handle sensitive data, many of them seem to be too complex. As a conclusion to various examples [19, 40, 41], Green and Smith [31] presented ten principles for crypto APIs in order to reduce developer errors. Further, Gorski et al. [29] evaluated studies concerning API usability. They proposed eleven usability characteristics they consider necessary for secure APIs.

Lazar et al. [40] studied cryptographic vulnerabilities that were reported in the Common Vulnerabilities and Exposures (CVE) database. Of these, 83% were found to be a consequence of API misuse. They stated that no existing technique could prevent certain classes of mistakes.

3. METHODOLOGY

The aim of our study was to gain insight into the design of developer studies. To that end, we used two different kinds of independent variables (IVs). The first was on the meta-level, i.e., variables concerning study design. In our case, we had two meta-IVs: task design (priming and deception) and type of study (qualitative and quantitative). We refer to these as meta-variables of the meta-study. We also have an independent variable concerning the actual study subject, in our case the framework used to store passwords (JavaServer Faces [JSF] or Spring). We refer to this variable as the primary variable of the primary study.

The study presented in this paper is an extension of our previous qualitative study on password storage [42]. This quantitative study was planned at the same time to facilitate the analysis of the study type meta-variable, comparing the qualitative and quantitative approaches.

To summarize the qualitative study: participants were told that they should implement the user registration functionality for a social networking platform. Half the participants were instructed to use the Spring framework, which has built-in features for secure password storage. The other half was instructed to use JSF, a framework with manual support for password storage. This part of the design addressed the primary study. Additionally, half the participants were told the purpose of the study was to examine their password behavior and that they should store the passwords securely. The other half received a deceptive study description, which stated that the study was about the usability of APIs. For more detailed information and the exact phrasing of the tasks, see [42]. After the task was completed, a questionnaire was administered and semi-structured interviews were conducted. For the task description and the interviews, participants could choose their preferred language, either English or German. The survey, however, was in English and had to be answered in English. The study was set up for 8 h.

The main difference between the two studies was that in the qualitative study, the exit interviews were used to gain qualitative insights into the development process, while in the quantitative study, we used the survey responses and data gathered by the platform to conduct statistical testing. The hypotheses for this paper were developed before the qualitative analysis in [42] was started. This approach allowed us to gain insights into how a qualitative approach compares to a more quantitative approach.

In the combined study, we examined the following independent variables: for the primary study, the IV was the *framework* used for development (either Spring or JSF). For the meta-study, we used the IVs *priming* (deception or true purpose) and the *type of study* (qualitative or quantitative).

Participants for both studies were recruited together via a prescreening survey advertised through the computer science email list of the University of Bonn and flyers on the computer science campus. In total, 82 computer science students completed the questionnaire. Of these, 67 were invited to take part in the study. Seven of these were used for pilot studies, leaving 60 invited participants.

The first 20 participants were used for the qualitative study published in [42]. The remaining participants were used to extend the participant pool for this study. Although we had not planned to do a qualitative analysis on the remaining candidates, we conducted the exit interviews with all participants. This was done to treat all participants equally and to enable extending the qualitative analysis beyond the initially planned 20 in the event we did not reach saturation.

We removed two participants from the dataset of [42], JN1 and SP2. Due to a technical fault, the code history was not stored for JN1, and SP2 misunderstood the task so completely that no useful data was collected. This was not a big problem for the qualitative analysis but would have made the quantitative comparisons more complicated. Two random participants with a similar skill profile were selected as replacements. Of the remaining 30 invited participants, only 22 showed up. This left us with a total of 40 participants. Participation was compensated with 100 euros. Table 1 shows the demographics of all 40 participants. In the rest of the paper, we will present the quantitative analysis based on these 40 participants. The four conditions being tested are shown in section 3.1. In addition, we will contrast the qualitative findings in [42] with the quantitative findings.

3.1 Conditions

We conducted an experiment with 40 computer science students in order to explore whether task framing and different levels of framework support for password storage affect the security of software. Participants were asked to implement a registration process for a web application in a social network context, as described in [42]. The experiment was conducted under the following four conditions:

1. **Priming** - Participants were explicitly told to store the user

Gender	Male: 77.5%	Female: 15%	Prefer not to say: 7.5%
Ages	mean = 24.89	median = 25	sd = 2.89
Level of education	Bachelor: 30%	Master: 65%	Other: 5%
Study program	Computer Science: 82.5%	Media Informatics: 15%	Other: 2.5%
Country of Origin	Germany: 32.5% Iran: 5% Indonesia: 2.5% Finnland: 2.5%	India: 27.5% United States: 2.5% Turkey: 2.5% Uzbekistan: 2.5%	Syria: 5% Korea: 2.5% Pakistan: 2.5% Prefer not to say: 2.5%
Java experience	< 1 year : 42.5% 6-10 years: 5%	1-2 years: 27.5%	3-5 years: 25%

Table 1: Demographics of 40 participants.

passwords securely in the *Introductory Text* and in the *Task Description*.

2. **Non-priming** - Participants were told the study is about API usability, but were not explicitly asked for *secure* password storage.

3. **Framework with opt-in support for password storage** - Participants were advised to use a framework offering a secure implementation option, which could be used if they thought of it or found it. Spring was chosen as a representative framework [42].

4. **Framework with manual support for password storage** - Participants were advised to use a framework with the weakest level of support for password storage. Thus, they had to write their own salting and hashing code using just crypto primitives. JSF was considered as a suitable web framework in this case [42].

Java was selected as the programming language because it is one of the most popular and widely used programming languages for applications and web development [1, 7, 5, 3, 4, 6]; in addition, it is regularly taught at our university. Therefore, we reasoned that we would be able to recruit a sufficient sample of computer science students for our study.

Since a related study [11] has shown that self-reported skills of developers affect the study results, we used randomized condition assignments and counterbalanced for participants' skills reported in the pre-questionnaire (this is known as Randomized Block Design [37]). The pre-questionnaire can be found in Appendix A.

3.2 Deception

We examined the effect that concealing the true purpose of the study had as opposed to openly making it about secure password storage. Kimmel indicated three stages in which deception can be integrated: *subject recruitment, research procedure and post-research/application* [36, p.65]. In our study, we investigated whether participants made sure to store end-user passwords securely, if they were not explicitly told to do so in either the *Introductory Text* or the *Task Description* (non-primed group). In the recruiting phase, all candidates (primed and non-primed) were told the purpose of the study is API usability research (*"The goal of the study is to test the usability of different Java web development APIs."*). Consequently, we used deception in the *subject recruitment* and *research procedure* stages.

3.3 Experimental environment

The experiment was performed in an in-person laboratory, which allowed us to control the study environment and the participants' behavior. We created an instrumented Ubuntu distribution designed for developer studies that included code-specific tracking features. Thus, we were able to collect all data produced by the participants within the 8 h sessions (e.g., the web history and program code history). Every code snippet that was compiled was secured in a history folder. In addition to a video recording of the participants' desktops, the setup also allowed us to take frequent snapshots of their progress. In order to capture copy/paste events, we used *Glipper* [2], a clipboard manager for GNOME, which we modified slightly to meet our requirements (e.g., adding a time stamp to the events in a log file). In this manner, the study environment allowed us to identify all participants who copied and pasted code for password storage and the websites from which they received the code.

3.4 Survey

Before working on the task, participants filled out a short entry survey regarding their expectations for task difficulty. They also completed a self-assessment of their programming skills (see Appendix B). After finishing the implementation task, participants were required to complete an exit survey (see Appendix C). We asked participants for their demographics, security background knowledge, programming experience, and experience with the task and APIs. Furthermore, we asked open questions that could be answered with free text. Two coders independently coded the participants' answers by using Grounded-Theory and compared their final code books using the inter-coder agreement. The Cohen's kappa coefficient (κ) [18] for all themes was 0.78. A value above 0.75 is considered a good level of coding agreement [27].

To analyze the usability of the APIs, we applied the 11-question scale suggested by Acar et al. [9] (Appendix C.1), since it is more developer-oriented than the standard System Usability Scale (SUS) [16], which is more end-user oriented. Acar et al.'s usability scale is a combination of the cognitive dimensions framework [17], usability suggestions from Nielsen [43], and developer-related recommendations from Green and Smith [31].

3.5 Scoring code security

We used the same scoring system as was used in [42]. For each solution, we examined its **functionality and security**. We rated participants' solutions as functional if "an end user was able to register the Web application, meaning that his/her data provided through the interface was stored to a database" [42].

We used two measures to record the security of a participant's

solution. Every solution was rated on a scale from 0 to 7, according to the security score introduced in [42] (see Appendix D). This value is referred to as the *security score*. In addition, we used a binary variable called *secure* which was given if participants used at least a hash function in their *final* solutions and thus did not store the passwords in plain text.

We were also interested in participants who attempted to store user passwords securely, but struggled and then deleted their attempts from their solutions (this was coded as *attempted but failed*, or *ABF*). For this, we collected and analyzed participants' code history. In order to identify security attempts, we used the Unix *grep* utility. With grep, we searched for security-relevant terms based on the frameworks and best practices (see Appendix F). When a term was found, we analyzed the code snippets manually.

It is important to note that we still gave security scores to participants who implemented secure password storage but failed to create a functional solution, i.e., the user registration did not work. The rationale for this was that we were interested in how participants stored passwords. All other parts of the task were distraction tasks and thus of less relevance.

4. LIMITATIONS

Our study has several limitations that need to be considered when interpreting the results.

The most noteworthy limitation is that we used a convenience sample comprising 40 computer science students from a single university. Despite having a pool of 1600 computer science students at our university and offering fairly high compensation, we did not get more volunteers. We will discuss this limitation in the context of both the primary study and the meta-study. For the meta-study, we wanted a homogeneous sample so we could attribute any changes in outcome to the difference in task design. The limitation of this decision is that our results are not currently transferable to other participant groups. While we believe it is likely that other student samples will produce similar results, we expect bigger differences when working professionals are considered. It is also likely that there will be big differences between different groups of working professionals. These differences will need to be explored in future work.

The primary study is limited in the same way. Here, it would have been more desirable to have a more diverse sample; however, this would have conflicted with the need for a homogeneous sample for the meta-study. Since the meta-study was our main goal, we accepted the limitation of the primary study. That being said, there are early indications that computer science students can be acceptable proxies for professionals in developer studies [51, 38, 10, 11, 33, 14, 49, 39, 46]. This sample was not selected for its representatives and thus should not be used to infer anything about non-students.

Since our study was performed in a laboratory environment with laboratory PCs, we have an unknown amount of bias in our results. This is particularly relevant to the meta-study. While the low amount of attempted security in the non-priming condition seems plausible in light of the many password database compromises, we have no way of confirming that we are measuring the same effect. It is possible that the low amount of attempted security in the non-primed group is not due to participants' lack of awareness that passwords should be hashed and salted, but rather to a lack of concern for passwords in a study environment. In fact, we received statements to this effect in the exit survey. Of the 20 non-primed participants,

- two attempted to implement a secure solution but failed,
- two thought it was secure despite not having done anything to secure it themselves,
- two stated that they did not implement security because it was not part of the task,
- three stated that the functionality was more important to them than security,
- three were aware that security was needed but did not give any reason why they did not implement it, and
- eight were not aware that hashing and salting were important for password storage.

We must point out that the above statements are based on self-reporting by the participants. False reporting is possible in both directions. Participants who might not have been aware of the need for security might have felt embarrassed and stated that they did know but chose not to implement it and made up a reason for it. It is also possible that a participant who did know stated otherwise so as not to have to explain why security was not implemented. We must acknowledge these limitations.

We only conducted Bonferroni-Holm correction for our main hypotheses. For the rest of the exploratory analysis, we accepted the higher probability of type 1 errors to lower the risk of type 2 errors. Thus, new findings need to be confirmed by replication before they are used.

5. ETHICS

At the time of the study, our institution did not have an IRB for computer science studies. The study design was instead discussed and cleared with our independent project ethics officer. Our study also complied with the local privacy regulations. Participants gave written informed consent before participating in the study. Since half our participants underwent a deception condition, the study ended with an in-person debriefing, where all participants were informed of the true purpose of the study. Most participants were not bothered by the deception condition at all. However, some participants felt that they were judged unfairly and they stated that they would have included security if we had asked for it. After re-stating that this was completely fine, that we were interested in the APIs' ability to nudge developers toward security, and that they were not at fault, there did not seem to be any lingering negative feelings. There were also positive reactions to the deception. The majority of participants remarked that they learned a lot through the deception and will be more aware of security in future tasks and jobs, even if they are not explicitly asked to think of security.

6. HYPOTHESES & TESTS

We examined seven main hypotheses in our experiment. Two concerned the meta-focus of this paper, namely, the effect of priming/deception, denoted by P(riming). Two further concerned the A/B test comparing the two frameworks, denoted by F(ramework), and the final three were general tests concerning password storage security, denoted by G(eneral).

H-P1 Priming has an effect on the likelihood of participants attempting security.

H-P2 Priming does not have an effect on achieving a secure solution once the attempt is made.

H-F1 Framework has an effect on the security score of participants attempting security.

H-F2 Framework has an effect on the likelihood of achieving functional solutions.

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.

H-G3 Copying/pasting has an effect on the security score.

6.1 Meta-study

It is natural to assume that requesting a secure solution will lead to more attempts at security (H-P1). The interesting aspect here was how many of the non-primed participants attempted a secure solution. While we expected priming to increase the number of attempts, we did not expect a different failure rate between the priming and non-priming group (H-P2), i.e., if non-primed participants attempted security, they should not have failed more often than primed members.

6.2 Primary study

While only indirectly linked to security, we also considered the possibility that the differences between the two frameworks (JSF and Spring) could lead to different rates of functionality (H-F2). We also expected that the greater level of support offered by Spring would increase the security score of Spring participants (H-F1).

6.3 General

The above hypotheses are novel to this work. However, we also wanted to confirm findings from related studies. In their study, Acar et al. [11] observed that the programming language experience had an effect on the security of participants' solutions. Therefore, we also assumed we would observe an effect regarding experience with the Java programming language and the security score of participants' solutions (H-G1). In addition, we assumed that if participants had experience with storing user passwords in a database backend, they would be more likely to create a secure solution in the study (H-G2). Finally, several studies noted the effects of copy/paste on study results [9], especially in terms of security [10, 21, 22, 23]. Thus, we assumed that copy/paste events would affect the security code of our participants as well (H-G3).

6.4 Statistical testing

We chose the common significance level of $\alpha = 0.05$. When conducting tests on all 40 participants, we labeled the group as "all". When only testing subgroups, these were also labeled for easy interpretation. We used the Fisher's Exact Test (FET) for categorical data. For numeric data, we considered linear regression and logistic regression for binary data if the data was normally distributed. In order to test for normality, we used the Kolmogorov-Smirnov test and plotted the data for manual inspection. For data that was not normally distributed, we used the following non-parametric tests: in order to find differences between all four conditions, we used the Kruskal-Wallis test; for two groups, we used the Mann-Whitney U test. In both cases, the Levene's median-based homogeneity of variance test showed the distributions among the groups to be similar. Statistically significant values are indicated with an asterisk (*).

Of the seven main hypotheses, three concerned the security score, two concerned the binary secure value, one concerned the attempted security, and one concerned functionality. Since all but the functionality tests were closely related, we applied family-wise error correction using the Bonferroni-Holm method with a family-wise correction of 6. These analysis sections are marked with the relevant hypothesis label. However, we did not apply multiple testing correction in the exploratory part of our analysis (sections not marked with a hypothesis label). Since virtually no research has been conducted on study design for developer studies, we thought it was more important to discover interesting effects for future research to explore than it was to avoid type 1 errors while potentially dismissing an important effect merely because our sample size was not big enough (i.e., type 2 errors). For more information on family-wise errors, see [12]. To ease identification, we labelled Bonferroni-Holm corrected tests with "family = N", where N was the family size, and reported both the initial and corrected p-values.

7. RESULTS

While our main goal was analyzing the meta-study results, we began by analyzing the functionality and security of the primary study since these results were needed for the meta-analysis. Sections analyzing one of the seven main hypotheses are marked with the hypothesis label.

7.1 Functionality

Here, we discuss the functionality of the code the participants produced. We considered a solution as functional if an end-user account could be created. Figure 1 shows the distribution of functional solutions across our conditions. Of all 40 participants, 26 (65%) produced a functional solution. As shown in Figure 1, the number of participants who were able to solve the functional task is a bit higher in the Spring group compared to the JSF group.

7.1.1 Framework effects functional solution (H-F2)

Eleven of 20 (55%) participants using JSF and 15 of 20 (75%) using Spring managed to solve the task functionally. These differences were not statistically significant (sub-sample = all, FET: $p = 0.32$, odds ratio = 2.40, CI = [0.54, 11.93]). Thus, we do not reject H-F2. However, we only had a power of 0.17, so this effect is worth looking at in follow-up studies. Interestingly, a significant result would mean that the more complex framework actually has better usability with respect to functional solutions.

7.1.2 Part-time job in computer science

We asked our participants whether they had a part-time job in computer science. In prior research, Acar et al. counted students who had part-time jobs as professionals [11]. We, however, found no significant effect between having a part-time job in computer science and a functional solution (sub-sample = all, FET, $p = 1.0$, odds ratio = 0.84, CI = [0.19, 3.89]).

7.2 Security

Figure 2 shows the distribution of secure solutions across our conditions. Twelve (30%) of our 40 participants implemented some level of security for their password storage. Of the 20 participants in the non-primed groups, 0% stored the passwords securely. While we had expected significantly fewer secure solutions in the non-primed groups, we were surprised by this extreme result. From the primed group using JSF, 5 of 10 (50%) implemented some level of security (mean security score = 2.15, median = 1, sd = 2.67). From the primed group with the Spring framework, 7 of 10 (70%) participants implemented some level of security (mean security score = 4.2, median = 6.0, sd = 2.9). Table 4 shows an overview of the security scores achieved by our participants (Appendix E).

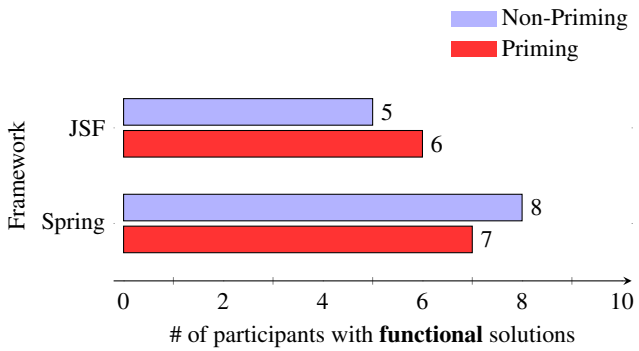


Figure 1: Functionality results per framework, split by primed vs. non-primed groups.

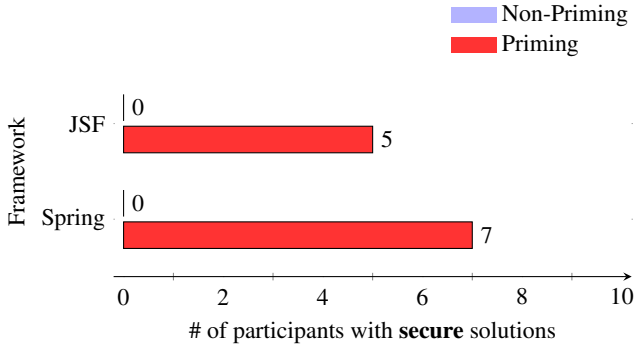


Figure 2: Security results per framework, split by primed vs. non-primed groups.

7.2.1 More Java experience, more security (H-G1)

In prior research, Acar et al. found that more Python experience leads to more security [11]. We wanted to examine this effect within our sample. We found no significant differences in our study (sub-sample = all, Kruskal-Wallis: $\chi^2 = 4.118$, $p = 0.249$, *cor-p* = 0.498, family = 6). Between the different groups of Java experience, the security score showed no significant effect. However, it must be mentioned that Acar et al. studied student and professional volunteers recruited on GitHub without compensation. Their participants had a wide range of years of experience in Python compared to our students in Java. So we have a number of differences in the samples. It is important to note this difference since it makes sense to take skills into account during condition assignment in randomized control trials. In short, we failed to confirm H-G1.

7.2.2 Previous password experience (H-G2)

We hypothesized that participants who had previous experience storing user passwords in a database backend would be more likely to add security in the study. Therefore, we wanted to test whether participants who reported having stored passwords before performed differently regarding security compared to participants who had never stored passwords before. Nine non-primed and 15 primed participants reported having stored passwords prior to the study. We found no significant differences in security in comparing the different groups of participants (sub-sample = all, FET: $p = 0.297$, *cor-p* = 0.498, odds ratio = 2.54, C.I = [0.49, 17.72], family = 6). We thus fail to reject the null hypothesis of H-G2 and cannot draw conclusions on this hypothesis. Furthermore, we calculated a power of 0.19, indicating that the effect is not reliable.

7.2.3 Framework effects security score (H-F1)

In this section, we only consider those participants who attempted security. We wanted to examine whether the framework used affected the security score (including ABF scores). We expected that Spring might score better because, in contrast to JSF, it offers built-in functions for storing passwords securely by using hashing, salting and iterations.

The descriptive statistics for the JSF group are Min 2, Median 5.5, Mean 4.3, and Max 6. The descriptive statistics for the Spring group are Min 6, Median 6, Mean 6, and Max 6. Due to the Bonferroni-Holm correction, the difference between the two groups is not flagged as significant (sub-sample = all $\wedge$ attempted security = 1 , Mann-Whitney U = 15, $p = 0.051$, *cor-p* = 0.20, family = 6). It does seem likely, though, that a larger sample would confirm the trend that Spring participants earned higher scores than JSF participants. This will be put into further context in section 7.4.1.

7.2.4 Usability of frameworks

We used the usability score from Acar et al. [9] (see Appendix C.1) to evaluate how participants perceived the usability of the two frameworks. We compared the values of the usability score for all four groups: non-primed JSF (mean = 48.25, median = 51.25, sd = 13.54), primed JSF (mean = 50.50, median = 53.75, sd = 9.78), non-primed Spring (mean = 50.50, median = 55.00, sd = 20.10), primed Spring (mean = 58.75, median = 57.50, sd = 15.65). We found no significant effect comparing all four groups (sub-sample = all, Kruskal-Wallis: $\chi^2 = 3.169$, $p = 0.37$). Furthermore, we examined whether the frameworks had different usability scores when the participants attempted to solve the task securely. We did not find a significant effect in this case either (sub-sample = all $\wedge$ attempted security = 1, Mann-Whitney U = 21.5, $p = 0.29$).

7.2.5 Security awareness

Fourteen primed and two non-primed participants believed that they managed to store user passwords securely. The two non-primed participants erroneously believed they had stored passwords securely (JN5, JN7). Their given survey answers suggested that neither had any background knowledge of password storage security at all. The primed participants were additionally asked whether they would have been aware of security if we had not explicitly ask them for it. Nine of the 14 participants indicated they would have stored user passwords securely, even if they had not been explicitly asked to do so. The fact that only two out of 20 non-primed participants attempted security suggests this is overly optimistic.

7.2.6 Security classes

In prior work, Acar et al. found that security courses had a significant effect on security [11]; therefore, we asked our participants which courses they had attended at our university in the past. We gave one point per security-relevant course. Since not all Masters students had completed their undergraduate studies at the same university, we also asked for other courses. None of the participants added a security-relevant class in the open question space. Participants reported they had attended between 0 and 4 security classes (mean = 0.8, median = 1, sd = 0.99). We found no significant evidence in the overall group (sub-sample = all, FET: $p = 0.737$, odds ratio = 1.39, CI = [0.29, 7.022]).

7.2.7 Part-time job in computer science

We found no effect between having a part-time job in computer science and a secure solution (sub-sample = all, FET, $p = 1.0$, odds ratio = 1.10, CI = [0.22, 5.30]).

7.2.8 Web browser history & task completion time

In order to analyze the web browser history, we aggregated all our participants' browser history. We assessed the visit count of all participants (mean = 179.0, median = 174.5, sd = 97.02). We could not analyze the browser history of one of our participants, because he had deleted it after completing the task. We found a total of 6224 distinct web pages for all participants. We also measured the time our participants needed to solve the task [hours] (mean = 5.11, median = 5.35, sd = 1.72). On average, participants visited 36.2 pages per hour (mean = 36.2, median = 31.52, sd = 16.44). We tested whether there was a difference in security that depended on the number of websites participants used. The results show that the website count was not significantly relevant (logistic regression, odds ratio = 1.0 , C.I = [0.99, 1.00], $p = 0.423$).

7.3 Priming

7.3.1 Priming leads to more attempts to store user passwords securely (H-P1)

The main goal of our study was to measure the effect of priming. Only two of 20 non-primed participants attempted to store the passwords securely, compared to 14 of 20 in the primed groups. This difference is statistically significant (sub-sample = all, FET: $p = 0.000$*, *cor-p* = 0.001*, odds ratio = 19.02, C.I = [3.10, 219.79], family = 6). Thus, we can reject the null of H-P1 and conclude that priming has a significant effect. We already stated we were surprised that no non-primed participant achieved a secure solution. This is mirrored in the very low number of participants who attempted to create a solution. However, we were also surprised that six participants in the primed group did not attempt a secure solution, since it was explicitly asked of them. Of these, though, three also did not manage to create a functional solution. In the exit survey, all six participants stated that they had not achieved an optimal solution and cited technical difficulties that prevented them from attempting to create a secure solution. For instance, SP6 noted: "[I] encountered errors in connecting with the DB through Spring JPA and was not able to come up with the solution. As a result [I] could not focus on implementing an algorithm to securely store the password."

It is interesting to note that even when security was explicitly stated as the goal of the study, these participants still wanted to create the functional solution before adding the security code.

7.3.2 Priming effect on achieving a secure solution once the attempt is made (H-P2)

We had hypothesized that the priming effect would only influence whether a participant would think of adding security, but once a participant had made the decision to add security, the will to follow through would be independent of priming. Now, it is very difficult to make a convincing case of no-effect using frequentist statistics with a small sample size; however, this may not be a concern. It turned out that there might actually be an effect. In the non-primed group, two of 20 attempted security but did not follow through to achieve a secure solution. In the priming group, 14 of 20 attempted and 12 achieved a secure solution. The difference between the groups is significant before correcting for multiple testing (sub-sample = all $\wedge$ attempted security = 1, FET: $p = 0.05$, *cor-p* = 0.20, odds ratio = Inf, CI = [0.64, Inf], family = 6). The same goes for the security scores (sub-sample = all $\wedge$ attempted security = 1, Mann-Whitney U: 2.0, $p = 0.034$*). Although this effect was not significant after correction, we think this is an important observation which should be examined in future studies. While it is possible that the small number of attempts in the non-primed group skewed our results, it is also possible that the failure to mention security in the task not only meant participants were not explicitly informed that security is important for password storage, but potentially discouraged participants who knew this from implementing it. This could have implications outside of study design since this effect is likely to occur in everyday life as well where developers might not be explicitly asked to secure their code and thus be dissuaded from doing so even if they know they should.

While we fail to reject the null of H-P2 due to the Bonferroni-Holm correction, we find the data to be highly interesting and suggest examining this effect in future studies.

7.4 Copy/Paste

7.4.1 Security and copy/paste (H-G3)

Our analysis of the copy/paste behavior of our participants showed another interesting result.

Of the 40 participants, only 17 copied and pasted code. Of these, 12 created a secure solution. The surprising aspect is that all secure solutions come from participants who copied and pasted security code. Not a single "non-copy/paste" participant achieved security. This difference was statistically significant (sub-sample = all, Mann-Whitney U = 57.5 $p = 0.000$*, *cor-p* = 0.000*, family = 6). Thus, we reject the null of H-G3. However, it is noteworthy that we see a positive effect of copy/paste. This is in contrast to previous work by Acar et al. [10] and Fischer et al [26]. For example, Acar et al. stated in their discussion: "Because Stack Overflow contains many insecure answers, Android developers who rely on this resource are likely to create less secure code" [10]. And Fisher et al. stated in their conclusion: "We show that 196,403 (15%) of the 1.3 million Android applications contain vulnerable code snippets that were very likely copied from Stack Overflow" [26].

These negative views are in stark contrast to our findings that 0% of participants who did not use copy/paste created a secure solution. We do not dispute the findings of Acar et al. and Fisher et al., but we do show that there is also a significant positive effect of copy/paste.

This finding also changes how we must interpret the difference in security scores between the two framework conditions presented in section 7.2.3. All secure Spring participants scored 6 points, while the JSF scores varied between 2 and 6. This could indicate that the Spring API has better usability, because it has safer defaults. However, this usability advantage seems to only affect our participants indirectly, via the web sources they use. This suggests that it is worth considering testing the usability of APIs not only with software developers but also with those who create web content. In the following section, we take a closer look at the websites used by our participants.

7.4.2 Websites used for copy/paste

Almost half of the participants (42.5%; 17/40) copied password storage examples from various websites on the Internet and pasted it to their program code. Of these, 82% (14/17) were primed participants. In all other cases, participants copied code from websites covering storage of user data in general (e.g., name, gender, email), adapting it for passwords. These websites were not considered for further analysis, since we were only interested in password storage examples.

Table 5 (Appendix F) shows all websites from which participants copied and pasted code for password storage into their solutions. The table also considers participants who attempted to store user passwords securely but did not include the security code in their final

solutions (ABF). We manually analyzed all proposed examples for password storage on these websites by using the same security scale as applied to the evaluation of participants' code (see Appendix D). If websites introduced generic solutions without predefined parameters for secure password storage, but discussed how these should be chosen in order to achieve security (e.g., OWASP: General Hashing Example (Appendix F)), we still awarded points for these parameters according to the security scale. Additionally, we compared the security scores participants received for their solutions with the scores of password storage examples from the websites they used. Since websites often contain more than one code snippet, we manually scored all of them and then used the following classification of snippets:

- **Most insecure example** - The worst solution we found on the page.
- **Obvious example** - The most obvious solution in our subjective assessment, e.g., answers on Stack Overflow that are rated with a high score by the community. For all other websites, we classified examples as obvious if they were posted at the beginning of the website.
- **Most secure example** - The solution with the highest security score.

We found that all participants who implemented password storage security (100%, 12/12) copied their program code from websites on the Internet. The majority, 75% (9/12) of participants, achieved the almost maximum score of 6/7 points in our study. These participants copied and pasted code from websites introducing up-to-date, strong algorithms. One thing the websites had in common was that all solutions had good security scores. Only one participant was on a website where the least secure example was "only" a 5.5 score. However, the most obvious example was scored with a 6 and taken by the participant.

The other three participants came across blog posts and tutorials with outdated or unsecure implementation (JP2, JP3, and JP10). For instance, JP2 copied code from a tutorial that was published in 2013 (see Appendix F, Blog Post: Hashing Example). Thus, he adopted an iteration count of 1 000 for PBKDF2, although 10 000 iterations are recommended by NIST today [30]. Interestingly, this tutorial also discussed the usage of `MD5`, `bcrypt`, and even `scrypt` with associated program code examples. The example for `MD5` was listed at the top of the website; we therefore classified it as the *obvious* example. But the author did state that this solution is vulnerable to diverse attacks and should be used with a salt. The blog post also discussed a program example for `scrypt`, which we classified as *most secure*. This was the only website visited by our participants where an example scored 7/7 points. However, JP2 decided to use `PBKDF2`, for which he found a general hashing example at the Open Web Application Security Project (OWASP) website (see Appendix F, OWASP: General Hashing Example). Although properties of parameters are discussed on the OWASP website in general, they are not applied in the code example. Therefore, JP2 searched for a similar implementation with predefined parameters and ended up with an outdated iteration count.

JP3 copied code that only contained a weak SHA1-based example. More interestingly, JP10 merged program code from four websites. Although one website included code with three points for an *obvious* example and five points for a *most-secure* example, he received only two points for his final solution. He did not use a salt, despite the fact that he copied code from an *obvious* example on Stack Overflow that considered a function with a salt as an input parameter. However, the example did not include a predefined implementation of the salt and was not implemented by our participant.

An interesting priming effect can be seen between the two participants, JP7 and SN8, who both copied code from websites in which user credentials were stored in plain text. The primed participant, JP7, used the unsecure blog post for gaining a functional solution and afterward installed a Java implementation of OpenBSD's Blowfish password hashing scheme, jBCrypt, and received six points. In contrast, the non-primed participant, SN8, did not take any further action to implement security.

Only two of the 20 non-primed participants considered security while programming, though they did not provide secure solutions in the end (JN9, SN4). JN9 was able to implement a functional solution storing user passwords securely. However, he accidentally deleted parts of his code, resulting in errors he was unable to correct. At the end, he provided a functional solution without including secure password storage. In terms of copy/paste, JN9 is interesting since the solution he implemented had, at one point, a security score of 3, although the website he used for copy/paste was scored with 2 points. He was the only participant who used a salt that he did not copy and paste from a website, but rather included it by himself. However, he used the user's email address as the salt, which is not considered a security best practice.

In summary, no participant who copied/pasted code used the *most unsecure example* on websites. Whenever the *obvious* security score differed from the *most secure* examples (true for 3/21 websites), participants used the latter. If participants' code was merged from more than one website (JP2, JP7, and JP10), participants' security score was always higher compared to the lowest-scored website, considering *most secure examples*.

7.5 Statistical testing summary

Table 2 gives an overview of the seven main hypotheses and the results of our statistical tests, with both the original and Bonferroni-Holm corrected p-values. We have two very clear results. First, concerning the meta-study: priming has a huge effect. Second, concerning the primary study: copy/paste has a strong positive effect on code security.

The effects of H-P2 and H-F1 were not statistically significant after correcting for multiple testing, but seem promising enough to examine in future work. It is also noteworthy that we did not find a significant effect for H-G1, which has been found in other studies. This is likely due to the fact that with only a student sample, the range of experience was so small that the effect is not large enough. This is important to know since it simplifies study design for developer studies conducted with students.

7.6 Examining survey open questions

We analyzed open questions of the exit survey for trends rather than for statistical significance, to gather deeper insights into the rationale behind participants' behavior.

Before mentioning security at all, we asked our participants whether they solved the task in an optimal way (see Appendix C, Q2). Thus, we were able to observe whether non-primed participants based their answers on functionality rather than on security. Seven out of 40 participants believed their solution was optimal (JP3, JN10, SN1, SN2, SN5, SN7, and SP1). In fact, most of the participants were non-primed and solved the task functionally but not securely. Some

H	Sub-sample	IV	DV	Test	O.R.	C.I.	p-value	$cor-p$-value
H-P1	-	Priming	Attempted security	FET	19.02	[3.10, 219.79]	0.000*	0.001*
H-P2	Attempted security = 1	Priming	Secure	FET	Inf	[0.64, Inf]	0.05*	0.20
H-F1	Attempted security = 1	Framework	Security score (incl. ABF)	Mann-Whitney	-	-	0.051*	0.20
H-F2	-	Framework	Functional	FET	2.40	[0.54, 11.93]	0.32	-
H-G1	-	Java experience	Security score	Kruskal-Wallis	-	-	0.249	0.498
H-G2	-	Stored passwords before	Secure	FET	2.54	[0.49, 17.72]	0.297	0.498
H-G3	-	Copy/Paste	Security score	Mann-Whitney	-	-	0.000*	0.000*

IV: Independent variable, DV: Dependent variable, O.R.: Odds ratio, C.I.: Confidence interval
Corrected with Bonferroni-Holm correction, except for H-F2.
Significant tests are marked with *.

Table 2: Summary of main hypotheses.

even stated that all requirements were functionally solved and thus their solution was optimal (JN10, SN7, SN1, SN2, and SN5). SN1, for instance, noted: "My [manually performed] tests [...] worked as expected, I should have covered everything." His answer shows that he invested some time in testing his implementation. Still, since SN1 did not think about storing the user credentials securely, it might be interesting to involve security in the testing process as well. The primed participant SP1, though, argued that his solution was optimal because the security part was sufficiently solved: "It uses `bcrypt` [with the] highest vote on [Stack Overflow link]." In contrast, a number of participants said that the quality of their code was not optimal because it did not rely on best practices, e.g., SP11: "I have probably not used best practices for Spring/Hibernate as it is the first time I used them." Other participants mentioned that exceptions and warnings need to be caught and the code can be written more cleanly and clearly (SP4, SP9, SP10, SN4, and SN9).

If participants believed they stored the user password securely, they were asked whether they solved the task in an optimal way with regard to security (see Appendix C, Q9). Only 7 of 40 participants believed that their security code was optimal (JN5, JN7, JP4, JP7, JP10, SP7, and SP11). SP7, for instance, noted that he used an "industry standard way of storing passwords" and assumed that his solution was therefore optimal. While JP3 and SP1 indicated they solved the task in an optimal way at first, they changed their minds when the question was asked in terms of security. While JP3 noted "everything is implemented", thus indicating his solution was optimal, he changed his mind with regard to security, "because the [iteration count] is not implemented yet." SP1 listed three reasons explaining why his solution is not optimal in terms of security: (1) "User is not enforced to use symbol, combination of numbers, etc.," (2) "Storing the password securely does not mean that one [person] cannot hack into another's account," and (3) "Lacking [...] 2 step validation (by phone, for example)." First, SP1 assumed that security should be implemented involving the end-user. This assumption was also made by other participants, who noticed that password validation for the end-user was missing in their solutions (SN1, SN2, SN4, SP5, JP7, SP9, and SP11). Second, SP1 did not trust password security at all, although he suggested a method for improvement (two-factor authentication). Interestingly, the non-primed participants, JN5 and JN7, indicated they stored the user password securely in an optimal way. However, we did not find any evidence of security at all, in either their solutions or in their attempts. Their answers suggested a general lack of knowledge of password storage security.

8. METHODOLOGICAL CONTRIBUTIONS

8.1 Deception

While Fahl et al. [20] found no significant difference in password studies in the behavior of end-users who were primed that the study was about passwords or received deceptive treatment, we see a very strong effect on the behavior of developers. Both design choices offer interesting insights into the problem of storing passwords securely.

If researchers wish to study the usability of a security API, priming participants is clearly the best choice, since the majority of participants in the non-primed group had no contact with the API at all and thus do not produce any data to analyze. The majority of developer user studies fall into this category.

However, these studies only look at one aspect of a much larger problem. In [21] Fahl et al. analyzed the misuse of transport layer security (TLS) APIs in Android. They found that 17% of applications using HTTPS contained dangerous code. However, 53.8% of apps did not use the TLS API at all, exposing a wealth of data to the Internet without any protection. We think it is important to study this aspect as well, and help developers become aware they need to think about security. Our results suggest that deception in studies is a promising way of studying this. It can be argued that the students simply did not include secure storage because they were in a study environment. Some participants even stated this in the exit survey and interviews. However, since there are many cases in the real world in which security is not explicitly stipulated, we think that the non-priming condition can be a valuable design for studies. This is definitively an area in which more research is needed before a reliable statement can be made.

For now, we do suggest that the usable security community also conducts developer studies using deception instead of focusing only on API use on its own. It is, however, important to conduct a full debriefing at the end to ensure the well-being of participants. In our case, we did not see any issues with the debriefing that were not addressed to the satisfaction of the participants.

8.2 Task length

The most difficult aspect of designing a deception study for developers is that distraction tasks are necessary to avoid tipping off the participants.

Short tasks Most related studies are very short [9, 10, 11, 48]. As noticed by Acar et al. [11], tasks for uncompensated developers should be designed in a way that *"participants would be likely to complete them before losing interest, but still complex enough to be interesting and allow for some mistakes."* Acar et al. [11] conducted an online experiment with 307 uncompensated GitHub users, who were asked to complete three different tasks: (1) URL shortener, (2) credential storage, and (3) string encryption. Each participant was assigned the tasks in random order. For the user credential storage task, only one function was given, which had to be completed by developers. The task was formulated in a straight forward way and

it was clear where to insert the needed code and why. Additionally, clear instructions were given to the participants, answering the question when the problem was solved. The participants were not explicitly asked to consider security. In their study, only a small number, 17.4%, stored the user passwords in plain text. A direct comparison cannot be made since the GitHub users were more experienced than the students in our study; however, the short task time and the direct instruction to store the passwords is likely to have an effect as well.

One-day time frame In contrast to tasks completed over a short time frame, longer studies are more realistic since developers have long-lasting projects and tasks they work on in the real world. In particular, it is possible to create competing requirements, pitting functionality against security in a way that is not possible in short, focused tasks. In [42], we discussed the design process of the task used in this paper in detail and how the 8 h time frame was calibrated with several pilot studies. The rational was that 8 hours is the longest time we could reasonably ask participants to remain in a lab setting. In addition, there are a number of benefits to having the participants in a controlled environment. In particular, we could fully configure the lab computers to gather a wealth of information, including full-screen capture, history of all code, copy/paste events, search history, and websites visited. Remote studies could easily use web-based editors to capture code and copy/paste events; however, gathering the rest of the information would be much more intrusive.

Multi-day time frame In a one-day time frame, we were able to conduct a task that was sufficiently long and complex that participants could perceive security as a secondary task. A multi-day time frame also offers this benefit. For instance, Bau et al. [13] investigated web application vulnerability in a multi-day experiment with eight freelancers. They were asked to develop an identity site for youth sports photo-sharing with login and different permission levels for coaches, parents, and administrators. The freelancers were primed for security by mentioning that the website "was mandated by 'legal regulations' to be 'secure', due to hosting photos of minors as well as storing sensitive contact information" [13]. The developers promised a delivery period of 35 days. Participants were compensated from three different price ranges (< $1 000, $1 000 - $2 500, and > $2 500). Two of the eight freelancers stored passwords in plain text, showing a similar distribution as in our priming condition. This design offers higher ecological validity; however, far less detailed information about the code creation process can be gathered. Both our study and the studies conducted by Acar et al. [10] have shown that information sources play a vital role in code security, which is much trickier to gather in this kind of study. So there is a trade-off between ecological validity and the ability to gather high-fidelity data.

In short, we see benefits in all three time frames and researchers now have initial data to help choose which is most appropriate for their setting.

8.3 Laboratory setting

Many developer studies are conducted online due to the difficulty of recruiting enough participants to come to a lab study. However, we found the information gathered by our instrument OS very valuable. Most developer studies contain both coding tasks and questionnaires. The questionnaires are used both for pre-screening and for gathering information on the task. While it is possible to detect the use of web sources indirectly through paste events, it is also critical to be able to detect the use of online sources during the administration of surveys.

We manually analyzed all the screen capture videos of our participants while they were answering the surveys. We could only analyze the videos of 38 participants due to technical difficulties, which meant that we were missing two videos (JN8 and SN5).[1]

We found that half the participants (20/38) used Google when answering the survey, either searching for framework-related topics (6/38) or for password storage-related topics (14/38; see Table 3). Interestingly, half the non-primed participants who did not attempt to store user passwords securely (4/8) started to search how this could be done while answering the survey. SN1, for instance, copied a survey answer from Wikipedia, explaining what hashing functions are defending against.

Of the primed participants with secure solutions, 58% (7/12) searched for additional password storage security details, e.g., in order to explain why the used algorithms were optimal or not.

Since our laboratory setting captured this information, we could take it into account during data analysis. In most online settings, this information is not available and thus there can be no certainty that the answers reflect the knowledge of the participant or just their ability to use Google.

This is particularly critical in the use of pre-screening surveys, as is done in most studies (including this one). It is common to try to screen out unsuitable candidates who do not have the technical skills needed to take part. Luckily, we only used self-assessment and reported experience to conduct the counter-balancing. However, there are also expert studies which used content-based questions for participant selection, such as the study by Krombholz et al. [38]. Here, the researchers had to be aware that a potentially large number of the participants used Google to answer the questions, which might not properly reflect their actual skills.

Being able to see all searches and information sources in direct relation to questions and answers was very valuable and is an important strength of lab-based studies. We will be releasing the study OS as an open source project, so other studies can easily capture the same information.

8.4 Qualitative vs. Quantitative study design

Finally, we want to share some observations contrasting the qualitative approach from [42] with our quantitative extension. Here, we need to distinguish between the primary study and the meta-study.

Concerning the meta-variable priming, the qualitative study already delivered a good indication that there was a significant effect, with 0 of 10 non-primed participants and 7 out of 10 primed participants achieving a secure solution. However, since small samples tend to produce more extreme results, we would not have recommended basing study design decisions on these results. With a sample size of 40 participants in the present study, we are confident this is not a fluke and that the use of deception changes the behavior of participants dramatically. It would be useful to conduct even larger studies since we currently can only expect to find large effects. However, with regard to study design, we would very much want to catch medium or even small effects as well.

For the primary study, extending the sample size allowed us to conduct an A/B test to compare two frameworks. While H-F1 was not significant in this study due to the addition of the meta-variables and consequent correction for multiple tests, even the relatively small sample size in a normal developer study would be sufficient

[1] We later discovered there was a keyboard shortcut that participants seemed to have used by accident which stopped the recording.

Group		Search	Security search
Primed	Non-Secure (6)	3	1
	ABF (2)	1	1
	Secure (12)	8	7
Non-Primed	Non-Secure (16)	7	4
	ABF (2)	1	1

Table 3: # of participants who searched on the Internet in order to fill out the survey.

to get good results. That being said, the qualitative study already highlighted many of the problems faced by developers, and the interviews were very valuable in gaining deeper insights. We did not find much to add to the conclusions of the primary study of [42] other than having stronger evidence that library support as offered by Spring has tangible benefits.

A particularly salient benefit to qualitative developer studies is that fewer participants are needed. As such, unless rigorous evidence in the context of an A/B test is needed, we think that usable security research into developers is at a stage where qualitative studies have a lot to offer and encourage the community to be more accepting of them.

9. TAKE-AWAYS

Below, we summarize the main take-aways from our study.

- Task design has a huge effect on participant behavior and deception studies seem to be a promising method for examining a previously overlooked component of developer behavior when using student participants. That said, we must reiterate important limitations to this finding. We cannot make any claims concerning studies with professionals. It seems likely that even within a group comprising professionals, there will be multiple sub-groups that will react differently under priming. This will need to be examined in future work. It is also possible that a large portion of this effect is a study artifact. In any case, we recommend more experimentation concerning the design of developer studies. Currently, researchers base task and study design mostly on gut feelings. Since we have shown that one gets vastly different outcomes, we believe it is worth investing the effort into testing multiple designs in pilot studies instead of just going with one design as is currently often the case. We also believe more effort needs to be invested in understanding what motivates developers to implement security instead of focusing too narrowly on the easier measure of API usability.

- The use of Google by participants during surveys is problematic and researchers should not rely on answers reflecting the internal knowledge of the participants. This is particularly relevant for pre-screening surveys and we strongly recommend avoiding use of answers that can be googled for participant selection or condition assignment. If at all possible, we recommend that search behavior and web usage should be tracked, because a) thus, researchers can distinguish between internal knowledge and the ability to search for knowledge; and b) seeing when and what participants google is very enlightening in itself and a valuable research instrument.

- It is our belief that qualitative research into developer behavior offers a good cost/benefit trade-off and that many valuable insights can be gained without the need for large(r) sample sizes. In addition, the use of interviews as opposed to surveys avoids the googling problem. We hope that our comparison of quantitative and qualitative examination of the same topic encourages more qualitative studies and lowers the barriers to entering into this field, since recruitment of participants is one of the biggest challenges.

- While Acar et al. have found that programming language experience has a significant effect on the security of code produced in developer studies [11], we did not find a significant effect for this. In contrast to their study, our student sample had a much smaller range of programming skills; this could explain the lack of a measurable effect. This suggests that it might not be necessary to balance programming experience when working with students, thus simplifying random condition assignment. However, our power on this test was low so this result should be replicated before it is used confidently.

- We found copy/paste has a significant positive effect on the security of our participants' code. The way previous work was set up meant that they mainly found negative effects, thus potentially skewing the perception. We think highlighting the positive side of copy/paste behavior is important.

10. CONCLUSION

In this paper, we presented an extension of our qualitative developer study on password storage [42]. The extension had the dual goal of generating insights into the effect of design for developer studies, as well as furthering the understanding of why developers struggle to store passwords securely. We examined seven main hypotheses concerning both the primary study and the meta-study. We also compared our quantitative extension to the qualitative results of [42]. Our results suggest that priming or not priming participants allows us to study different aspects of student developer behavior. Priming can be used to discover usability problems of security APIs and test improvements with a straightforward study setup. Non-priming (i.e., deception), though, might be used to research why developers do not add security without study countermeasures or being prompted. However, more work is needed to validate the ecological validity of deception in this context. We also found many participants use Google to answer survey questions. This is potentially very damaging to studies that do not account for this effect and one of many reasons we see for using qualitative research methods such as interviews to study developers.

The next step in this research endeavor is designing an experiment to study the priming effect with professionals. Since it is unrealistic to expect even a small number of working professionals to sacrifice a full day to take part in a lab study, a different study design will be needed. We also plan to study additional design variables for developer studies to create a stronger foundation for conducting usable security and privacy research with professionals.

11. ACKNOWLEDGMENTS

This work was partially funded by the ERC Grant 678341: Frontiers of Usable Security.

12. REFERENCES

[1] Githut: A small place to discover languages in github. http://githut.info/, February 6, 2018 visited.

[2] Glipper is a clipboardmanager for gnome. https://launchpad.net/glipper, February 6, 2018 visited.

[3] Pypl popularity of programming language. http://pypl.github.io/PYPL.html, February 6, 2018 visited.

[4] The redmonk programming language rankings. http://redmonk.com/sogrady/2017/06/08/language-rankings-6-17/, February 6, 2018 visited.

[5] Tiobe index. http://www.tiobe.com/tiobe-index/, February 6, 2018 visited.

[6] Trendy skills: Extracting skills that employers seek in the it industry. http://trendyskills.com/, February 6, 2018 visited.

[7] W3techs web technology surveys: 'usage of server-side programming languages for websites'. https://w3techs.com/technologies/overview/programming_language/all, February 6, 2018 visited.

[8] R. Abu-Salma, M. A. Sasse, J. Bonneau, A. Danilova, A. Naiakshina, and M. Smith. Obstacles to the adoption of secure communication tools. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 137–153. IEEE, 2017.

[9] Y. Acar, M. Backes, S. Fahl, S. Garfinkel, D. Kim, M. L. Mazurek, and C. Stransky. Comparing the usability of cryptographic apis. In *Proceedings of the 38th IEEE Symposium on Security and Privacy*, 2017.

[10] Y. Acar, M. Backes, S. Fahl, D. Kim, M. L. Mazurek, and C. Stransky. You get where you're looking for: The impact of information sources on code security. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 289–305. IEEE, 2016.

[11] Y. Acar, C. Stransky, D. Wermke, M. L. Mazurek, and S. Fahl. Security developer studies with github users: Exploring a convenience sample. In *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[12] R. A. Armstrong. When to use the Bonferroni correction. 34:502–508, 2014.

[13] J. Bau, F. Wang, E. Bursztein, P. Mutchler, and J. C. Mitchell. Vulnerability factors in new web applications: Audit tools, developer selection & languages. *Stanford, Tech. Rep*, 2012.

[14] P. Berander. Using students as subjects in requirements prioritization. In *Empirical Software Engineering, 2004. ISESE'04. Proceedings. 2004 International Symposium on*, pages 167–176. IEEE, 2004.

[15] J. Bonneau and S. Preibusch. The password thicket: Technical and market failures in human authentication on the web. In *WEIS*, 2010.

[16] J. Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[17] S. Clarke. Using the cognitive dimensions framework to design usable apis.

[18] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[19] M. Egele, D. Brumley, Y. Fratantonio, and C. Kruegel. An empirical study of cryptographic misuse in android applications. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 73–84. ACM, 2013.

[20] S. Fahl, M. Harbach, Y. Acar, and M. Smith. On the ecological validity of a password study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, page 13. ACM, 2013.

[21] S. Fahl, M. Harbach, T. Muders, L. Baumgärtner, B. Freisleben, and M. Smith. Why eve and mallory love android: An analysis of android ssl (in) security. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 50–61. ACM, 2012.

[22] S. Fahl, M. Harbach, M. Oltrogge, T. Muders, and M. Smith. Hey, you, get off of my clipboard. In *International Conference on Financial Cryptography and Data Security*, pages 144–161. Springer, 2013.

[23] S. Fahl, M. Harbach, H. Perl, M. Koetter, and M. Smith. Rethinking ssl development in an appified world. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 49–60. ACM, 2013.

[24] M. Finifter and D. Wagner. Exploring the relationship between web application development tools and security. In *USENIX conference on Web application development*, 2011.

[25] K. Finstad. Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of Usability Studies*, 5(3):104–110, 2010.

[26] F. Fischer, K. Böttinger, H. Xiao, C. Stransky, Y. Acar, M. Backes, and S. Fahl. Stack overflow considered harmful? *The Impact of Copy & Paste on Android Application Security. CoRR abs/1710.03135*, 2017.

[27] J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[28] A. Forget, S. Chiasson, P. C. Van Oorschot, and R. Biddle. Improving Text Passwords Through Persuasion. In *Proceedings of the 4th Symposium on Usable Privacy and Security*, pages 1–12. ACM, jul 2008.

[29] P. Gorski and L. L. Iacono. Towards the usability evaluation of security apis. In *Proceedings of the Tenth International Symposium on Human Aspects of Information Security & Assurance (HAISA 2016)*, page 252. Lulu. com, 2016.

[30] P. A. Grassi, E. M. Newton, R. A. Perlner, A. R. Regenscheid, W. E. Burr, J. P. Richer, N. B. Lefkovitz, J. M. Danker, Y.-Y. Choong, K. Greene, et al. Digital identity guidelines: Authentication and lifecycle management. *Special Publication (NIST SP)-800-63B*, 2017.

[31] M. Green and M. Smith. Developers are not the enemy!: The need for usable security apis. *IEEE Security & Privacy*, 14(5):40–46, 2016.

[32] S. M. T. Haque, M. Wright, and S. Scielzo. A Study of User Password Strategy for Multiple Accounts. pages 1–3.

[33] M. Höst, B. Regnell, and C. Wohlin. Using students as subjects - a comparative study of students and professionals in lead-time impact assessment. *Empirical Software Engineering*, 5(3):201–214, 2000.

[34] I. Ion, R. Reeder, and S. Consolvo. "... no one can hack my mind": Comparing expert and non-expert security practices. In *SOUPS*, volume 15, pages 1–20, 2015.

[35] B. Kaliski. Pkcs# 5: Password-based cryptography specification version 2.0, Sept. 2000.

[36] A. J. Kimmel. *Ethical issues in behavioral research: Basic and applied perspectives*. John Wiley & Sons, 2009.

[37] R. E. Kirk. *Experimental design*. Wiley Online Library, 1982.

[38] K. Krombholz, W. Mayer, M. Schmiedecker, and E. Weippl. "I Have No Idea What I'm Doing" âĂŞ On the Usability of Deploying HTTPS. *USENIX Security*, pages 1–18, jun 2017.

[39] T. D. LaToza, G. Venolia, and R. DeLine. Maintaining mental

models: a study of developer work habits. In *Proceedings of the 28th international conference on Software engineering*, pages 492–501. ACM, 2006.

[40] D. Lazar, H. Chen, X. Wang, and N. Zeldovich. Why does cryptographic software fail?: a case study and open problems. In *Proceedings of 5th Asia-Pacific Workshop on Systems*, page 7. ACM, 2014.

[41] S. Nadi, S. Krüger, M. Mezini, and E. Bodden. Jumping through hoops: why do java developers struggle with cryptography apis? In *Proceedings of the 38th International Conference on Software Engineering*, pages 935–946. ACM, 2016.

[42] A. Naiakshina, A. Danilova, C. Tiefenau, M. Herzog, S. Dechand, and M. 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*, pages 311–328. ACM, 2017.

[43] J. Nielsen. *Usability engineering*. Elsevier, 1994.

[44] L. 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, 2011.

[45] N. Provos and D. Mazieres. A future-adaptable password scheme. In *USENIX Annual Technical Conference, FREENIX Track*, pages 81–91, 1999.

[46] I. Salman, A. T. Misirli, and N. Juristo. Are students representatives of professionals in software engineering experiments? In *Proceedings of the 37th International Conference on Software Engineering-Volume 1*, pages 666–676. IEEE Press, 2015.

[47] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. The emperor's new security indicators. In *Security and Privacy, 2007. SP'07. IEEE Symposium on*, pages 51–65. IEEE, 2007.

[48] J. Stylos and B. A. Myers. The implications of method placement on api learnability. In *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering*, pages 105–112. ACM, 2008.

[49] M. Svahnberg, A. Aurum, and C. Wohlin. *Using students as subjects - an empirical evaluation*. ACM, New York, New York, USA, Oct. 2008.

[50] R. Wash and E. Rader. Influencing mental models of security: a research agenda. In *Proceedings of the 2011 New Security Paradigms Workshop*, pages 57–66. ACM, 2011.

[51] K. Yakdan, S. Dechand, E. Gerhards-Padilla, and M. 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. IEEE, 2016.

APPENDIX

We used a seven-point rating scale according to [25].

A. PRE-SCREENING QUESTIONNAIRE

1. Gender: *Female/Male/Other/Prefer not to say*
2. Which university are you at? *University of Bonn/Other: [free text]*
3. In which program are you currently enrolled? *Bachelor Computer Science/Master Computer Science/Other: [free text]*
4. Your semester: [free text]
5. How familiar are you with Java?
   1 - Not familiar at all - 7 - Very familiar
6. How familiar are you with PostgreSQL?
   1 - Not familiar at all - 7 - Very familiar
7. How familiar are you with Hibernate?
   1 - Not familiar at all - 7 - Very familiar
8. How familiar are you with Eclipse IDE?
   1 -Not familiar at all - 7 - Very familiar

B. ENTRY SURVEY

Before solving the task, participants were asked questions Q5 - Q8 from the pre-screening questionnaire (Appendix A) one more time for consistency reasons. Additionally, they were asked two further questions:

1. *Expectation*:
   What is your expectation? Overall, this task is
   1 - Very difficult - 7 - Very easy
2. How familiar are you with JavaServer Faces (JSF)/Spring?
   1 - Not familiar at all - 7 - Very familiar

C. EXIT SURVEY

Questions asked after solving the task:

1. *Experience*
   Overall, this task was
   1 - Very difficult - 7 - Very easy
2. Do you think your solution is optimal? *No / Yes*
   - Why do you think your solution is (not) optimal? [free text]
3. I have a good understanding of security concepts.
   1 - Strongly disagree - 7- Strongly agree
4. How often do you ask for help facing security problems?
   1- Never - 7 - Every time
5. How often are you asked for help when somebody is facing security problems?
   1- Never - 7 - Every time
6. How often do you need to add security to the software you develop in general (Primed group: apart from this study)?
   1- Never - 7 - Every time
7. How often have you stored passwords in the software you have developed (Primed group: apart from this study)?
8. How would you rate your background/knowledge with regard to secure password storage in a database?
   1- Not knowledgeable at all - 7 Very knowledgeable
9. Do you think that you stored the end-user passwords securely? *No / Yes*
   - If Yes:
     - What did you do to store the passwords securely? [free text]
     - Do you think your solution is optimal? *No / Yes*
       - Why do you think your solution is (not) optimal? [free text]
     - Primed group: Do you think you would have stored end-user passwords securely, if you had not been told about it? Please explain your decision. [free text]
   - If No:
     - Why do you think that you did not store the passwords securely? [free text]

  - Non-Primed group: Were you aware that the task needed a secure solution? *No / Yes*
  - What would you do, if you needed to store the end-user passwords securely? [free text]

10. Did you use libraries to store the end-user passwords securely? *No / Yes*
    - If Yes:
      - Which libraries did you use to store the end-user passwords securely (in this study)? [free text]
      - Please name the most relevant library you have used to store the end-user passwords securely (in this study). [free text]
      - You have identified {*participant's answer*} as the most relevant library to store end-user passwords securely. How would you rate its ease of use in terms of accomplishing your tasks functionally / securely? *1- Very Difficult - 7- Very Easy*
        Please explain your decision. [free text]
      - Usability scale for {*participant's answer*} (see C.1)
11. JSF/ Spring supported me in storing the end-user password securely. *1 - Strongly disagree - 7- Strongly agree*
    Please explain your decision. [free text]
12. JSF/ Spring prevented me in storing the end-user password securely. *1 - Strongly disagree - 7- Strongly agree*
    Please explain your decision. [free text]
13. JSF/ Spring: Usability scale (see C.1); the term *library* was replaced by *framework*.
14. Have you used Java APIs / libraries to store end-user passwords securely before? *No / Yes*
    - If Yes:
      - Which Java APIs / libraries to store end-user passwords securely have you used before? [free text]
      - What is your most-used API / library for secure password storage? [free text]
      - How would you rate its ease of use in terms of accomplishing your tasks functionally? *1- Very Difficult - 7- Very Easy*
        Please explain your decision. [free text]
      - How would you rate its ease of use in terms of accomplishing your tasks securely? *1- Very Difficult - 7- Very Easy*
        Please explain your decision. [free text]

C.1 Usability scale from [9]

By contrast to [9] we dropped the option "does not apply" for the last two questions, Q10 and Q11. Used scale in our study:

Please rate your agreement to the following questions on a scale from 'strongly agree' to 'strongly disagree.' (Strongly agree; agree; neutral; disagree; strongly disagree). Calculate the 0-100 score as follows: 2.5 * (5-Q_1 + $\sum_{i=2..10}$ (Q_i-1)); for the score, Q11 is omitted.

- I had to understand how most of the assigned library works in order to complete the tasks.
- It would be easy and require only small changes to change parameters or configuration later without breaking my code.
- After doing these tasks, I think I have a good understanding of the assigned library overall.
- I only had to read a little of the documentation for the assigned library to understand the concepts that I needed for these task.
- The names of classes and methods in the assigned library corresponded well to the functions they provided.
- It was straightforward and easy to implement the given tasks using the assigned library.
- When I accessed the assigned library documentation, it was easy to find useful help.
- In the documentation, I found helpful explanations.
- In the documentation, I found helpful code examples.

Please rate your agreement to the following questions on a scale from 'strongly agree' to 'strongly disagree'. (Strongly agree; agree; neutral; disagree; strongly disagree).

- When I made a mistake, I got a meaningful error message/exception.
- Using the information from the error message/ exception, it was easy to fix my mistake.

C.2 Demographics

- Please select your gender. *Female/Male/Other/Prefer not to say*
- Age: [free text]
- What is your current occupation? *Student Undergraduate/Student Graduate/Other: [free text]*
- At which university are you currently enrolled? *University of Bonn / University of Aachen*
- Which security lectures did you pass in your Bachelor/Master programme? *(To select)/Other: [free text]*
- Currently, do you have a part-time job in the field of Computer Science? If yes, please specify: [free text]
- How many years of experience do you have with Java development? *< 1 year/ 1 - 2 years/ 3 - 5 years/ 6 - 10 years/ 11+ year*
- What is your nationality? [free text]
- Thank you for answering the questions! If you have any comments or suggestions, please leave them here: [free text]

D. SECURITY SCORE

We used the following security score from Naiakshina et al. [42] for the evaluation of participants' solutions:

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` [35] and at least $2^{10} = 1\,024$ for `bcrypt` [45] (+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).

E. SECURITY RESULTS

Table 4 summarizes the security evaluation for participants' implemented solutions as introduced in [42] with slightly modifications, e.g., the digest size of `bcrypt` was changed from 192 bits to 184 bits, reasonable by practical implementation standards. Table 4 also considers participants who attempted to store end-user passwords securely during programming, but removed the security code from their final solutions (ABF = attempted but failed).

	Time	Functionality	Security					
			Hashing			Salt		
	(hh:mm)	Storage working	Hashing function (at most +2)	Digest size (bits) (+1 if $\geq$ 160)	Iteration count (at most +1)	Generation (at most +2)	Length (bits) (+1 if $\geq$ 32)	**Total** (7)
JN2*	04:05	✓						
JN3*	03:01	✓						
JN4*	04:11	✗						
JN5*	05:30	✗						
JN6	05:13	✓						
JN7	07:33	✗						
JN8	07:33	✗						
JN9	06:08	✓	SHA1	160	1	end-user email address	8	3 (ABF)
JN10	03:45	✓						
JN11	06:36	✗						
JP1*	04:55	✓						
JP2*	03:12	✓	PBKDF2(SHA256)	512	1000	SecureRandom	256	5.5
JP3*	05:29	✓	SHA256	256	1			2
JP4*	04:12	✓	PBKDF2(SHA1)	160	20000	SecureRandom	64	6
JP5*	06:32	✓						
JP6	07:33	✗						
JP7	06:08	✗	BCrypt	184	2^{12}	SecureRandom	128	6
JP8	07:22	✗						
JP9	07:18	✗	BCryp	184	2^{8}	pgcrypto	128	5 (ABF)
JP10	04:45	✓	SHA256	256	1			2
SN1*	03:15	✓						
SN2*	02:24	✓						
SN3*	02:01	✓						
SN4*	04:01	✓	BCrypt	184	2^{10}	SecureRandom	128	6 (ABF)
SN5*	04:50	✓						
SN6	07:03	✗						
SN7	05:35	✓						
SN8	07:33	✗						
SN9	05:31	✓						
SN10	03:23	✓						
SP1*	03:15	✓	BCrypt	184	2^{10}	SecureRandom	128	6
SP3*	07:00	✗	BCrypt	184	2^{10}	SecureRandom	128	6
SP4*	03:39	✓	BCrypt	184	2^{10}	SecureRandom	128	6
SP5*	03:44	✗						
SP6	07:33	✓						
SP7	01:49	✓	BCrypt	184	2^{11}	SecureRandom	128	6
SP8	05:59	✗	#					0 (ABF)
SP9	05:50	✓	BCrypt	184	2^{10}	SecureRandom	128	6
SP10	05:53	✓	BCrypt	184	2^{10}	SecureRandom	128	6
SP11	03:15	✓	BCrypt	184	2^{10}	SecureRandom	128	6

Table 4: Password security evaluation, including participants who attempted to implement security but failed (ABF).
Labeling of participants: S = Spring, J = JSF, P = Priming, N = Non-Priming
* = Used for the qualitative study in [42].
# = Used Spring Security's `PasswordEncoder` interface without deciding for an algorithm.

F. COPY/PASTE WEBSITES

Table 5 lists all websites used by participants who implemented user password storage security. We also examined websites used by participants who attempted to store passwords securely, but removed all security-relevant code from their solutions (ABF = attempted but failed). In order to search for programming security attempts we used the Unix utility *grep*. The following search words were used for security attempt identification: encode, sha, pbkdf2, scrypt, hashpw, salt, MD5, passwordencoder, iterations, pbekeyspec, argon2, bcrypt, messagedigest, crypt.

Participant	Security score	Website	Description	Most insecure example	Obvious example	Most secure example
JN9	3 (ABF)	www.sha1-online.com/sha1-java/	Blog Post: SHA1 Java	2	2	2
JP2	5.5	https://www.owasp.org/index.php/Hashing_Java	OWASP: General Hashing Example	6	6	6
		https://stackoverflow.com/questions/18268502/how-to-generate-salt-value-in-java	Stack Overflow: Salt Example	3	3	3
		https://howtodoinjava.com/security/how-to-generate-secure-password-hash-md5-sha-pbkdf2-bcrypt-examples/#PBKDF2WithHmacSHA1	Blog Post: Hashing Example	1	1	7
JP3	2	www.mkyong.com/java/java-sha-hashing-example/	Blog Post: Hashing Example	2	2	2
JP4	6	http://blog.jerryorr.com/2012/05/secure-password-storage-lots-of-donts.html	Blog Post: Hashing Example	5.5	6	6
JP7	6	http://javaandj2eetutor.blogspot.de/2014/01/jsf-login-and-register-application.html	Blog Post: Hashing Example	0	0	0
		www.mindrot.org/projects/jBCrypt/	Documentation: Java Implementation jBCrypt	6	6	6
JP9	5.5 (ABF)	https://www.meetspaceapp.com/2016/04/12/passwords-postgresql-pgcrypto.html	Blog Post: Hashed Passwords with PostgreSQL's pgcrypto	5.5	5.5	5.5
JP10	2	https://stackoverflow.com/questions/33085493/hash-a-password-with-sha-512-in-java	Stack Overflow: Hashing Example	3	3	5
		https://stackoverflow.com/questions/3103652/hash-string-via-sha-256-in-java	Stack Overflow: Hashing Example	2	2	2
		https://docs.oracle.com/javase/7/docs/api/java/security/MessageDigest.html	Documentation: Class MessageDigest	1	2	2
		https://stackoverflow.com/questions/11665360/convert-md5-into-string-in-java	Stack Overflow: Convert MD5 into String in Java	1	1	1
SN4	6 (ABF)	http://websystique.com/spring-security/spring-security-4-password-encoder-bcrypt-example-with-hibernate/	Blog Post: Hashing Example	6	6	6
SN8	0	https://dzone.com/articles/spring-mvc-example-for-user-registration-and-login-1	Blog Post: Hashing Example	0	0	0
SP1	6	https://stackoverflow.com/questions/25844419/spring-bcryptpasswordencoder-generate-different-password-for-same-input	Stack Overflow: Hashing Example	6	6	6
SP3, SP4, SP11	6	www.mkyong.com/spring-security/spring-security-password-hashing-example/	Blog Post: Hashing Example	6	6	6
SP7	6	https://hellokoding.com/registration-and-login-example-with-spring-xml-configuration-maven-jsp-and-mysql/	Blog Post: Hashing Example	6	6	6
SP8	0 (ABF)	www.websystique.com/springmvc/spring-mvc-4-and-spring-security-4-integration-example/	Blog Post: Hashing Example	6	6	6
SP9	6	https://stackoverflow.com/questions/18653294/how-to-correctly-encode-password-using-shapasswordencoder	Stack Overflow: Hashing Example	5.5	6	6
SP10	6	https://stackoverflow.com/questions/42431208/password-encryption-in-spring-mvc	Stack Overflow: Password Encryption in Spring MVC	6	6	6

Table 5: Websites from which participants copied and pasted code for password storage.

API Blindspots: Why Experienced Developers Write Vulnerable Code

Daniela Seabra Oliveira, Tian Lin, Muhammad Sajidur Rahman, Rad Akefirad[α],
Donovan Ellis, Eliany Perez, Rahul Bobhate, Lois A. DeLong[ν],
Justin Cappos[ν], Yuriy Brun[μ], Natalie C. Ebner
University of Florida [α]Auto1.inc [ν]New York University [μ]University of Massachusetts Amherst
daniela@ece.ufl.edu, {lintian0527, rahmanm, donovanmellis, elianyperez, rabo, natalie.ebner}@ufl.edu,
rad@akefirad.com, lad278@nyu.edu, jcappos@nyu.edu, brun@cs.umass.edu

ABSTRACT

Despite the best efforts of the security community, security vulnerabilities in software are still prevalent, with new vulnerabilities reported daily and older ones stubbornly repeating themselves. One potential source of these vulnerabilities is shortcomings in the used language and library APIs. Developers tend to trust APIs, but can misunderstand or misuse them, introducing vulnerabilities. We call the causes of such misuse blindspots. In this paper, we study API blindspots from the developers' perspective to: (1) determine the extent to which developers can detect API blindspots in code and (2) examine the extent to which developer characteristics (i.e., perception of code correctness, familiarity with code, confidence, professional experience, cognitive function, and personality) affect this capability. We conducted a study with 109 developers from four countries solving programming puzzles that involve Java APIs known to contain blindspots. We find that (1) The presence of blindspots correlated negatively with the developers' accuracy in answering implicit security questions and the developers' ability to identify potential security concerns in the code. This effect was more pronounced for I/O-related APIs and for puzzles with higher cyclomatic complexity. (2) Higher cognitive functioning and more programming experience did not predict better ability to detect API blindspots. (3) Developers exhibiting greater openness as a personality trait were more likely to detect API blindspots. This study has the potential to advance API security in (1) design, implementation, and testing of new APIs; (2) addressing blindspots in legacy APIs; (3) development of novel methods for developer recruitment and training based on cognitive and personality assessments; and (4) improvement of software development processes (e.g., establishment of security and functionality teams).

1. INTRODUCTION

Despite efforts by the security community, software vulnerabilities are still prevalent in all types of computer devices [56]. Symantec Internet Security reported that 76% of all websites scanned in 2016 contained software vulnerabilities and 9% of those vulnerabilities were deemed critical [56]. According to a 2016 Veracode report [53] on software security risk, 61% of all web applications contained vulnerabilities that fell into the Open Web Application Security Project (OWASP) Top 10 2013 vulnerability categories [39] (e.g., information leakage: 72%, flawed cryptographic implementations: 65%, carriage-return-line-feed (CRLF) injection: 53%). Further, 66% of the vulnerabilities represented programming practices that failed to avoid the "top 25 most dangerous programming errors" identified by CWE/SANS [12]. In addition, new instances of existing, well-known vulnerabilities, such as SQL injections and buffer overflows, are still frequently reported in vulnerability databases [50, 26]. These data affirm that current software security awareness efforts have not eradicated these problems in practice.

A contributing factor in the introduction of software vulnerabilities may be the way developers view the programming language resources they routinely use. APIs provide developers with high-level abstractions of complex functionalities and are crucial in scaling software development. Yet, studies on API usability [46, 47] and code comprehension [25] show that developers experience a number of challenges while using APIs, such as mapping developer-specific requirements to proper usage protocols, making sense of internal implementation and related side effects, and deciding between expert opinions. Further, misunderstandings in developers' use of APIs are frequently the cause of security vulnerabilities [9, 14, 45]. Developers often blindly trust APIs and their misunderstanding of the way API functions are called may lead to blindspots, or oversights regarding a particular function usage (e.g., assumptions, results, limitations, exceptions). More significantly, when developers use an API function, they may behave as if they are outsourcing any security implications of its use [37]. That is, they do not see themselves as responsible for the correct usage of the function and any possible resulting security consequences.

An API security blindspot is a misconception, misunderstanding, or oversight [9] on the part of the developer when using an API function, which leads to a violation of the recommended API usage protocol with possible introduction of security vulnerabilities. Blindspots can be caused by API functions whose invocations have security implications that are not readily apparent to the developer. It is analogous to the concept of a car blindspot, an area on the side of a car that is not visible to the driver that can lead to accidents. For an example of an API blindspot, consider the `strcpy()` function from the C standard library. For almost three decades [41], this function has been known to lead to a buffer overflow vulnerability if developers do not check and match sizes of the destination and source arrays. Yet, developers tend to have a blindspot with respect

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

to this function. In a recent study [37], a developer who could not detect a buffer overflow in a programming scenario mentioned that *"It's not straightforward that misusing strcpy() can lead to very serious problems. Since it's part of the standard library, developers will assume it's OK to use. It's not called unsafe_strcpy() or anything, so it's not immediately clear that that problem is there."*

In this paper, we present an empirical study of API blindspots from the developers' perspective, and consider personal characteristics that may contribute to the development of these blindspots. Our study goals were to: (1) determine developers' ability to detect API blindspots in code and (2) examine the extent to which developer characteristics (i.e., perception of code correctness, familiarity with code, confidence in correctly solving the code, professional experience, cognitive function, personality) affected this capability. We also explored the extent to which API function or programming scenario characteristics (i.e., category of API function and cyclomatic complexity of the scenario) contributed to developers' ability to detect blindspots.

We recruited 109 developers, including professional developers and senior undergraduate and graduate students (professionals = 70, students = 39, mean age = 26.4, 80.7% male). Developers worked online on six programming scenarios (called puzzles) in Java. Each puzzle contained a short code snippet simulating a real-world programming scenario. Four of the six puzzles contained one API function known to cause developers to experience blindspots. The other two puzzles involved an innocuous API function. Puzzles were developed by our team and were based on API functions commonly reported in vulnerability databases [36, 49] or frequently discussed in developer forums [51]. The API functions considered addressed file and stream handling, cryptography, logging, SQL operations, directory access, regular expressions (regex), and process manipulation. Following completion of each puzzle, developers responded to one open-ended question about the functionality of the code and one multiple-choice question that captured developers' understanding of (or lack thereof) the security implication of using the specific API function. After completing all puzzles, each developer provided demographic information and reported their experience and skills levels in programming languages and technical concepts. Developers then indicated endorsement of personality statements based on the Five Factor Personality Traits model [13] and completed a set of cognitive tasks from the NIH Cognition Toolbox [21] and the Brief Test of Adult Cognition by Telephone (BTACT; modified auditory version for remote use) [58].

Using quantitative statistics, we generated the following novel findings:

1. Presence of API blindspots in puzzles reduced developers' accuracy in answering the puzzles' implicit security question and reduced developers' ability to identify potential security concerns in the code.
2. API functions involving I/O were particularly likely to cause security blindspots in developers.
3. Developers were more susceptible to API blindspots for more complex puzzles, as measured by cyclomatic complexity.
4. Developers' cognitive function and their expertise and experience with programming did not predict their ability to detect API blindspots.
5. Developers exhibiting greater openness as a personality trait were more likely to detect API blindspots.

These results have the potential to inform the design of APIs that are inherently more secure. For example, testing and validation of API functions should take into account potential security blindspots developers may have, particularly for certain types of API functions (e.g., I/O). Furthermore, since our data suggest that experience and cognition may not predict developers' ability to detect API blindspots, it corroborates the validity of the emerging practice of establishing separate functionality and security development teams. Separate teams for these domains may be a better strategy to assure secure software development than sole reliance on one group of experts to simultaneously address both aspects.

The remainder of this paper is organized as follows. Section 2 reports on the study methodology and the development of the puzzles. Section 3 assesses the results, while Section 4 discusses some of the implications of these findings. Section 5 places this study in the context of related work, and Section 6 summarizes its primary contributions.

2. METHODOLOGY

This section presents the study methodology, describing recruitment, participant management, and procedures. Data collection took place between December 2016 and November 2017.

2.1 Participants

This study, approved by the University of Florida IRB, targeted developers who actively worked with Java. These individuals were recruited from the United States, Brazil, India, Bangladesh, and Malaysia via a number of recruitment mechanisms, including flyers and handouts disseminated throughout the university campus, particularly in locations frequented by students and professionals with programming experience (e.g., Computer Science and Engineering departments), social media advertisements (i.e., Facebook, Twitter, and LinkedIn), ads on online computer programming forums, Computer Science/Engineering department groups, and contacts via the authors' personal networks of computer programmers at universities and software development companies in the United States, Brazil, India, Bangladesh, and Malaysia. We also used a word-of-mouth recruiting technique, which gave participants the option to refer friends or colleagues. Participants were informed that the purpose of the study was to investigate how developers interpret and reason about code. As we aimed to have developers work on the programming tasks as naturally as possible, without any priming or nudging towards software security aspects, we did not explicitly mention that code security was the metric of interest. Figure 1 summarizes the demographic information of participating developers.

As shown in Figure 1, developers in the final sample size ($N = 109$) ranged between the ages of 21 and 52 years ($M = 26.67$, $SD = 5.28$) and were largely male ($n = 88$, 80.7%). The sample was composed of 70 (64.2%) professional developers and 39 (35.8%) senior undergraduate or graduate students in Computer Science and Computer Engineering though in this paper, we collectively refer to all participants as "developers". The large majority of developers ($n = 83$, 82.5%) had been programming in Java for two or more years, and almost all developers reported at least a working knowledge of Java ($n = 101$, 97.1%). Student participants self-reported a relatively high programming experience ($M = 5.8$ years, $SD = 5.8$), probably because they had been programming before entering university or had been students for more than six years (e.g., PhD students).

We received a total of 168 emails from interested developers, 33 (19.6%) of which were not included in the study because they never

	Professionals ($n = 70$) Mean (SD)/ %	Students ($n = 39$) Mean (SD)/ %
Gender		
Male (88)	81.4	79.5
Female (21)	18.6	20.5
Age		
Male (88)	28.0 (6.0)	24.4 (2.1)
Female (21)	27.8 (6.2)	24.4 (2.2)
Years of Programming		
	6.3 (3.5)	5.8 (5.8)
Highest Degree Earned		
High School	1.4	0.0
Some College	0.0	2.6
Associates	1.4	2.6
Bachelor's	40.0	56.4
Some Graduate School	11.4	5.1
Graduate-Level Degree	45.7	33.3
Annual Income		
0–$39,999	45.7	69.2
$40,000–$70,000	22.9	15.4
$70,001–$100,000	20.0	12.8
$100,001–$200,000	11.4	2.6
>$200,000	0.0	0.0
Race/Ethnicity		
American Indian/Alaskan	1.4	2.6
Asian	81.4	92.3
African American	2.9	0.0
Hawaiian/Pacific Islander	0.0	0.0
White	10.0	2.6
Other/Multi-racial	4.3	2.6
Country of Residence		
United States	72.3	94.9
Bangladesh	15.7	2.6
Brazil	8.6	2.6
Malaysia	1.4	0.0

Figure 1: Demographic and professional expertise/experience information about participating developers by professional group.

signed the informed consent form or signed the form but did not continue with the assessment. The remaining 135 developers received a personalized link to the study assessment, which was hosted online on the Qualtrics platform. We had to discard data from 26 (19.3%) developers because of incomplete entries or technical/browser incompatibility issues related to the audio recording (see details below). Unless otherwise stated, we report our results based on a sample of 109 developers, who proceeded through all study procedures as instructed and completed the tasks with valid responses.

2.2 Procedure

After initial contact with interested developers, an online screening questionnaire determined study eligibility (e.g., sufficient knowledge with Java, fluency in the English language, age over 18 years). Eligible developers received a digital informed consent form, which disclosed study procedures, the minimal risk from participating, and potential data privacy and anonymity issues. After providing their digital signature, developers received a personalized link to the online instrument. Each developer was assigned a unique identifier to assure confidentiality. Developers were strongly encouraged to complete the study in two separate sittings to counteract possible fatigue effects (one sitting to work on the puzzles and complete the demographic questionnaire, and the other sitting to complete the psychological/cognitive assessment). Student developers were compensated with a US$20 Amazon gift card, while professionals received a US$50 Amazon gift card, as professional developers had a larger financial incentive in consideration of their relatively high-paying jobs and their more limited availability, as approved by our IRB. The study procedure comprised five parts. The first part (Puzzles) involved responding to the programming puzzles and related questions (see Section 2.3). The second part (Demographics) asked basic demographic questions about the subject, including age, gender, race/ethnicity, education, field of study, employment status, and primary language.

The third part (Professional Experience and Expertise) included questions about the developers' technical proficiency and years of programming experience in six commonly used programming languages (i.e., Java, Python, C/C++, PHP, Visual Basic.Net, and JavaScript). A free-text response field was provided for developers to record their preferred programming language, if it was not listed. Developers also indicated their level of knowledge in and experience with 17 programming concepts and technologies identified from the literature and via job postings for software developers [6, 30] (e.g., SQL/MySQL, Cryptography, File compression, Networking, HTTP/HTTPS, I/O operations).

The fourth part (Personality Assessment) used the Big Five Inventory (BFI) questionnaire to measure aspects of personality [29]. This questionnaire contains 44 items to assess five personality dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Developers rated the extent to which they endorsed each personality statement on a Likert scale (1 = disagree strongly; 5 = agree strongly). We computed the sum score across all items for each of the five personality dimensions.

The fifth part (Cognitive Assessment) comprised two instruments: the Oral Symbol Digit Test from the NIH Toolbox [21] and the Brief Test of Adult Cognition by Telephone (BTACT) [58]. Figure 2 illustrates the Oral Symbol Digit Test. This test is a brief measure of processing speed and working memory. In this task, developers were presented with a coding key containing nine abstract symbols, each paired with a number between 1 and 9. They were then given 120 seconds to call out as many numbers that went with the corresponding symbols, in the order presented and without skipping any. The BTACT is a battery of cognitive processing tasks for adults of different ages and takes approximately 20 minutes to complete. The BTACT sub-tests refer to episodic verbal memory, working memory, verbal fluency, inductive reasoning, and processing speed. Figure 3 presents instructions for the BTACT Word List Recall task, which measures immediate and delayed episodic memory for verbal material. This particular task asked developers to recall a number of spoken words. The Oral Symbol Digit Test and the BTACT were chosen based on the cognitive processes (e.g., reasoning, working memory, processing speed) developers likely use when working on code. Traditionally, the Oral Symbol Digit Test and the BTACT are administered in-person and over the phone, respectively. Given the online format of our study, we implemented browser-based audio recordings of the two measures. In particular, audio narrations for all the tasks instructions were created, with calculated timings, vocal inflections, and pauses. Formal time limits were maintained. To capture oral responses, we built an audio recording plugin leveraging the Qualtrics JavaScript API. All task modifications underwent pilot testing to ensure that content and response sensitivity was maintained. As part of the study infrastructure, recorded audio files were sent to a secure and encrypted study server and were stored in an anonymized fashion. Trained coders coded these files for performance in the various tasks. For

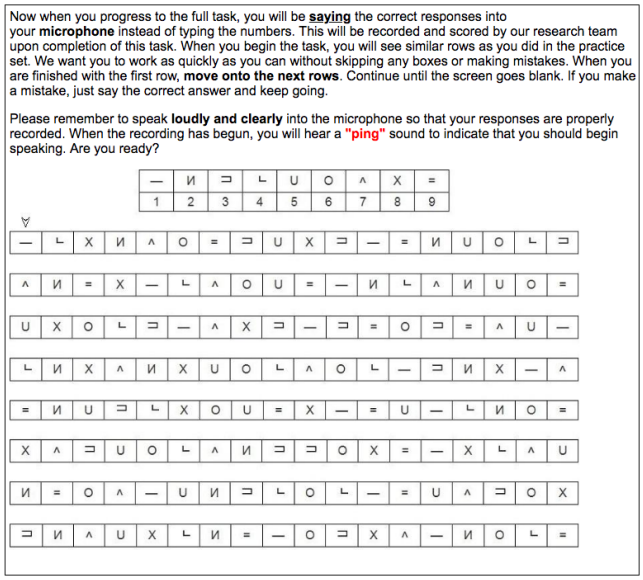


Figure 2: Oral Symbol Digit task.

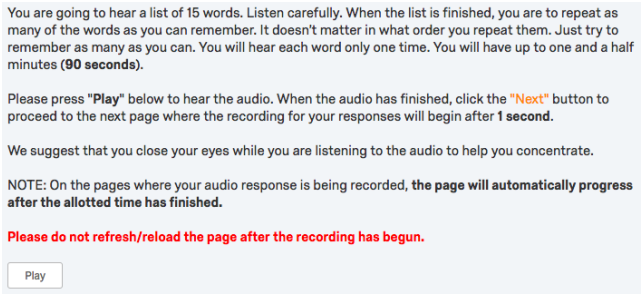


Figure 3: BTACT Word List Recall task.

the Oral Symbol Digit task we computed a sum score of correct responses. For the BTACT, we aggregated the total number of correct responses in each of the cognitive subset tasks.

2.3 Programming puzzles

This section describes the development of the programming puzzles and their characteristics. We defined a puzzle as a snippet of code simulating a real-world programming scenario. Our goal was to create concise, clear, and unambiguous puzzles that related to real-life programming tasks, while removed of code or functionality not needed for understanding the primary functionality of the code snippet. We developed two types of puzzles: those with and those without a blindspot. A blindspot puzzle targeted one particular Java API function, known to cause developers to misunderstand the security implications of its usage [40, 32]. The non-blindspot puzzles involved innocuous API functions in code context that strictly followed standard API usage protocol and code security guidelines. Puzzle development involved a two-phase, iterative process, which lasted from April 2016 to December 2016.

We chose Java as the programming language because of its rich and well-developed set of libraries and API functionalities, which can perform a diverse set of operations (including security tasks), such as I/O, multithreading, networking, random number generation, cryptography, and hashing. Java has a long-standing popularity within developer communities who use it to work on software products for different platforms, including web, mobile, and enterprise [15]. Besides being popular among professional developers, Java's wide availability of toolkits, tutorials, and online/offline resources has made it a popular choice for those learning object-oriented programming. It is the second most used programming language in GitHub repositories after Javascript [22] and was voted the third most popular technology by developers who frequently visit Stack Overflow with programming related Q&As [52]. These features made Java a good choice of programming language for our study, as we aimed to recruit from a diverse pool of developers.

Puzzle creation. This process began with a literature review to determine secure Java coding practices and the potential risks of misusing Java APIs. For puzzle selection and design principles, we were guided by the Open Web Application Security Project (OWASP) [40], CERT's secure coding guidelines [32], vulnerability databases [36, 49], HPE's Software Security Taxonomy [19, 57], and the Java API official documentation [28]. We also leveraged programming Q&A forums, such as Stack Overflow [51] to select commonly discussed API functions. We did not to look for candidate blindspot functions in bug repositories because we did not want developers in our study to fix bugs in code. Instead, our aim was to analyze whether developers would detect improper API usage to infer the insecure behaviors to which it may lead. Thus, all the code snippets were free from bugs and compilation errors, and were compatible with Java standard edition version 7 or higher. Our API function selection process included functions from different categories, including I/O, cryptography, SQL, and string.

We initially identified 61 API function candidates and created 61 corresponding puzzles, each targeting one particular function. This pool encompassed a variety of Java API misuse scenarios, including file I/O operations, garbage collection, de/serialization, cryptography, secure connection establishment, command line arguments/user inputs processing for database query, logging, user authentication, and multithreading.

Each puzzle contained four parts: (1) the puzzle scenario itself; (2) an accompanying code snippet; (3) a question about the puzzle's functionality, and (4) a multiple-choice question, which, for blindspot puzzles was implicitly related to code security and for non-blindspot puzzles related to code functionality. Developers' accuracy on the multiple-choice question served as the central outcome measure. It captured the developers' understanding of the blindspot in the code.

Puzzle review and final selection. Three co-authors, who had not created the puzzles, independently reviewed the initial set of 61 blindspot puzzles, together with eight non-blindspot puzzles, to ensure puzzle accuracy, legibility, coherence, and relevance to real-life programming situations. The specific criteria used for puzzle approval were:

1. Is the scenario clear and realistic?
2. Is the code snippet clear and concise (maximum one screen)?
3. Does the code snippet compile and run if provided with the necessary Java packages?
4. Does the choice of API function contribute to diversity in the puzzle set (API function category, blindspot vs. non-blindspot function, blindspot by function omission vs. presence, and number of parameters)?
5. Does the multiple-choice question have only one answer without ambiguity?

6. For blindspot puzzles, does the multiple-choice question address the security implications subtly without priming developers about security concerns?

7. For blindspot puzzles, is there a way to rewrite the puzzle to address the security vulnerability, thus avoiding the blindspot?

To be contained in the final pool, puzzles had to be independently approved by all three reviewers.

The final set comprised 16 blindspot puzzles and eight non-blindspot puzzles, which varied in the following categories: (1) blindspot vs. non-blindspot; (2) API usage category; and (3) cyclomatic complexity.

Blindspot vs. non-blindspot. We included non-blindspot puzzles as a control and to cover the security focus of the study. Blindspot puzzles were bug-free and functionally correct, but could cause a blindspot in developers when they used them, thus having the potential to cause developers to introduce one of the following vulnerabilities in code: (1) arbitrary code/command injection; (2) DoS (exhaustion of local resources); (3) time-of-check-to-time-of-use (TOCTTOU); (4) sensitive data disclosure; (5) broken or flawed cryptographic implementation, and (6) insecure file and I/O operations.

API usage category. The puzzles referred to three different API usage contexts: (1) I/O, involving operations, such as reading and writing from/to streams and files, internal memory buffers, and networking activity; (2) Crypto, involving functions handling cryptographic operations, such as encryption, decryption, and key agreement, and (3) String, involving functions that perform string processing or manipulation, or queries and user input.

Cyclomatic complexity [31]. Puzzles varied in their cyclomatic complexity, defined as a quantitative measure of the number of linearly independent paths in the source code. We classified the cyclomatic complexity of each puzzle into one of three levels: an integer value of low (cyclomatic complexity of 1–2), medium (cyclomatic complexity of 3–4), or high (cyclomatic complexity > 4) complexity.

We divided the final set of 24 puzzles into four subsets, each set containing six puzzles, four with a blindspot and two without a blindspot. This counterbalancing scheme ensured that each puzzle set was comparable regarding representation of API category and cyclomatic complexity. Statistical analysis found no effects for puzzle sets as covariate, confirming successful counterbalancing. We assigned each developer randomly to one of the four puzzle sets.

Figure 4 illustrates a blindspot puzzle, involving a Java Runtime API usage. The puzzle scenario was presented to developers as follows:

"You are asked to review a utility method written for a web application. The method, `setDate`, changes the date of the server. It takes a `String` as the new date ("dd-mm-yyyy" format), attempts to change the date of the server, and returns `true` if it succeeded, and `false` otherwise. Consider the snippet of code below (assuming the code runs on a Windows operating system) and answer the following questions, assuming that the code has all required permissions to execute."

After presenting the code snippet, developers were asked which of the following statements would be correct if the `setDate()` method was invoked with an arbitrary `String` value as the new date:

```
1  // OMITTED: Import whatever is needed
2  public final class SystemUtils {
3    public static boolean setDate (String date)
4        throws Exception {
5      return run("DATE " + date);
6    }
7
8    private static boolean run (String cmd)
9        throws Exception {
10     Process process = Runtime.getRuntime().
       exec("CMD /C " + cmd);
11     int exit = process.waitFor();
12
13     if (exit == 0)
14       return true;
15     else
16       return false;
17   }
18 }
```

Figure 4: Sample blindspot puzzle targeting a Java Runtime API usage.

a. If the given `String` value does not conform to the "dd-mm-yyyy" format, an exception is thrown.

b. The `setDate()` method cannot change the date.

c. The `setDate()` method might do more than change the date.

d. The return value of the `waitFor()` method is not interpreted correctly (lines 14–17).

e. The web application will crash.

The correct answer is option 'c'. A close inspection of the code shows that the `Runtime.getRuntime.exec()` method executes, in a separate process, the specified string command (line 10) which is provided by the `setDate()` method. The `setDate()` method takes a `String` type argument and does not implement any input sanitization and validation, which makes it vulnerable to format string injection attacks. For example, calling the `setDate()` method with "`10-12-2015 && shutdown /s`" as the argument changes the date and turns off the server. Either the argument for `setDate()` method has to be sanitized or its type should be an instance of the Java `Date` class, which can be formatted as a `String` type before passing to the `Runtime.getRuntime.exec()` method. As the outcome of the program (executing in a benign or malicious fashion) depends solely on the (un)sanitized input of the `Runtime.exec()` method, the blindspot API function for this puzzle is `Runtime.exec()`.

Table 1 details the complete list of puzzles used in the study with information about the puzzle's vulnerability, the API usage context, and the Java API function targeted for both blindspot and non-blindspot puzzles.

After completion of a puzzle and related security questions, developers responded to the following four questions about their puzzle perceptions using a Likert scale (1 = not at all to 10 = very) : (1) **Difficulty** (How difficult was this scenario?); (2) **Clarity** (How clear was this scenario?); (3) **Familiarity** with the API functions presented in the code snippet (How familiar were you with the functions in this scenario?), and (4) **Confidence** (How confident were you that you solved the scenario correctly?).

Table 1: Overview of the final puzzle set with information about puzzle vulnerability, API usage context, and Java API function targeted in each puzzle.

Has Blindspot	Vulnerability (if any)	Description	API Usage Context	Targeted (non) Blindspot API function
YES	TOCTTOU race condition	A program that performs two or more file operations on a single file name or path name creates a race window between the two file operations. Thus, `File.createNewFile()` may overwrite an existing file even after the overwrite flag is set to false.	I/O	`java.io.File.createNewFile()`
YES	TOCTTOU race condition	`File.renameTo()` relies solely on file names for identification, which does not guarantee that the file renamed is the same file that was opened, processed, and closed, thus being vulnerable to the TOCTTOU vulnerability.	I/O	`java.io.File.renameto()`
YES	Resurrectable object	JVM does not guarantee the timing for garbage collection of an object. Malicious subclasses that override the `Object.finalize()` method can resurrect objects meant for garbage collection.	I/O	`java.lang.Object.finalize()` in the context of `java.io.File.delete()`
YES	Ambiguous return value	The `getSize()` method of the `ZipEntry` class is not reliable because it returns -1 when the size of the entry (file) is unknown. It allows an attacker to forge the field in the zip entry, which can lead to a DoS or data corruption attack.	I/O	`java.util.zip.Zipentry.getSize()`
YES	Flawed cryptographic implementation	Forgetting to call `Cipher.doFinal()` causes a Cipher object not to flush the bytes it is holding on to as the object tries to assemble a block for encrypted text. This will lead to truncated data in the final output.	Crypto	`javax.crypto.Cipher.doFinal()`
YES	Flawed cryptographic implementation	After calling `Cipher.update()`, an inappropriate selection of `Cipher.doFinal()` overloaded method (in this case, `Cipher.doFinal (byte[] input)` instead of `Cipher.doFinal(byte[] output, int outputOffset)`) creates an invalid ciphertext.	Crypto	`javax.crypto.Cipher.update()`
YES	Flawed cryptographic implementation	Failing to call `Cipher.getOutputSize()` does not guarantee the allocation of sufficient space for an output buffer, thus creating an invalid ciphertext.	Crypto	`javax.crypto.Cipher.getOutputSize()`
YES	Flawed cryptographic implementation	Failing to call `CipherOutputStream.close()` produces an invalid ciphertext, which cannot be decrypted into the original text.	Crypto	`javax.crypto.CipherOutputStream.close()`
YES	Improper input validation	Without proper input/argument sanitization, `Runtime.exec()` is vulnerable to command injection attacks.	String	`java.lang.Runtime.exec()`
YES	Improper input validation	Susceptible to inline command injection attacks without proper input sanitization.	String	`new java.lang.ProcessBuilder()`
YES	Improper input validation	Only using the `PreparedStatement` class cannot stop SQL injection attacks if string concatenation is used to build an SQL query.	String	`java.sql.PreparedStatement.setString()`
YES	Improper input validation	Inadequate input sanitization and validation allow malicious users to glean restricted information using the directory service.	String	`javax.naming.directory.DirContext.search()`
YES	Improper input validation	By using an evil regex, an attacker can make a program enter a prolonged unresponsive condition, thus enabling DoS attacks.	String	`java.util.regex.Matcher.matches()`
YES	Improper input validation	Without verifying sources, an attacker can make a program write false/unverified information into log files.	String	`java.util.logging.Logger.info()`
YES	Disclosure of sensitive information	Temporary file deletion by invoking `File.deleteOnExit()` occurs only in the case of a normal JVM shutdown, but not when the JVM crashes or is killed.	I/O	`java.io.File.deleteOnExit()`
YES	Disclosure of sensitive information	An implementation with `StandardOpenOption.DELETE_ON_CLOSE` may be unable to guarantee that it deletes the expected file when replaced by an attacker while the file is open. Consequently, sensitive data may be leaked.	I/O	`java.nio.file.Files.write()`
NO	N/A	N/A	I/O	`java.io.File.createNewFile()`
NO	N/A	N/A	I/O	`java.io.File.renameto()`
NO	N/A	N/A	I/O	`java.io.InputStream.read()`
NO	N/A	N/A	I/O	`java.util.zip.Zipentry.getSize()`
NO	N/A	N/A	I/O	`java.io.File.listFiles()`
NO	N/A	N/A	Crypto	`javax.crypto.Cipher.getOutputsize()`
NO	N/A	N/A	Crypto	`javax.crypto.CipherOutputStream.close()`
NO	N/A	N/A	String	`java.sql.PreparedStatement.setString()`

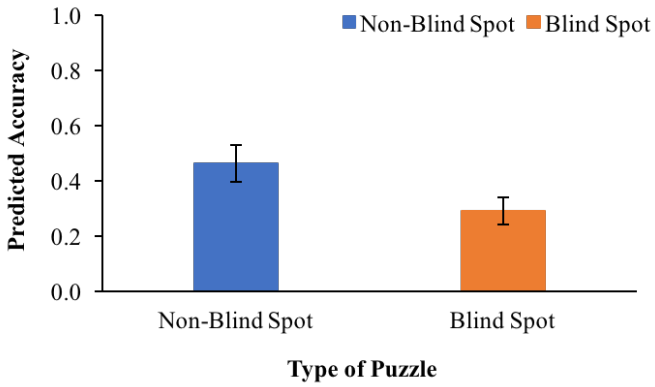


Figure 5: Developers were more likely to solve non-blindspot than blindspot puzzles. Error bars represent 95% confidence intervals.

In sum, we collected the following measures from the developers: (1) responses to puzzles; (2) developer-reported perceptions of puzzle difficulty, clarity, familiarity with puzzle functions, and confidence in solving the puzzle; (3) demographic information; (4) programming experience and skills, (5) personality traits, and (6) cognitive functioning scores.

Debriefing. All developers were debriefed at the end of the study about its true purpose and presented with the correct solutions for each puzzle they had worked on, including the rationale for the correct answer. The study ended by soliciting feedback about the study and processing compensation.

3. DATA ANALYSIS AND RESULTS

This section presents the results of the study and the findings that emerged from the data. We used the statistical software package STATA 14.0 for data analysis. As described in Section 1, the study goals were to (1) determine developers' ability to detect API blindspots in code and (2) examine the extent to which developer characteristics affected this capability. In particular, we tested the following hypotheses:

H1: Developers are less likely to correctly solve puzzles with API functions containing blindspots than puzzles with innocuous functions (non-blindspot puzzles).

H2:
- **a:** Developers perceive puzzles with API functions containing blindspots as more difficult than non-blindspot puzzles.
- **b:** Developers perceive puzzles with API functions containing blindspots as less clear than non-blindspot puzzles.
- **c:** Developers perceive puzzles with API functions containing blindspots as less familiar than non-blindspot puzzles.
- **d:** Developers are less confident about their puzzle solution when working on puzzles with API functions containing blindspots than non-blindspot puzzles.

H3: Higher cognitive functioning (reasoning, working memory, processing speed) in developers is associated with greater accuracy in solving puzzles with API functions containing blindspots.

H4: Higher levels of professional experience and expertise in developers are associated with greater accuracy in solving puzzles with API functions containing blindspots.

H5: Higher levels of conscientiousness and openness, and lower levels of neuroticism and agreeableness in developers are associated with greater accuracy in solving puzzles with API functions containing blindspots.

We used multilevel modeling to test H1 and H2a–d and ordinal logistic regression to test H3, H4, and H5 (see details below).

The main purpose of our analyses for all hypotheses was to determine the significance of specific effects (e.g., effect of a given personality trait on accuracy for blindspot puzzles), rather than identifying the best model to represent our data. Therefore, we did not apply a model comparison approach in our central analyses. In the exploratory analyses in Section 3.1, however, we were interested in determining the extent to which adding moderators (i.e., API usage type, cyclomatic complexity) enhanced the fit of our model, compared to the model originally tested under H1. In these instances, we report relevant goodness of fit indices (Akaike Information Criterion [AIC] and Bayesian Information Criteria [BIC] [8]).

Unless mentioned otherwise, we considered effects with p-values smaller than 0.05 as significant.

3.1 H1: Puzzle accuracy for blindspot vs. non-blindspot puzzles

We used multilevel logistic regression to test H1, accommodating for (1) the hierarchical data structure in which each set of six puzzles (level-1) was nested within each developer (level-2) and (2) the dichotomous outcome variable puzzle accuracy (1 = correct answer, 0 = incorrect answer). The independent variable was the presence of a blindspot (0 = no blindspot; 1 = blindspot). In this model, we also considered the random effect of the intercept to accommodate for inter-individual differences in overall puzzle accuracy. Presence of a blindspot had a significant effect on puzzle accuracy (*Wald* $\chi^2(2) = 20.60, p < .001$, Table 2), supporting H1 that developers were less likely to correctly solve puzzles with API functions containing blindspots than in those puzzles without blindspots.

In an exploratory fashion, we examined the extent to which (1) API usage type (i.e., I/O, Crypto, and String, see Section 2.3) and (2) puzzle cyclomatic complexity qualified the observed effect of the presence of blindspot on puzzle accuracy. The small number of puzzles in each set limited our capability to examine those two predictors in a single model. Therefore, we ran these exploratory analyses in two separate models, one for API usage type and the other for puzzle cyclomatic complexity. We used Wald tests to determine the significance of the main effects and interactions. To control for family-wise type-I error inflation due to multiple dependent models (i.e., models that share the same dependent variable), we applied Bonferroni correction for the threshold of the p-value to determine statistical significance in these exploratory analyses ($p < 0.025$).

API usage type. We added the categorical variable API usage type (1 = I/O, 2 = Crypto, 3 = String) and its interaction with the presence of blindspot as predictors in the model. Both the AIC and BIC were smaller for this model with the added moderator than for the H1 model (Table 2), suggesting a better goodness of fit when adding API usage as a moderator into the model. The main effect of presence of blindspot was not significant (*Wald* $\chi^2(1) = 0.91, p = 0.34$), but the main effect of API usage type

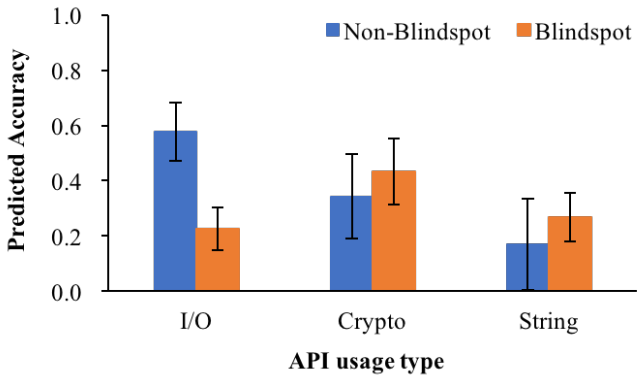


Figure 6: Interaction effect of API usage type and presence of blindspot on puzzle accuracy. The x-axis shows the three types of API usage: I/O, Crypto, and String. The y-axis shows predicted accuracy (predicted probability of correctly solving a puzzle). Error bars represent 95% confidence intervals after Bonferroni correction of the p-value.

(*Wald* $\chi^2(2) = 10.64, p = 0.005$) and its interaction with the presence of blindspot (*Wald* $\chi^2(2) = 24.81, p < 0.001$) was significant. As shown in Figure 6, accuracy was higher for non-blindspot puzzles than for blindspot ones with an API function that involved I/O. Accuracy was comparable in both non-blindspot and blindspot puzzles with API functions that involved the other two usage types (i.e., Crypto, String).

Cyclomatic complexity. We added the categorical variable cyclomatic complexity (1 = low, 2 = medium, 3 = high) and its interaction with the presence of blindspot as predictors in the model. Both the AIC and BIC were smaller for this model with the added moderator than for the H1 model (Table 2), suggesting a better goodness of fit when adding puzzle cyclomatic complexity as an additional predictor. The main effect of cyclomatic complexity was not significant (*Wald* $\chi^2(1) = 0.74, p < 0.69$), but the main effect of the presence of blindspot (*Wald* $\chi^2(1) = 23.95, p < 0.001$) and its interaction with cyclomatic complexity (*Wald* $\chi^2(2) = 30.1, p < 0.001$) was significant. As shown in Figure 7, accuracy was higher for non-blindspot than for blindspot puzzles at medium cyclomatic complexity, and, even more pronounced at high cyclomatic complexity. That is, the higher the cyclomatic complexity of the code in a puzzle containing blindspots, the less likely developers were to correctly solve the puzzle.

3.2 H2: Developers' perceptions for blindspot vs. non-blindspot puzzles

For H2a–d, we again used multilevel modeling to accommodate for the hierarchical data structure. The dependent variables for H2a–d were the four continuous rating dimensions (i.e., difficulty, clarity, familiarity, confidence), respectively, which we submitted to four separate multilevel regression models to examine the effect of the presence of blindspot on each of the four rating dimensions. In each model, we also considered the random effect of the intercept to accommodate for the inter-individual differences in the overall ratings for the respective dimension. As shown in Table 3, developers' perceptions did not differ as a function of the presence of blindspot in puzzles. Thus, the data did not support H2a–d.

3.3 H3–5: Cognitive function, technical expertise/experience, and personality traits

For H3, H4, and H5, the number of correctly solved blindspot puzzles across the four blindspot puzzles constituted the ordinal outcome variable blindspot puzzle accuracy, with a range from 0 to 4. Given this ordinal outcome variable, we conducted ordinal logistic regressions to test these hypotheses.

For various reasons (e.g., audio recording failure, incompatibility between the developers' browser version and our audio recording plugin and survey software), four cognitive measures from the BTACT (i.e., immediate recall, delay recall, verbal fluency, backward counting) had more than 25% data points missing. These four measures were therefore not analyzed. In addition, only 80 out of 109 developers had complete data on the Series and Digit-Back task from the BTACT, and the Oral Symbol Digit Task from the NIH toolbox. Given this missing data on the cognitive measures, which would have largely reduced the sample size, and thus power to detect significant effects, if collapsed across predictor variables (i.e., the three cognitive measures for H3 as well as the experience/expertise measures for H4 and the personality traits for H5), we conducted three separate models for H3, but tested H4 and H5 in one single model. For testing H4 and H5, three measures of professional expertise (Years of programming, Technical score, Java skills) and five personality traits (Agreeableness, Conscientiousness, Extraversion, Neuroticism, Openness) served as independent variables. As all four models (three for the cognitive measures and one for experience/expertise and personality) referred to the same dependent variable (i.e., blindspot puzzle accuracy), we applied Bonferroni correction on the threshold of the p-values ($p < 0.008$ for H3 and $p < 0.025$ for H4 and H5).

Cognitive Function. Our analyses pertaining to H3 resulted in no significant effects for any of the three cognitive measures on blindspot puzzle accuracy (all $ps > 0.10$, Table 4). Thus, the data did not support H3.

Technical Experience/Expertise. As shown in Table 5, none of the three predictors of experience/expertise predicted blindspot puzzle accuracy (all $ps > 0.10$). Thus, the data did not support H4.

Personality Traits. As shown in Table 5, the effect of openness on blindspot puzzle accuracy was significant ($p < 0.001$). That is, greater openness as a personality trait in developers was associated with greater accuracy in solving blindspot puzzles. None of the other personality dimensions showed significant effects (all $p > 0.09$).

4. DISCUSSION

This section summarizes the study findings, discusses study strengths and limitations, and offers actionable recommendations.

4.1 Summary of findings

The goal of this study was to examine API blindspots from the developers' perspective to: (1) determine the extent to which developers can detect API blindspots in code with the goal to improve understanding of the implication blindspots have on software security, and (2) determine the extent to which developer characteristics (i.e., difficulties with code, perceptions of code clarity, familiarity with code, confidence in solving puzzles, developers' level of cognitive functioning, their professional experience and expertise, and their personality traits) influenced developers' ability to detect blindspots. We also explored the extent to which API usage category and cyclomatic complexity of the puzzles impacted developers' ability to detect blindspots.

Table 2: Effect of presence of blindspot on puzzle accuracy (H1) and results of exploratory analyses on the moderation of API usage type and cyclomatic complexity on puzzle accuracy.

	Hypothesis 1		Expl. Anal. – API Usage Type		Expl. Anal. – Cyclomatic Complexity	
Fixed Effect	O.R. (SE)	95% CI	O.R. (SE)	95% CI	O.R. (SE)	95% CI
Presence of blindspot						
Blindspot	**0.44 (0.08)**	**[0.31, 0.63]**	**0.16 (0.05)**	**[0.09, 0.31]**	1.72 (0.55)	[0.92, 3.21]
API usage type						
Crypto			**0.33 (0.13)**	**[0.15, 0.71]**		
String			**0.11 (0.07)**	**[0.04, 0.37]**		
Presence of blindspot × API usage type						
Blindspot × *Crypto*			**9.10 (4.50)**	**[3.45, 23.98]**		
Blindspot × *String*			**11.35 (7.85)**	**[2.92, 44.04]**		
Cyclomatic complexity						
Medium					1.52 (0.68)	[0.64, 3.63]
High					**6.88 (2.62)**	**[3.26, 14.53]**
Presence of blindspot × Cyclomatic complexity						
Blindspot × *Medium*					**0.29 (0.15)**	**[0.10, 0.82]**
Blindspot × *High*					**0.02 (0.01)**	**[0.005, 0.08]**
Random Effect	σ^2 (SE)	95% CI	σ^2 (SE)	95% CI	σ^2 (SE)	95% CI
Intercept	**0.43 (0.20)**	**[0.17, 1.09]**	**0.72 (0.28)**	**[0.34, 1.52]**	**0.54 (0.25)**	**[0.22, 1.33]**
Goodness of Fit						
AIC	824.13		794.63		773.26	
BIC	837.58		826.01		804.64	

Note. O. R. = odds ratio; SE = standard error; CI = confidence interval. We used robust standard errors to accommodate for the hierarchical data structure. The reference category is *non-blindspot* for "presence of blindspot", *I/O* for "API usage type", and *low* for "cyclomatic complexity". Bonferroni correction was applied to p-values in the simple effect analyses for the main effect of API usage type and cyclomatic complexity and the follow-up analyses to counter inflation of type-I errors due to multiple comparison. **Bold** indicates significant effects at $p < .05$.

Table 3: Effect of presence of blindspot on developers' perception of puzzles.

	H2a: Difficulty		H2b: Clarity		H2c: Familiarity		H2d: Confidence	
Fixed Effect	B (SE)	95% CI	B (SE)	95% CI	B (SE)	95% CI	B (SE)	95% CI
Presence of Blindspot								
Blindspot	0.16 (0.14)	[-0.12, 0.43]	-0.01 (0.12)	[-0.25, 0.23]	-0.10 (0.15)	[-0.40, 0.19]	-0.11 (0.13)	[-0.36, 0.15]
Random Effect								
Intercept	**2.27 (0.33)**	**[1.31, 3.01]**	**2.22 (0.37)**	**[1.61, 3.07]**	**1.67 (0.32)**	**[1.15, 2.43]**	**1.72 (0.37)**	**[1.13, 2.60]**

Note. B = unstandardized regression coefficient; SE = standard error; CI = confidence interval. The reference category is *non-blindspot* for "presence of blindspot". **Bold** indicates significant effects at $p < .05$.

Table 4: Effect of developers' level of cognitive function on puzzle accuracy.

Cognitive	Blindspot Puzzles		Non-Blindspot Puzzles	
Function	**O.R. (SE)**	**95% CI**	**O.R. (SE)**	**95% CI**
Reasoning	1.16 (0.17)	[0.87, 1.54]	1.31 (0.21)	[0.96, 1.80]
Working Memory	1.12 (0.08)	[0.97, 1.28]	1.09 (0.11)	[0.90, 1.33]
Processing Speed	1.00 (0.01)	[0.99, 1.02]	1.01 (0.01)	[0.99, 1.03]

Note. O.R.= odds ratio; SE = standard error; CI = confidence interval. 91 developers were included in the analysis for reasoning, 90 for working memory, and 89 developers for processing speed.

Our results confirmed **H1** that developers are less likely to correctly solve puzzles with blindspots compared to puzzles without blindspots. This finding suggests that developers experience security blindspots while using certain API functions. Oliveira et al. [37] interviewed professional developers and found that they generally trust APIs. Given this general trust, even security-minded developers may not explicitly look for vulnerabilities in API functions, with the result that blindspots cause security vulnerabilities.

Our exploratory analyses suggested that the presence of blindspot particularly impacts accuracy in solving puzzles with I/O-related API functions, and with more complex programming scenarios (i.e., high cyclomatic complexity).

Our data did not support **H2a-H2d**, that posited developers' perceptions of puzzle difficulty, clarity, familiarity, and confidence are associated with their ability to detect blindspots. Our results also did not support **H3** that developers' level of cognitive functioning could predict their ability to detect blindspots.

We also found no support for **H4** that professional and technical experience were associated with developers' ability to detect blindspots. This finding is in line with research on code review that showed a developer's amount of experience does not correlate with greater accuracy or effectiveness in detecting security issues in code [16].

Our results partially support **H5** as more openness as a personality trait in developers does appear to be associated with a higher likelihood to detect blindspots. Openness relates to intellectual curiosity and the ability to use one's imagination [29]. It is plausible

Table 5: Effect of developers' professional expertise and personality traits on puzzle accuracy.

Factor	Blindspot Puzzles		Non-Blindspot Puzzles	
	O.R. (SE)	95% CI	O.R. (SE)	95% CI
Professional Expertise				
Years of programming	0.81 (0.70)	[0.15, 4.45]	3.47 (2.82)	[0.71, 17.06]
Technical expertise	0.93 (0.12)	[0.72, 1.19]	1.08 (0.12)	[0.87, 1.34]
Java skills	1.11 (0.15)	[0.85, 1.45]	0.96 (0.13)	[0.74, 1.24]
Personality Traits				
Agreeableness	0.95 (0.05)	[0.85, 1.05]	0.98 (0.04)	[0.90, 1.07]
Conscientiousness	0.97 (0.05)	[0.88, 1.07]	0.94 (0.05)	[0.85, 1.04]
Extraversion	0.94 (0.04)	[0.87, 1.01]	0.99 (0.04)	[0.91, 1.08]
Neuroticism	0.93 (0.04)	[0.86, 1.01]	**0.91 (0.04)**	**[0.83, 0.99]**
Openness	**1.18 (0.05)**	**[1.09, 1.29]**	1.08 (0.04)	[0.99, 1.17]

Note. O.R.= odds ratio; SE = standard error; CI = confidence interval. **Bold** indicates significant effects at $p < .05$.

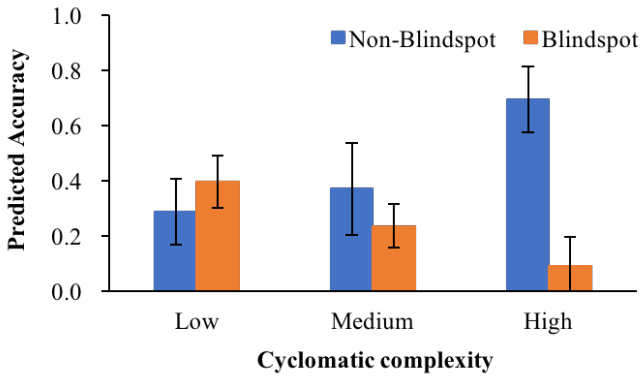


Figure 7: Interaction effect of presence of blindspot and cyclomatic complexity (CC) on puzzle accuracy. X-axis shows the three levels of CC: low ($\leq$ 2), medium (3 – 4) and high ($>$ 4). Y-axis shows predicted accuracy (predicted probability of correctly solving a puzzle). Error bars represent 95% confidence intervals after Bonferroni correction of the p-value.

that detection of security vulnerabilities benefits from a developer's ability and willingness to think of different scenarios and program inputs that might cause a piece of software to generate unexpected results. None of the other tested personality traits showed any significant effect. This finding is in line with previous research [23] that programming aptitude was not associated with agreeableness or neuroticism.

4.2 Strengths and limitations

Our work takes a novel approach by analyzing blindspots in API functions from the developers' perspective, thereby considering variables such as perception of code, level of cognitive function, experience, and personality. This interdisciplinary approach joints forces from computer science and psychology to understand how API blindspots cause security vulnerabilities.

A strength of our study was that it used a behavioral approach in addition to self-reporting by providing developers actual programming scenarios and assessing their ability to solve them. Our study also assessed performance-based cognitive functioning levels as possible predictors of puzzle accuracy.

Our sample was diverse, comprising 109 developers, made up of mostly professionals from different countries.For recruitment, we used snowball sampling [5], which meant participants could refer other developers. This word-of-mouth technique is often applied in research, particularly when targeting a specific group of individuals (i.e., developers). It was advantageous in allowing our team to reach developers we could not have otherwise found using our standard recruiting techniques (flyers, forums, social media groups, personal networks). However, it can also introduce bias by reducing random sampling and adding possible interdependence to the data. In our study, 41.3% of the participants chose the referral option with only 8.3% of the referred individuals enrolling in the study.

We conducted an a-priori power analysis to determine the appropriate sample size and number of puzzles needed considering our factorial design and with regard to our primary study aims. However, to counter possible fatigue effects, as suggested during the piloting phase of this research, we asked developers to only complete six puzzles. This resulted in a limited number of observations, thus not allowing a robust examination of some of the effects (i.e., API usage type, cyclomatic complexity). Therefore, we conducted exploratory analysis on these puzzle features to generate preliminary results, which we hope will spur future research. These preliminary results suggested that developers' detection of blindspots was particularly difficult for puzzles with I/O usage function and with high cyclomatic complexity. Increasing the number of puzzles each developer solves would, in future research, enhance the analytic power and allow a more comprehensive analysis of diverse puzzle subtypes. However, to avoid fatigue and attrition, future studies should focus on a few such categories at a time. For example, to examine the moderation effect of I/O functions on developers' ability to detect blindspots, I/O functions could be varied between puzzles, while keeping cyclomatic complexity and number of parameters consistent.

Because of compatibility issues between some developers' browser versions and our audio recording system software, we were not able to collect complete cognitive data for all participants. This missing data reduced the sample size in the analyses pertaining to the cognitive measures, thus reducing power to detect significant effects. Also, even though the cognitive tasks administered in the present study are widely used, they may not have been sensitive enough to differentiate between developers and/or may not have targeted cognitive processes that are particularly relevant for detection of blindspots in API functions.

4.3 Recommendations

Our results provide important insights for the software and API development community and corroborates aspects of related research in code review and developers' perceptions of code. Our data supports the notion that blindspots in API functions lead to the introduction of vulnerabilities in software, even when used by experienced developers. Given these findings, API designers should consider addressing developers' misconceptions and flawed assumptions when working with APIs to increase code security. For example, before release to the public, new or updated API functions should undergo pilot testing with developers not involved in the function's design and implementation. This pilot testing could be modeled after the approach used in our study. Furthermore, developer-centric testing should be conducted with existing APIs, so that misconceptions of specific categories of APIs can be better documented. In this context, given our preliminary findings regarding the more pronounced effect of blindspots for I/O-related API func-

tions, greater effort should be invested in improving the design and documentation of I/O-related functions, especially considering the high prevalence of I/O operations in today's software.

Our data did not provide support for the claim that developers' ability to detect blindspots could be associated with their perceptions of problem difficulty, code clarity, function familiarity, confidence in their ability to solve code, their experience, expertise, and cognitive functioning, or any tested personality traits, with the exception of openness. It could be assumed that a developer who is confident and familiar with the programming scenario and API functions at hand, who has many years of programming experience, especially with a particular programming language, is cognitively high functioning, and is self-disciplined (high conscientiousness), suspicious of situations in general (low agreeableness) and emotionally stable (low neuroticism), would be better in detecting security blindspots, and would, consequently, write more secure code. These assumptions were not supported by our data. Rather, our data suggests that cognitively high functioning, experienced, confident developers can still fall for security blindspots. Software security awareness education may be a useful approach to educate developers about these risks. Such educational approaches could train developers not to rely on beliefs and gut feelings when using API functions. Increased risk awareness could lead to developers asking themselves more questions about how API function usage may result in unexpected outcomes, and could motivate them to rely more on diagnostic tools.

In large software development companies, it has become common to assign different teams to work on the various aspects of code. For example, within Google [42, 24], three distinct groups may work on functionality, security, and privacy aspects of the software separately. Such a diversified approach has the potential to minimize the introduction of vulnerabilities in code because there will be a group of developers whose primary task would be to identify how an adversary can exploit source code and cause security and privacy breaches. However, not many companies can afford to hire developers to address security alone. The common rationale is that all developers should create secure functionality. However, as discussed in Section 1, and supported by our data, this mindset maybe misleading. Both of these tasks are cognitively demanding and thus, one team to address both might be a zero-sum game.

Another practice often applied in companies is to hire an expert who is highly familiar with security vulnerabilities and has good knowledge of programming languages to decrease the chance of code vulnerabilities. Our results suggest that this rationale might also be misleading, in that even highly experienced, cognitively high functioning developers experience difficulties in detecting security blindspots in API functions.

Taken together, our study findings are applicable in the following areas: (1) design, implementation, and evaluation of new APIs; (2) addressing of blindspots in legacy APIs; (3) development of novel methods for developer recruitment/training based on personality assessment; and (4) improvement of software development processes in organizations (e.g., establishment of separate security vs. functionality teams).

5. RELATED WORK

Our work intersects the areas of API usability, programming language design, and developers' practices and perceptions of security. In this section we provide a discussion of related work, and position our work with respect to these earlier initiatives.

5.1 API usability

Our work falls into the still young, but growing topic of API usability, which focuses on how to design APIs in a manner that reduces the likelihood of developer errors that can create software vulnerabilities. A recent article presents an overview of this field [34]. For example, Ellis et al. [17] showed that, despite its popularity, the factory design pattern [20] was detrimental to API usability because when incorporated into an API it was difficult to use.

Most studies of API usability have focused on non-security considerations, such as examining how well programmers can use the functionality that an API intends to provide. Our work is, thus, a significant departure from this research direction, although it shares many of the same methodologies.

Two of the few existing studies on security-related API usability were conducted by Coblenz et al. [10, 11] and by Weber et al. [61]. Stylos and Clarke [55] had concluded that the immutability feature of a programming language (i.e., complete restriction on an object to change its state once it is created) was detrimental to API usability. Since this perspective contradicted the standard security guidance (*"Mutability, whilst appearing innocuous, can cause a surprising variety of security problems"* [48, 32]), Coblenz et al. investigated the impact of immutability on API usability and security. From a series of empirical studies, they concluded that immutability had positive effects on both security and usability [11]. Based on these findings they designed and implemented a Java language extension to realize these benefits [10].

Recent work has investigated the usability of cryptographic APIs. Nadi et al. [35] identified challenges developers face when using Java Crypto APIs, namely poor documentation, lack of cryptography knowledge by the developers, and poor API design. Acar et al. [1] conducted an online study with open source Python developers about the usability of the Python Crypto API. In this study, developers reported the need for simpler interfaces and easier-to-consult documentation with secure, easy-to-use code examples.

In contrast to previous work, our study focused on understanding blindspots that developers experience while working with general classes of API functions.

5.2 Programming language design

Usability in programming language design has been a long-standing concern. Initially, most of the related literature was non-empirical, but empirical studies of programming language design have become more popular. For example, Stefik and Siebert [54] showed that syntax used in a programming language was a significant barrier for novices. Our work has the potential to contribute to programming language design, since our focus is on understanding security blindspots in API function usage, and the function traits that exacerbate the problem.

5.3 Developer practices and perceptions of security and privacy

Balebako et al. discussed the relationship between the security and privacy mindsets of mobile app developers and company characteristics (e.g., company size, having a Chief Privacy Officer, etc.). They found that developers tend to prioritize security tools over privacy policies, mostly because of the language of privacy policies is so obscure [7].

Xie et al. [66] conducted interviews with professional developers to understand secure coding practices. They reported a disconnect between developers' conceptual understanding of security and their

attitudes regarding personal responsibility and practices for software security. Developers also often hold a "not-my-problem" attitude when it comes to securing the software they are developing; that is, they appear to rely on other processes, people, or organizations to handle software security.

Witschey et al. [63] conducted a survey with professional developers to understand factors contributing to the adoption of security tools. They found that peer effects and the frequency of interaction with security experts were more important than security education, office policy, easy-to-use tools, personal inquisitiveness, and better job performance to promote security tool adoption.

Acar et al. [4] and Green and Smith [27] suggest a research agenda to achieve usable security for developers. They proposed several research questions to elicit developers' attitudes, needs, and priorities in the area of security. Oltrogge et al. [38] asked for developers' feedback on TLS certificate pinning strategy in non-browser based mobile applications. They found a wide conceptual gap about pinning and its proper implementation in software due to API complexity.

A survey conducted by Acar et al. [2] with 295 app developers concluded that developers learned security through web search and peers. The authors also conducted an experiment with over 50 Android developers to evaluate the effectiveness of different strategies to learn about app security. Programmers who used digital books achieved better security than those who used web searches. Recent research corroborates this finding by showing that the use of code-snippets from online developer forums (e.g., Stack Overflow) can lead to software vulnerabilities [3, 18, 59].

Recent studies have investigated the need and type of interventions required for developers to adopt secure software development practices. Xie et al. [65] found that developers needed to be motivated to fix software bugs. There has also been some work on how to create this motivation and encourage use of security tools. Several surveys identified the importance of social proof for developers' adoption of security tools [33, 62, 64].

Research on the effects of external software security consultancy suggests [43] that a single time-limited involvement of developers with security awareness programs is generally ineffective in the long-term. Poller et al. [44] explored the effect of organizational practices and priorities on the adoption of developers' secure programming. They found that security vulnerability patching is done as a stand-alone procedure, rather than being part of product feature development. In an interview-based study by Votipka et al. [60] with a group of 25 white-hat hackers and software testers on bug finding related issues, hackers were more adept and efficient in finding software vulnerabilities than testers, but they had more difficulty in communicating such issues to developers because of a lack of shared vocabulary.

In a position paper, Cappos et al. [9] proposed that software vulnerabilities are a blindspot in developers' heuristic-based decision making mental models. Oliveira et al. [37] further showed that security is not a priority in the developers' mindsets while coding. They found, however, that developers did adopt a security mindset once primed about the topic.

Our work complements and extends previous investigations on the effect of API blindspots on writing secure code, and in determining the extent to which developers' characteristics (perceptions, expertise/experience, cognitive function, and personality) influence such capabilities.

6. CONCLUSIONS

In this paper, we report the results of an empirical study on understanding blindspots in API functions from the perspective of the developer. We evaluated developers' ability to perceive blindspots in a variety of code scenarios and examined how personal characteristics, such as perceptions of the correctness of their answers, familiarity with the code, years of professional experience, level of cognitive functioning, and personality, affected this capability. We also explored the influence of programming scenario characteristics (API usage type, cyclomatic complexity) on developers' performance in detecting blindspots.

Our study asked 109 developers to work on a set of six naturalistic programming scenarios (puzzles), comprising four puzzles with blindspots and two without blindspots. Developers were not informed about the security focus of this investigation. Our results showed that: (1) developers were less likely to correctly solve puzzles with blindspots than puzzles without blindspots, with this effect more pronounced for I/O API functions and complex code scenarios; (2) developers' level of cognitive functioning and (3) their expertise and experience did not predict their ability to detect blindspots; however, (4) those who exhibited more openness as a personality trait did show a greater ability to detect blindspots.

Our findings have the potential to inform the design of more secure APIs. Our data suggests that API design, implementation, and testing should take into account the potential security blindspots developers may have, particularly when using I/O functions. Further, our findings that experience and cognition may not predict developers' ability to detect blindspots, suggest that the emerging practice of establishing separate functionality vs. security teams in a given project may be a promising strategy to improve software security. This strategy may also constitute a more cost-effective paradigm for secure software development than solely relying on one group of experts, expected to simultaneously address both functionality and security.

7. ACKNOWLEDGEMENTS

We thank our shepherd Michael Reiter for guidance in writing the final version of the paper and the SOUPS 2018 anonymous reviewers for valuable feedback. We thank Sam Weber and Yanyan Zhuang for discussions related to our work. This work was supported by the National Science Foundation under grants no. CNS-1513055, CNS-1513457, and CNS-1513572.

8. REFERENCES

[1] Y. Acar, M. Backes, S. Fahl, S. Garfinkel, D. Kim, M. L. Mazurek, and C. Stransky. Comparing the Usability of Cryptographic APIs. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 154–171, May 2017.

[2] Y. Acar, M. Backes, S. Fahl, D. Kim, M. L. Mazurek, and C. 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, May 2016.

[3] Y. Acar, M. Backes, S. Fahl, D. Kim, M. L. Mazurek, and C. Stransky. How Internet Resources Might Be Helping You Develop Faster but Less Securely. *IEEE Security Privacy*, 15(2):50–60, March 2017.

[4] Y. Acar, S. Fahl, and M. L. Mazurek. You are Not Your Developer, Either: A Research Agenda for Usable Security and Privacy Research Beyond End Users. In *2016 IEEE Cybersecurity Development (SecDev)*, pages 3–8, Nov 2016.

[5] R. Atkinson and J. Flint. Accessing Hidden and Hard-to-Reach Populations: Snowball Research Strategies. *Social research update*, 33(1):1–4, 2001.

[6] J. Bailey and R. B. Mitchell. Industry Perceptions of the Competencies Needed by Computer Programmers: Technical, Business, and Soft Skills. *Journal of Computer Information Systems*, 47(2):28–33, 2006.

[7] R. Balebako, A. Marsh, J. Lin, J. I. Hong, and L. F. Cranor. The Privacy and Security Behaviors of Smartphone App Developers. In *Proceedings of 2014 Workshop on Usable Security*. Internet Society, 2014.

[8] K. P. Burnham and D. R. Anderson. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological methods & research*, 33(2):261–304, 2004.

[9] J. Cappos, Y. Zhuang, D. Oliveira, M. Rosenthal, and K.-C. Yeh. Vulnerabilities As Blind Spots in Developer's Heuristic-Based Decision-Making Processes. In *Proceedings of the 2014 New Security Paradigms Workshop*, NSPW '14, pages 53–62, New York, NY, USA, 2014. ACM.

[10] M. Coblenz, W. Nelson, J. Aldrich, B. Myers, and J. Sunshine. Glacier: Transitive Class Immutability for Java. In *Proceedings of the 39th International Conference on Software Engineering*, ICSE '17, pages 496–506, Piscataway, NJ, USA, 2017. IEEE Press.

[11] M. Coblenz, J. Sunshine, J. Aldrich, B. Myers, S. Weber, and F. Shull. Exploring Language Support for Immutability. In *Proceedings of the 38th International Conference on Software Engineering*, ICSE '16, pages 736–747, New York, NY, USA, 2016. ACM.

[12] Common Weakness Enumeration (CWE)/SANS Top 25 Most Dangerous Software Errors, 2011. Available at `http://cwe.mitre.org/top25/`.

[13] P. T. Costa and R. R. MacCrae. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual*. Psychological Assessment Resources, Incorporated, 1992.

[14] B. Dagenais and M. P. Robillard. Creating and Evolving Developer Documentation: Understanding the Decisions of Open Source Contributors. In *Proceedings of the Eighteenth ACM SIGSOFT International Symposium on Foundations of Software Engineering*, FSE '10, pages 127–136, New York, NY, USA, 2010. ACM.

[15] N. Diakopoulos and S. Cass. The Top Programming Languages 2016, Jul 2016. Available at `https://spectrum.ieee.org/static/interactive-the-top-programming-languages-2016`.

[16] A. Edmundson, B. Holtkamp, E. Rivera, M. Finifter, A. Mettler, and D. Wagner. An Empirical Study on the Effectiveness of Security Code Review. In *Proceedings of the 5th International Conference on Engineering Secure Software and Systems*, ESSoS'13, pages 197–212, Berlin, Heidelberg, 2013. Springer-Verlag.

[17] B. Ellis, J. Stylos, and B. Myers. The Factory Pattern in API Design: A Usability Evaluation. In *Proceedings of the 29th International Conference on Software Engineering*, ICSE '07, pages 302–312, Washington, DC, USA, 2007. IEEE Computer Society.

[18] F. Fischer, K. Böttinger, H. Xiao, C. Stransky, Y. Acar, M. Backes, and S. Fahl. Stack Overflow Considered Harmful? The Impact of Copy Paste on Android Application Security. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 121–136, May 2017.

[19] A Taxonomy of Coding Errors that Affect Security. Available at `https://vulncat.hpefod.com/en`.

[20] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Patterns: Elements of Reusable Object-oriented Software*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1995.

[21] R. C. Gershon, M. V. Wagster, H. C. Hendrie, N. A. Fox, K. F. Cook, and C. J. Nowinski. NIH Toolbox for Assessment of Neurological and Behavioral Function. *Neurology*, 80(11 Supplement 3):S2–S6, 2013.

[22] GitHut : Discover Languages in Github, 2014. Available at `http://githut.info/`.

[23] T. Gnambs. What Makes a Computer Wiz? Linking Personality Traits and Programming Aptitude. *Journal of Research in Personality*, 58:31 – 34, 2015.

[24] Google's Approach to IT Security: A Google White Paper, 2016. Available at `https://static.googleusercontent.com/media/1.9.22.221/en//enterprise/pdf/whygoogle/google-common-security-whitepaper.pdf`.

[25] D. Gopstein, J. Iannacone, Y. Yan, L. DeLong, Y. Zhuang, M. K.-C. Yeh, and J. Cappos. Understanding Misunderstandings in Source Code. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, ESEC/FSE 2017, pages 129–139, New York, NY, USA, 2017. ACM.

[26] GraphicsMagick 1.4 Heap-Based Buffer Overflow Vulnerability, Dec. 2017. Available at `https://nvd.nist.gov/vuln/detail/CVE-2017-17915`.

[27] M. Green and M. Smith. Developers are Not the Enemy!: The Need for Usable Security APIs. *IEEE Security Privacy*, 14(5):40–46, Sept 2016.

[28] Java Platform SE 8 Documentation. Available at `https://docs.oracle.com/javase/8/docs/`.

[29] O. P. John and S. Srivastava. The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999.

[30] P. J. Kovacs and G. A. Davis. Determining Critical Skills and Knowledge Requirements of It Professionals by Analysing Keywords in Job Posting. In *48th Annual IACIS International Conference*. IACIS, 2008.

[31] T. J. McCabe. A Complexity Measure. *IEEE Trans. Softw. Eng.*, 2(4):308–320, July 1976.

[32] J. McManus and S. Shrum. SEI CERT Oracle Coding Standard for Java. Available at `https://wiki.sei.cmu.edu/confluence/display/java/Java Coding Guidelines`.

[33] E. Murphy-Hill, D. Y. Lee, G. C. Murphy, and J. McGrenere. How Do Users Discover New Tools in Software Development and Beyond? *Computer Supported Cooperative Work (CSCW)*, 24(5):389–422, Oct 2015.

[34] B. A. Myers and J. Stylos. Improving API Usability. *Commun. ACM*, 59(6):62–69, May 2016.

[35] S. Nadi, S. Krüger, M. Mezini, and E. Bodden. Jumping Through Hoops: Why Do Java Developers Struggle with Cryptography APIs? In *Proceedings of the 38th International Conference on Software Engineering*, ICSE '16, pages 935–946, New York, NY, USA, 2016. ACM.

[36] National Vulnerability Database. Available at `https://nvd.nist.gov/`.

[37] D. Oliveira, M. Rosenthal, N. Morin, K.-C. Yeh, J. Cappos, and Y. Zhuang. It's the Psychology Stupid: How Heuristics Explain Software Vulnerabilities and How Priming Can Illuminate Developer's Blind Spots. In *Proceedings of the 30th Annual Computer Security Applications Conference*, ACSAC '14, pages 296–305, New York, NY, USA, 2014. ACM.

[38] M. Oltrogge, Y. Acar, S. Dechand, M. Smith, and S. Fahl. To Pin or Not to Pin Helping App Developers Bullet Proof Their TLS Connections. In *Proceedings of the 24th USENIX Conference on Security Symposium*, SEC'15, pages 239–254, Berkeley, CA, USA, 2015. USENIX Association.

[39] The Open Web Application Security Project (OWASP) Top 10 Most Critical Web Application Security Risks, 2013. Available at `https://www.owasp.org/images/f/f8/OWASP_Top_10_-_2013.pdf`.

[40] OWASP Secure Coding Practices Checklist, 2016. Available at `https://www.owasp.org/index.php/OWASP_Secure_Coding_Practices_Checklist`.

[41] H. Orman. The Morris Worm: a Fifteen-year Perspective. *IEEE Security Privacy*, 1(5):35–43, Sept 2003.

[42] Personal communication with a Google project team leader.

[43] A. Poller, L. Kocksch, K. Kinder-Kurlanda, and F. A. Epp. First-time Security Audits As a Turning Point?: Challenges for Security Practices in an Industry Software Development Team. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 1288–1294, New York, NY, USA, 2016. ACM.

[44] A. Poller, L. Kocksch, S. Türpe, F. A. Epp, and K. Kinder-Kurlanda. Can Security Become a Routine?: A Study of Organizational Change in an Agile Software Development Group. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 2489–2503, New York, NY, USA, 2017. ACM.

[45] M. S. Rahman. An Empirical Case Study on Stack Overflow to Explore Developers' Security Challenges. Masters Report, Available at `http://krex.k-state.edu/dspace/handle/2097/34563`, 2016.

[46] M. P. Robillard. What Makes APIs Hard to Learn? Answers from Developers. *IEEE Software*, 26(6):27–34, Nov 2009.

[47] M. P. Robillard and R. Deline. A Field Study of API Learning Obstacles. *Empirical Softw. Engg.*, 16(6):703–732, Dec. 2011.

[48] Secure Coding Guidelines for Java SE. Available at `http://www.oracle.com/technetwork/java/seccodeguide-139067.html#6`.

[49] Security Focus Vulnerability Database. Available at `https://www.securityfocus.com/`.

[50] Security Vulnerabilities Published In 2017 (SQL Injection), 2017. Available at `https://www.cvedetails.com/vulnerability-list/opsqli-1/sql-injection.html`.

[51] Stack Overflow: A Q/A Site for Professional and Enthusiast Programmers. Available at `https://www.stackoverflow.com/`.

[52] Stack Overflow Developer Survey, 2016. Available at `https://insights.stackoverflow.com/survey/2016`.

[53] State of Software Security, 2016. Available at `https://www.veracode.com/sites/default/files/Resources/Reports/state-of-software-security-volume-7-veracode-report.pdf`.

[54] A. Stefik and S. Siebert. An Empirical Investigation into Programming Language Syntax. *Trans. Comput. Educ.*, 13(4):19:1–19:40, Nov. 2013.

[55] J. Stylos and S. Clarke. Usability Implications of Requiring Parameters in Objects' Constructors. In *Proceedings of the 29th International Conference on Software Engineering*, ICSE '07, pages 529–539, Washington, DC, USA, 2007. IEEE Computer Society.

[56] Symantec Internet Security Threat Report, 2017. Available at `https://www.symantec.com/content/dam/symantec/docs/reports/istr-22-2017-en.pdf`.

[57] K. Tsipenyuk, B. Chess, and G. McGraw. Seven Pernicious Kingdoms: a Taxonomy of Software Security Errors. *IEEE Security Privacy*, 3(6):81–84, Nov 2005.

[58] P. A. Tun and M. E. Lachman. Telephone Assessment of Cognitive Function in Adulthood: the Brief Test of Adult Cognition by Telephone. *Age and Ageing*, 35(6):629–632, 2006.

[59] T. Unruh, B. Shastry, M. Skoruppa, F. Maggi, K. Rieck, J.-P. Seifert, and F. Yamaguchi. Leveraging Flawed Tutorials for Seeding Large-Scale Web Vulnerability Discovery. In *11th USENIX Workshop on Offensive Technologies (WOOT 17)*, Vancouver, BC, 2017. USENIX Association.

[60] D. Votipka, R. Stevens, E. Redmiles, J. Hu, and M. Mazurek. Hackers vs. Testers: A Comparison of Software Vulnerability Discovery Processes. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 134–151, 2018.

[61] S. Weber, M. Coblenz, B. Myers, J. Aldrich, and J. Sunshine. Empirical Studies on the Security and Usability Impact of Immutability. In *2017 IEEE Cybersecurity Development (SecDev)*, pages 50–53, Sept 2017.

[62] J. Witschey, S. Xiao, and E. Murphy-Hill. Technical and Personal Factors Influencing Developers' Adoption of Security Tools. In *Proceedings of the 2014 ACM Workshop on Security Information Workers*, SIW '14, pages 23–26, New York, NY, USA, 2014. ACM.

[63] J. Witschey, O. Zielinska, A. Welk, E. Murphy-Hill, C. Mayhorn, and T. Zimmermann. Quantifying Developers' Adoption of Security Tools. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*, ESEC/FSE 2015, pages 260–271, New York, NY, USA, 2015. ACM.

[64] S. Xiao, J. Witschey, and E. Murphy-Hill. Social Influences on Secure Development Tool Adoption: Why Security Tools Spread. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 1095–1106, New York, NY, USA, 2014. ACM.

[65] J. Xie, H. Lipford, and B.-T. Chu. Evaluating Interactive Support for Secure Programming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2707–2716, New York, NY, USA, 2012. ACM.

[66] J. Xie, H. R. Lipford, and B. Chu. Why do Programmers Make Security Errors? In *Visual Languages and Human-Centric Computing (VL/HCC), 2011 IEEE Symposium on*, pages 161–164. IEEE, 2011.

"If I press delete, it's gone" - User Understanding of Online Data Deletion and Expiration

Ambar Murillo, Andreas Kramm, Sebastian Schnorf, Alexander De Luca
Google
{ambarm, akramm, sebschnorf, adeluca}@google.com

ABSTRACT

In this paper, we present the results of an interview study with 22 participants and two focus groups with 7 data deletion experts. The studies explored understanding of online data deletion and retention, as well as expiration of user data. We used different scenarios to shed light on what parts of the deletion process users understand and what they struggle with. As one of our results, we identified two major views on how online data deletion works: UI-Based and Backend-Aware (further divided into levels of detail). Their main difference is on whether users think beyond the user interface or not. The results indicate that communicating deletion based on components such as servers or "the cloud" has potential. Furthermore, generic expiration periods do not seem to work while controllable expiration periods are preferred.

1. INTRODUCTION

With growing storage capabilities and the large amounts of data[1] that people store online, data deletion is a common practice for internet users these days [12]. Reasons for deletion are manifold and range from simple things such as cleaning up your account to more critical tasks like getting data out of the reach of others, i.e. privacy [12].

We know that incomplete understanding of online data deletion can cause problems such as mishandling personal data due to misinterpretation of the process [12]. Ultimately, this can lead to issues with maintaining user privacy. Despite this importance, understanding online data deletion practices from a user perspective is still an understudied topic that deserves more attention. It is important to study what users actually know, need, and want when it comes to online data deletion.

To fill this gap, we conducted a user study with 22 participants with varying demographic backgrounds. In addition, we ran two focus groups with 7 data deletion experts. The main focus of this workstream was on *deletion*, *retention*, and *expiration*. In this work, we define *deletion* as the process of a user-invoked event to remove user generated content from an account. *Retention* refers to how long it takes until data is removed from all entities after it has been deleted. Finally, with *expiration*, we explored if it makes sense to have certain data automatically disappear after a certain period of time (think for instance about Snapchat messages that disappear after a user-defined timeframe).

In this paper, we provide insights into users' understanding of online data deletion, retention, and expiration. Our results can help with designing and communicating deletion in a way that is graspable for users, and as such, help the community to create better user interfaces and user education for online data deletion. For example, we identified two major views on online deletion: one solely based on the user interface and the second about what is going on in the background (with different levels of detail). We also found that data expiration does not follow a chronological order but is rather context-dependent. This means that data that is considered worthless at a certain point in time can become useful again later due to certain events.

[1]Please note that in this work, we focused on user-generated content as opposed to automatically generated data such as different types of metadata (e.g., log data).

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

2. RELATED WORK

For a long time, humans' ability to remember relied on biological memory and media with limited storage and sharing capacities. Most things were forgotten, and only few were remembered [9]. Even most acts violating social norms were forgotten after some time [4]. However, modern technology, and especially the internet, provides us with new abilities to overcome forgetting. Data can easily be stored, distributed, searched and used. Despite its benefits, this presents new challenges, especially with respect to an individual's privacy. For example, in 2006, a student teacher posted a picture of herself in a pirate costume with the caption "Drunken Pirate" on MySpace. Based on this picture, she was later denied her teaching degree [13].

A lot of research work in the past years has focused on helping people to protect their privacy while still being able to live a digital life. Not surprisingly, much of this work is centered around the content of online social networks, and more precisely, deletion and permanence of this content. For example, Wang et al. [17] showed that regret is a major factor for deletion in Facebook. Similar results were found in research on regrets on Twitter [14]. Interestingly, despite regret, a large scale study on deletion on Twitter [1] found that the majority of deletion cases are rather for corrections/edits. They also showed that content on public social networks like Twitter might not really be gone after deletion due to replies, comments, and internet archives storing them. For example, the meaning of a deleted tweet can, in many cases, be recreated based on replies and mentions. To mitigate this issue, Wang et al. proposed a system to support social network users to post fewer regrettable posts by providing them hints on who will be able to see their posts [16].

In a field trial, this system indeed significantly reduced sharing of potentially regrettable content.

Another important dimension of online data privacy is permanence. Much online content is designed to remain available until it is actively removed by the user, raising questions of how data sharing preferences might change longitudinally. Ayalon and Toch [2] looked at sharing preferences of Facebook content over time, finding a meaningful decrease in willingness to share content as it ages. User behavior, however, did not directly align with these stated preferences, since users did not tend to delete old posts to the same degree that their sharing preferences would have implied. The authors suggest expiration controls as a method to manage longitudinal privacy, with users setting expiration dates for content as they post it. This assumes that people will be able to predict their sharing preferences for content with some degree of certainty. However, past research [3], has found that participants were not particularly good at predicting their privacy preferences over time, therefore, raising questions as to whether setting an expiration date for content as it is created would be in line with users' evolving privacy needs. Bauer et al. [3] also found that participants wanted constant access to posts over time, even if only for reminiscing purposes. Posts associated with changing privacy preferences seemed to be the exception.

Considering that users might not accurately predict their future privacy preferences for their data, Mondal et al. [10] suggested an alternative online data privacy preserving mechanism for older data that moves away from time-based deletion. The authors suggest that, after a given period of inactivity, the user could receive suggestions to remove online content (e.g., Twitter posts).

This past research suggests the use of deletion mechanisms, in the form of expiration, to help users manage their online data privacy. Although users might not be very accurate in their predictions for desired expiration dates for their data, they are quite familiar with the use of deletion as a privacy preserving mechanism. A recent study on deletion practices in cloud storage [12] showed that one of the main motivators for deleting data in the cloud is privacy. Moreover, the paper showed that many problems that came with deletion are grounded in incomplete "mental models". Other research has also established the connection between "mental models" and their impact on user behavior. Wash [18] has also researched user "folk models" about security, finding that users relied on their models to guide their choice of security software, what expert security advice to follow, and how to justify ignoring certain advice. Other "mental models of security and privacy" research has found that knowledge of "mental models" can also be used as a foundation to create better user communication [5].

Most research has focused on the how and why of users' decision making process about deletion. That is, there is only little data on how users see deletion and whether this has consequences for their privacy. In addition, misunderstandings and unfounded expectations of deletion are grounded in incomplete understanding of the deletion process. With this work we provide first insights to fill this gap. This foundational research can then help us better design user education and implementation of data protection regulations.

3. STUDY

To uncover users' understanding of online data deletion and their expectations, we conducted semi-structured interviews in combination with a think-aloud drawing task. In addition, we conducted two expert focus groups to set a baseline to compare the interview study results against.

3.1 Interview Study

We conducted interviews in combinations with drawing tasks. Drawing tasks are a useful tool to uncover participants' understanding [8, 18]. They are particularly appropriate when researching underlying understandings which are hard to verbalize, as can be the case with abstract concepts where participants might lack technical vocabulary [11]. Additionally, drawing tasks are particularly well suited to generate reflective feedback as opposed to reactive feedback [15]. Drawing tasks are usually combined with the think-aloud protocol [7], meaning that participants verbalize what they are thinking as they are drawing, giving the observing researchers further insights into the meaning behind their drawings.

Each interview consisted of three main parts: *General Deletion*, *Deletion Scenarios*, and *Expiration*.

General Deletion - In this part, we explored the participants' online data use and understanding of deletion on a general level. For example, we asked them what online services they use that store data. At the end of this part, participants were asked to draw how they think online data deletion works in general (without a specific use case). As mentioned before, this included a think-aloud task.

Deletion Scenarios - This part explored two deletion scenarios: *Email* and *Social Media*[2]. We picked these two scenarios because they a) are very common, b) come with common deletion tasks, and c) are significantly distinct in how data is shown to users and how deletion works. This includes potential consequences such as the fact that social media data might still retain or be recoverable (literally or by meaning) after deletion due to shares, comments, archives etc. [1]. The two scenarios were counterbalanced, to mitigate learning effects.

Since there are plenty of different email and social media services, which could influence the results, we recruited for the following: We made sure that all participants used the online user interfaces of their respective email provider. All participants used either GMX, Web.de (the two most dominant email providers on the German market), or Gmail, or a combination of those. For social media, all participants were knowledgeable of Facebook (and referred to Facebook in their examples). Please be aware that this limits generalizability.

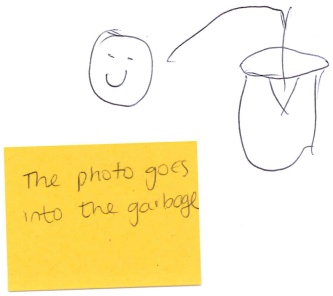
For each scenario, we asked the same questions, including why and when participants delete data on the respective platform. Similar to the general questions part, we also asked participants to create a drawing about how deletion works in each respective scenario, again, applying the think-aloud methodology. For details on the scenarios script, see Appendix A.

Three resulting drawings can be found in Figure 1.

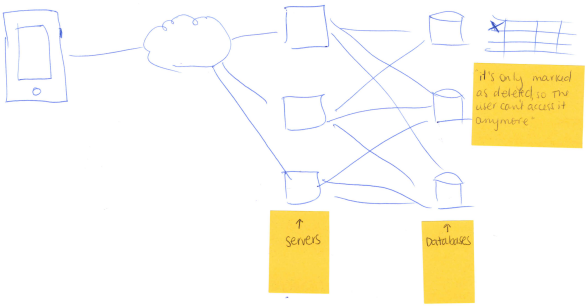
Expiration - The final part was about online data expiration. Here, we wanted to explore if and under what circumstances, participants thought specific data could or should be automatically deleted. We used four scenarios: Online shopping (data: address), email (data: email), social media (data: post/tweet), and search (data: search history). Online shopping and search were added in addition to the deletion scenarios to provide a wide spectrum of potential data. In addition, active deletion (as opposed to expiration) is rather rare in those two scenarios.

The main tool we used in this section was the graph shown in Figure 2. On the x-axis, participants were asked to add events which

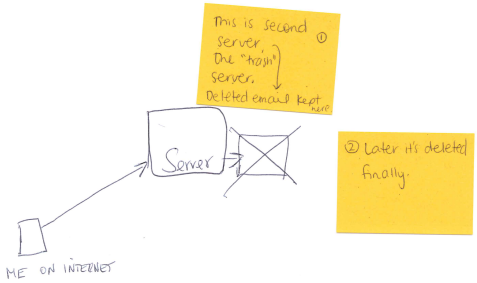
[2]All participants were recruited to be active email and social media users.



(a)



(b)



(c)

Figure 1: Participant diagrams explaining how deletion works: (a) in general, with no given scenario (Participant 6); (b) in an Email scenario (Participant 11); (c) in a Social Media scenario (Participant 8). Yellow notes were added by one of the researchers to clarify the diagrams.

would influence the usefulness of the data. On the y-axis, we asked them to add the respective usefulness rating (for them as users of the service). The question was: "*How useful it is for you that the service provider has this data?*". In the end, they were also asked if there would be an event, at which the data completely loses its usefulness (for details, see Appendix B).

3.1.1 Pilot Study

To verify and improve the study instrument, we ran an internal pilot study. To avoid or mitigate technical bias in the pilot study sessions, we recruited for co-workers from non-tech divisions of our company.

Figure 2: Graph used in the Expiration portion of the Interviews. Participants were given a blank graph (this one is from the Online Shopping scenario), and asked to add events that would influence the usefulness of the data on the x axis and the respective usefulness (for them) on the y axis.

In addition to simple wording improvements, the pilot study helped us to identify more significant changes: For instance, we used the results to identify appropriate scenarios for the expiration and deletion tasks, meaning the tasks covered a wide spectrum both in terms of how the service works and in how it is received by participants. The biggest change after the pilot was in the expiration graph which turned out to be much easier to understand with a time component involved in it as this seemed closer to how users perceive expiration.

3.1.2 Procedure

All interviews were conducted in-person at our premises. At the beginning of each session, participants were introduced to the study. First, they were asked to read and sign a consent form and NDA (was sent to all participants before the study so they had the chance to familiarize themselves with it). After this, the procedure was explained to them and the interviewer told them that they were free to stop the interview at any time or skip questions/parts they did not feel comfortable with (this option was not used by any participant). We also asked them for permission to make a video (and audio) recording of the session which was needed to analyze the data. To protect their privacy, the recordings were anonymized. For example, we only filmed participants' hands and drawings.

After the introduction, an anonymous ID was assigned to each participant, which was used during the analysis instead of their real data. This was followed by the interview. After the interview part was finished, the participants were debriefed and were given the chance to ask questions themselves. Each session lasted around 40 to 70 minutes. Since we always target to provide fair compensation for each respective country, participants received a compensation of around €60, which was based on their travel and time effort.

3.1.3 Participants

We recruited 22 interview participants from Germany. In order to recruit participants from the general population, we worked together with an external recruiting agency providing them with a detailed screener. The study was advertised as being about online data. The most important screening criteria were that they regularly engaged in online deletion activities and the categories of our scenarios: They had to use some sort of social network and own

ID	Age Range	M / F	Occupation
1	18 - 24	M	Student (Business Mgmt.)
2	45 - 54	F	Industrial Eng.
3	45 - 54	M	Self-employed (Tourism sector)
4	45 - 54	F	Freelance Manager in Public Health
5	45 - 54	M	Insurance Salesman
6	35 - 44	M	Hotel Clerk
7	35 - 44	M	Tour Guide
8	55 - 64	F	Clerk
9	25 - 34	M	Business Management
10	18 - 24	F	Student (Accounting)
11	25 - 34	M	Financial services
12	25 - 34	M	Electrician
13	45 - 54	M	Business Management
14	45 - 54	F	Office Manager
15	25 - 34	F	Automotive Engineer
16	45 - 54	M	Painter/Varnisher
17	35 - 44	F	Office Comm. Clerk
18	35 - 44	F	Real Estate Mgmt.
19	45 - 54	F	Florist
20	45 - 54	F	None
21	45 - 54	M	Insurance Salesman
22	35 - 44	F	Office Clerk

Table 1: Demographics of the interview study participants.

an email account and use it through its online interface. They were also familiar with online shopping and regularly performed online searches. With respect to diversity, we targeted for gender diversity, different professional backgrounds and education, as well as differing attitudes towards privacy.

Table 1 lists the demographics of all interview participants.

3.2 Expert Focus Groups

Instead of reproducing the interview study for the experts, we decided to run focus groups. This decision was made to enable discussion among the experts, which we identified as a vital step to come up with a solid baseline to compare the interview results against.

The focus groups were conducted in combination with a drawing task identical to the interview study. In contrast to the interview study, we asked focus group participants to do the 3 drawings (General (no scenario), Email scenario, Social Media scenario) as homework before the actual meeting. All necessary instructions were sent to them via email. They were also asked to bring these drawings with them to the focus group.

The actual focus group session consisted of 3 main parts (in this order): Data deletion in general, the Email scenario, and the Social Media scenario. For each part, each participant (counterbalanced per part) presented the respective drawing and discussed it with the rest of the group. Then, after everyone presented, the participants were asked to decide which parts of the presented drawings they thought were the most important ones that a lay person should know in order to have a good understanding about what is going on when deleting online data.

3.2.1 Procedure

Both focus groups were conducted at our premises in Switzerland. Two researchers conducted the focus groups together. One of them took notes and the other researcher was leading the focus group (including presentations and discussion).

Before attending the focus group, participants were introduced to the study via email. Moreover, they were asked to read and sign a consent form. At the beginning of the sessions, we ensured that all participants understood and signed the consent form, after which we explained the focus group procedure to them. The consent form mainly asked for permission to make a video (and audio) recording of the session which was needed to analyze the data. To protect participants' privacy, the recordings were anonymized like in the interview study. After the introduction, an anonymous ID was assigned to each participant, which was used during the analysis instead of their real data.

This was followed by the actual focus group. In the end, participants were debriefed and were given the chance to ask questions.

Both focus groups lasted around 60 minutes. Each expert received a compensation worth €30 with respect to the time they invested in being part of the study. Please note that they did not have to travel as we conducted the focus groups in their office spaces.

3.2.2 Participants

Overall, we recruited 7 participants from a major tech company, three for the first and four for the second focus group. Recruitment was done through the company's internal communication channels by specifically targeting pre-identified product areas that involve data deletion.

We targeted participants working in security and privacy and for whom online data deletion and retention are part of their daily job. Thus, we considered these participants experts in the technical parts of online data deletion. We aimed for a good mix of job level and nationalities. We also made sure they all worked on different types of products, and thus, types of online data deletion, to mitigate the influence of a certain type of application on the results. For example, occupations ranged from log specialists to data monitoring.

3.3 Data Analysis

Data analysis of the study results (both interviews and focus groups) took roughly two months from first to last session. Overall, three researchers were involved in the analysis process.

For both, the open-ended questions and the drawings, we used the same inductive coding approach: Two researchers independently coded the entire dataset and each separately came up with a codebook. Disagreements between both codebooks (<7%) were discussed by these researchers and resolved in two in-person sessions. The resulting codebook was then iterated on by both researchers by independently re-coding the dataset. Further disagreements were resolved in further in-person meetings. The final codebook was then used by one researcher to code the entire dataset.

Please note that for the drawings, the analysis did not only involve the actual drawings but also the transcripts of what participants said while drawing (think-aloud). Based on those two data sources (transcripts and drawings), we identified all elements that participants thought were part of the process as well as the elements' interdependencies (e.g. backup servers that are connected with each other). For the sketches, we did not differentiate between written elements (in words, e.g. "cloud") and drawn elements.

After the final codes were assigned, a third researcher joined the analysis process and took part in a two days analysis workshop and two additional refinement sessions. In those sessions, the data of the two studies (interviews and focus groups) was used by the three researchers to identify and discuss overarching themes. For instance, the final list of the expert focus group codes was used

Code	Email Scenario	Social Media Scenario	Total
Not needed anymore, old/outdated	10	6	16
Too much data, limited storage	10	–	10
Tidying inbox, avoiding cognitive overload	6	1	7
To remove Spam/Ads	6	–	6
To remove potentially embarrassing content	–	4	4
Don't delete data	3	7	10

Table 2: This table shows how many participants mentioned each of the following reasons for deletion in their responses for each scenario. Numbers do not add up to 22 because each participant can fall into several categories or none.

to iteratively go through the interview data (and codes) again to identify how they related to each other (i.e., how they were similar or different).

For all themes, saturation was reached after a maximum of 14 participants (excluding the experts), which indicates that we caught the main insights with the 22 participants that we recruited.

3.4 Results

In the following, we will outline the main themes that came out of the analysis. The results cover the interview study as well as the expert focus groups. For sake of ease, we will refer to the interview study participants as "participants" and to the focus group participants as "experts". For a discussion of the results, please refer to the discussion section.

3.4.1 Reasons for Deletion

We identified 5 main reasons for why participants delete data. Table 2 shows a frequency table of these themes split up by scenario, as well as the number of participants who stated they to not delete data in those two scenarios at all.

The data shows that deletion is much more frequent in the Email scenario with storage limitation being one of the major reasons. Participants also deleted emails because the data were no longer needed (10 participants). For 6 participants, deletion was carried out to keep a tidy inbox, and to avoid cognitive overload when checking emails.

In the case of Social Media, participants' main reason for deletion was to remove data which they considered outdated and no longer useful (mentioned by 6 participants), as well as potentially embarrassing content (mentioned by 4 participants and not mentioned at all in the Email scenario). For example, Participant 1 recalled deleting a few posts from a Social Media site because *"they were old, weird, embarrassing stuff I posted when I was 15"*. This is in line with reasons for deletion in social media as presented by Wang et al. [16] (we cover all of them under this category).

The number of participants who did not delete their online data differed in the two scenarios as well. While only 3 participants stated they did not delete emails, 7 participants stated that they do not delete social media data. For the Email scenario, essentially unlimited storage was one of the reasons mentioned why online data was not deleted, as stated by Participant 1: *"I just archive them [emails], in case I need them later on"*. For the Social Media scenario, data was not deleted because many participants declared to be passive users, therefore not having much of their own data added to the social media platforms they used, as exemplified by Participant 4: *"I mostly look at others' content, until now I have no need for that [referring to deletion], I don't have any information there that should be deleted"*.

Code	General Deletion	Email Scenario	Social Media Scenario	Total (Unique)
Components involved				
Servers	13	9	8	30 (15)
User Interface (e.g.,trash bin)	10	7	2	19 (11)
Databases (storage)	1	2	3	6 (3)
Internet	6	–	–	6 (6)
Cloud	3	–	1	4 (3)
User Account	–	1	2	3 (3)
Satellite	2	–	–	2 (2)
Finality				
Data is retained	10	8	4	22 (14)
Data is gone	6	3	9	18 (14)
Data remains in other places (e.g., recipient)	–	7	3	10 (10)
Only permanently deleted once deleted from Trash	–	7	1	8 (7)
Deletion is not entirely possible (permanent traces remain, data can be recovered)	4	2	–	6 (5)
Privacy Concerns Expressed				
Don't know if data is really gone from everywhere	–	3	10	13 (11)

Table 3: This table shows for each scenario, how many participants mentioned the following components, finality of deleted data, and whether they expressed privacy concerns regarding deletion. Numbers do not add up to 22 because each participant can fall into several categories or none.

3.4.2 Dimensions of Deletion

Participants mostly described deletion along two main dimensions: *components involved* (e.g., server, "the cloud"), and *finality of the deletion process* (e.g., the end state of the process). Table 3 shows the frequency with which participants mentioned each of the components involved, their understandings regarding the finality of the deletion, and if they expressed privacy concerns regarding the deletion process.

Ten participants in the general deletion scenario and 7 participants in the Social Media scenario associated the process of deletion with elements of the UI, using UI terminology (such as "trash can") to

Code	Email Scenario	Social Media Scenario	Total
Server/ Database/ Cloud/ Internet	13	18	31
With Recipient	11	3	14
Account	4	4	8
Device	4	3	7
No idea	2	1	3
Satellite	2	–	2

Table 4: This table shows how many participants thought data was stored at each of these locations. Numbers do not add up to 22 because each participant can fall into several categories or none.

explain deletion. The general understanding of deletion at the front end was that data is selected, a delete command is given (e.g., pushing a "delete" button), and then data is gone. We can see this deletion process explained by Participant 4: *"If I press delete, it's gone, not anymore inside, that's what I understand."* Four participants' view of online data deletion only included interactions which occurred at the UI front end.

The rest of the participants (18) described a second part of the deletion process which occurs in the back end. The major part of these participants were most familiar with servers as components, although they were not always clear on exactly what functions they served, often using the terms "server" and "cloud" interchangeably (please refer to Table 3 for the number of participants for each scenario). Participants who were aware of the backend also mentioned components such as databases (3 participants in the Social Media scenario), and the internet (6 participants, for the general scenario). In terms of the deletion process, these participants generally described a server or cloud as a place through where data transits, with data being stored on servers, the cloud or databases. Participant 6 describes this process: *"So my data is on the server, I am on the internet and I connect to the server and telling it to delete my data. Then the server isn't going to delete it completely. I think they have a second server, and they transfer the data there, the 'trash' server, and I don't know what will happen afterwards."*

Seven participants in the Email scenario and 3 in the Social Media scenario mentioned that data remains in other places, such as with the recipient, or on the provider's server. Ten participants mentioned that data is retained (general scenario), and four participants mentioned that deletion is not entirely possible (Email scenario), as explained by Participant 10: *"I think that no data is really deleted."* Most privacy concerns were mentioned for the Social Media scenario (10 participants). Also in the Social Media scenario, 9 participants mentioned that data is just gone after deletion.

3.4.3 Data Storage

Across the different interview parts, participants mentioned data storage before deletion as an essential part of the deletion process as its complexity influences whether or not data will be gone (immediately). The most common responses for both scenarios were server, database, cloud or "the internet." As we can see in Table 4, 13 (Email) and 18 participants (Social Media) thought that data was stored in these locations. Please note that participants often used the terms "server" and "cloud" interchangeably, so in their understanding they serve the same or similar purposes. Eleven participants also mentioned that data could be stored with the recipient in the Email scenario, but this was only mentioned by 3 participants in the Social Media scenario (please refer to Table 4). In both scenarios, participants referred to their accounts or their devices as places where data can be located as well.

Code	Email Scenario	Social Media Scenario	Total
Backups for provider, because they can store everything	3	11	14
Law enforcement	8	6	14
To learn about/profile users for marketing purposes	2	5	7
Data not stored indefinitely, provider keeps data for retention period	6	1	7
No idea/ no reason given	2	2	4
Data sold to 3rd parties	–	3	3
Backups to help user recover data	2	1	3
Deletion in the world wide web isn't possible	–	2	2

Table 5: This table shows how many participants thought that data was stored for these reasons, in each of the scenarios. Numbers do not add up to 22 because each participant can fall into several categories or none.

Participants also discussed reasons for data being stored at these locations. As shown in Table 5, for the Email scenario, 6 participants noted that data is stored for a given retention period (the exact duration of which could not be specified), but not indefinitely. In the case of Social Media, this was not the case, with only 1 participant mentioning that data was not stored indefinitely. In terms of reasons why Email data was stored, law enforcement (e.g., as evidence in a criminal case) was the most often mentioned reason (8 participants), such as stated by Participant 17: *"[data is stored] under certain circumstances like legal enforcement."* Backups were another prominent reason why data was kept by the provider (3 participants). Only two participants mentioned that Email data was retained to profile users, possibly for marketing purposes.

In the case of Social Media, as shown in Table 5, backups by the service provider were the most commonly cited reason (given by 11 participants) explaining why providers keep data. The next most commonly given reasons were law enforcement, mentioned by 6 participants, and 5 participants mentioned profiling users for marketing purposes, as explained by participant 8: *"I think they are collecting all data. I don't know where they store it, but they keep it for sending commercials or something like that to your profile, to see your habits and what you like."* In this scenario, two participants thought a consequence of this was that deletion in the world wide web is not possible, and three participants thought that their data was sold to third parties.

Although several participants in both scenarios mentioned data being kept by providers in the form of backups, only 2 participants in the Email scenario and 1 participant in the Social Media scenario thought that these backups were kept to help the user recover data which was accidentally deleted. The other participants saw backups as a part of business processes, and these backups were not necessarily accessible by users.

3.4.4 Automatic Deletion

As mentioned before, one part of the interview study was dedicated to data expiration, i.e. automatic deletion of data. While expiration was mentioned as a theme across the study, the results in this section are mainly based on the expiration exercise.

We explored expiration by having participants consider how useful it is to them that a specific service provider has their data, and how this value evolves over time. It turned out that all participants had major issues thinking about changes to this value over time, for all 4 different scenarios. The overwhelming majority of participants thought that changes to this value were related to specific events which were not time bound, for example canceling their account with the service provider.

In some instances, participants could think about specific situations in which it was no longer useful for them that the service provider held their data. However, this did not necessarily mean the overall end of its usefulness. Certain events were able to "revive" data and increase its value again. For example, several participants thought that it was always useful for websites from which they shop online to have their address for delivery. In the short term, after a particular delivery is received, it was no longer as immediately useful that the service provider has their address data. However, as participants put up another order with those shops, it was once again useful that the provider has their data. Therefore, as opposed to our assumption, the value to users that service providers have their data does not change in a linear fashion but comes in waves or short bursts of usefulness because it is highly context-dependent.

3.4.5 Supportive Deletion Knowledge

Based on their detailed knowledge of online data deletion, experts agreed on six major topics they thought would be beneficial for users to know. "Beneficial" refers to the fact that experts thought that knowing these things will help users to make appropriate decisions that help them better maintain their data privacy. They did not expect users to have such detailed knowledge. However, they assumed that users would benefit from this knowledge. How users could acquire this knowledge was not part of the discussion.

The six topics are:

Backend - This refers to knowing that something is happening beyond the interface. Data will be sent to different servers and will be stored. There will also be copies of the data. This was considered a crucial aspect for the understanding of online data deletion.

Time - Data is not immediately deleted after pressing the delete button. Data may still rest somewhere, even though the users might not be able to see it on their screen.

Backup - Identical data may exist in different places for data storage and data security reasons. In addition, the same information may be stored in different services except the service where it was deleted. For example, travel information might be deleted from an email account but could still be available in a calendar service.

Derived Information - If data is deleted, its essence might still exist. For instance, a user might have deleted a song from a playlist, but the musical interest profile still has this information. Unlearning of derived information like this takes time and thus, deleting data might not immediately change the corresponding profile.

Anonymization - In many cases, a first step of deletion is removing the connection between the data item and the user. After this point, the data might still exist for a while but cannot be related to the user anymore.

What experts consider helpful for users to know	Participants' mentions across all scenarios (N=22)
Backend	18
Time	16
Backup	7
Derived information	1
Anonymization	1
Shared Copy	7

Table 6: Concepts experts think users should know in order to better understand deletion, and the number of participants who are at least slightly aware of these.

Shared Copies - Experts added that users should know about shared copies. Other users might have a copy of the data, e.g., a deleted email. As one expert put it: "Better think before posting and regretting it later."

3.4.6 Expert and Participant Knowledge

In the last rounds of data analysis (e.g., during the workshop days), this list was used to analyze overlap with what participants mentioned throughout the study. We counted an overlap if the participant had mentioned the item at least once during the whole interview (including the drawing tasks). Please note that degrees of knowledge between participants varied significantly. Some participants briefly or inaccurately touched one of the topics and did not further elaborate - even when prompted to do so. However, we wanted to learn, if participants are generally aware of the topics experts mentioned. Thus, we did not differentiate whether participants thoroughly discussed these topics or only briefly mentioned them.

In this analysis, we observed a huge discrepancy between the different topics experts consider helpful for users to know about deletion. Most interview participants expressed awareness for two of the topics: 18 out of 22 participants were aware that something is happening in the Backend and 16 participants acknowledged that it will take some time until data is finally deleted (see Table 6). However, only few participants brought up the topics of Backup (7/ 22), Derived Information (1/ 22) and Anonymization (1/ 22). In the Email and Social Media scenarios, only 7 participants mentioned that other users might still hold a copy of the data they deleted.

3.4.7 Views and Understanding of Deletion

Overall, we found that participants differed in their view of online data deletion across five parameters. The first two were *components involved* in deletion, and the *terminology* used to refer to them. The third was how these *components interact*, and the fourth was whether a *backend* (anything beyond the UI) was identified. Finally, their understanding of online data deletion was also different in the *duration* of the deletion process.

It should be noted that these parameters are reflecting the complexity of the participants' views of deletion, and not their technical accuracy. Thus, the following should not be interpreted as a quality rating of the responses.

By analyzing participants' responses across these parameters, we identified two general distinct categories of understanding of deletion[3] as shown in Figure 3. The first category reflects a UI-centric understanding of deletion. Therefore, we refer to it as the *UI-Based*

[3]During the iterative analysis process, we at first identified 4 categories that we then narrowed down to the 2 presented here.

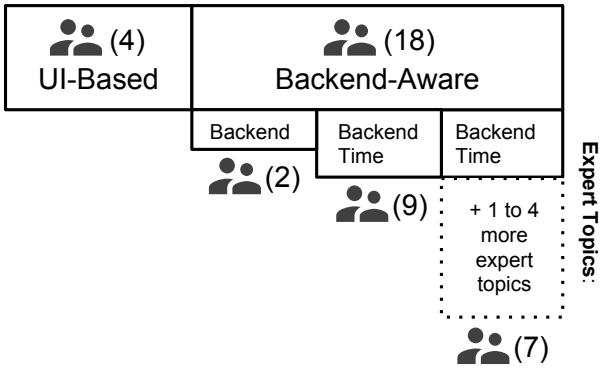


Figure 3: Two categories of user understanding of online data deletion: UI-Based and Backend-Aware. The second category can be subdivided using the topic list of the expert focus group: Backend, Time, Backup, Derived Information, Anonymization, Shared Copy.

Figure 4: A UI-Based view of online data deletion, explaining how email deletion works (Participant 4).

category. The 4 participants that fell into this category displayed an understanding of deletion and a terminology completely based on the UI components they were most familiar with, such as checkboxes to select emails, and then pressing the delete button, so that data ends up in the trash bin. Backend knowledge was not part of this view. Consequently, participants in the UI-Based category described the deletion process as being completed within seconds of clicking the delete button. Figure 4 shows a sample diagram of the UI-Based category.

The second category we identified was more complex. The 18 participants that fell in this category identified more components involved in the deletion process, particularly backend components such as servers or the cloud. Therefore, we refer to this as the *Backend-Aware* category. The components mentioned by the participants also interact with each other, such as sending a delete command to a server from a device where the user is accessing their email account. However, the terminology used to describe these components was often inconsistent, with terms such as "cloud" and "server" being used interchangeably. Since a backend to the deletion process was identified, participants tended to understand the deletion process as taking longer than a few seconds, even if the exact duration of the process could not be identified. Please refer to Figure 5 for two sample drawings of the Backend-Aware category.

Participants that fell into this category distinguished between deletion at the UI level as opposed to data being purged from backend. No participant mentioned the risk of data being stolen nor the advantages of retention after deletion for recovery. However, this advantage was mentioned for data in the trash.

Unsurprisingly, all experts fell into the Backend-Aware category expressing varying degrees of knowledge about what exactly goes on in the background. In general, experts' knowledge around deletion surpassed all interview participants' knowledge by far.

While the understandings and drawings categorized as the UI-Based category were rather homogeneous, this was not the case for the Backend-Aware category. Since all topics mentioned by the experts fall into this more complex view, we used those six topics to further divide this category into three sub-categories, based on how many of these dimensions were included.

Two participants had a view which only included "backend" (see Figure 3). Nine participants fell into the next sub-category, which includes both concepts of a backend and time. Seven participants fell into the more complex sub-category, which includes both a backend, time, and at least one of the other dimensions. Of all the participants, only one mentioned all the dimensions, and was the only one to include the more complex concepts of derived information and anonymization.

4. DISCUSSION

Our study results revealed a plethora of reasons, views and understanding, and needs when it comes to online data deletion. In this section, we will provide some lessons learned and implications based on these results. Please note that while the results are based on two specific use cases (plus two for expiration), our recommendations go beyond these two instances.

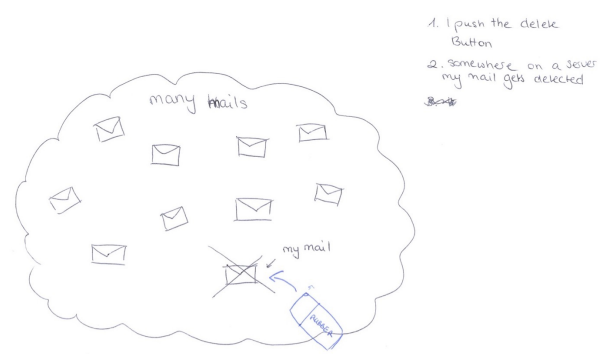
4.1 No One-Size-Fits-All Solution

As mentioned before, we used email and social media as scenarios because we hypothesized that they represent different ends of the deletion spectrum (i.e., different types of data generated in different ways). Our results show that this assumption held true. Understanding as well as views and needs for the two scenarios differed to a great extent. For instance, reasons for deletion had little to no overlap and were highly service-dependent.

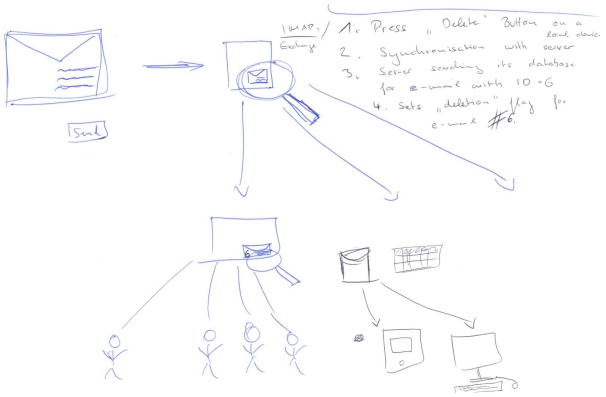
This shows that there is likely no one-size-fits-all solution when it comes to deletion strategies (from both a UI and technological point of view) which means that these individual differences need to be taken into account when designing deletion for a specific online service. For instance, understanding of (what happens during and after) deletion depends to a great extent on how a service handles its data and deletion should be reflective of this.

4.2 No Generalization of Data Deletion Needs

Related to this, we observed a great number of reasons to delete data, including privacy issues. The most prominent one (and also the only one consistent across the two scenarios) was getting rid of old or outdated data that is not needed anymore. Another interesting reason, which is related to the value of data, is deletion to tidy (or clean) an account. Participants mentioned that certain data would pollute their accounts and they wanted to get rid of this data.



(a)



(b)

Figure 5: The differing complexity of the Backend-Aware categories can be seen by comparing diagrams (a) and (b), drawn by Participants 10 and 11, respectively.

Whether data is still needed is decided on a highly individual level depending on different factors such as context, service, and usefulness, and it cannot be generalized. Even within a participant, no consistent reasons for deleting data across services (or cases) could be identified as they decide these issues on a case-to-case base.

Similar to the previous insight (no one-size-fits-all solution), a major consequence of this is that we cannot generalize user needs of online data deletion across services. For instance, while providing unlimited storage space can make many cases of deletion unnecessary in the email space, this does not translate to social networks in which publicness and embarrassment are much bigger factors.

4.3 Communicating Deletion

The study results showed that certain concepts related to online data deletion were highly present in the participants' view of deletion even though they were not necessarily correctly used from a technical point of view.

Terms and functionality related to different components in the backend (or the backend in general) were mentioned by the majority of participants. In many cases, they referred to them as the reason for increased data retention periods, i.e., the fact that data is not deleted immediately. The participants that connected these technical constraints with data retention were also more likely to find it acceptable or understandable as opposed to the participants who thought that data was solely retained for business purposes.

Other concepts were harder to understand and thus seldom part of their mindsets. One example is anonymization, which was only mentioned once.

A consequence of this is that communicating (and explaining) online data deletion using these more common concepts has the potential to positively affect users' attitude towards technological constraints of deletion. As an example, based on this, a promising direction for explaining retention periods might be to built it around technical complexity of removing it from servers, backup servers, and the like.

4.4 Deletion in the UI

Related to the previous section on communicating deletion, we think that our results can have direct influence on how deletion user interfaces are designed.

For instance, a common practice for services with a trash folder is to highlight that fact in the deletion dialogue (e.g., along the lines of "This file has been moved to the trash."). Similar to this, one could imagine that when deleting a file for good by removing it from the trash, the following procedure could be teased, again, based on parts of the process that users understand (e.g., "This file will now be deleted from our servers" to indicate technical complexity).

That said, we do not have data to judge how this should look like exactly and thus argue that this would have to be evaluated in further studies, especially with a focus on how upfront such messages would have to be to provide the best effect.

4.5 Control of Expiration

Expiration is a special use case of deletion: automatic deletion after a certain time. We worked based on the assumption that expiration for data could be represented on a timeline together with certain events that mark the end of its usefulness to a user.

However, the study results showed that this did not hold true for any of the scenarios. While there are single instances (or single participants) that could identify such an event, it was highly context-dependent. In addition, for each data item (and scenario), participants could identify events or situations which would give new value to information that was previously marked as useless.

This indicates that enforcing specific expiration periods on undeleted user data is likely to create situations in which useful (or wanted) data is not available anymore. A potential consequence of this would be a reduction in service quality from a user's point of view.

Participants indicated that control, especially self-selected expiration conditions, would be a better way to approach this issue. One participant proposed the following approach for email deletion: *"You could have a folder which allows you to set an expiration date for items in this folder. Like when I move an email in there, it could be automatically deleted after 30 days or whatever amount I decide."*

This type of control mechanism highlights another result of this research: users can relate even abstract concepts very well to the UI. Therefore, we can leverage this to communicate with users through well-known concepts and metaphors, such as the "trash can".

Summed up, this means that, instead of automatic (default) data expiration, allowing control over how data expiration is handled on an individual level is a more promising direction. This would also give users more control over their data (preferences).

4.6 Shared/Implicit Copies Not Well Understood

Participants understood the concept of shared copies for emails rather well. It is easy to comprehend that when you send out an email, a copy of it will exist with the recipients. As a consequence, they pay more attention to what they write due to the fact that control about the data will be lost [6].

As opposed to this, for social networks, this problem was not well understood, despite it being similarly likely as shown by related work on Twitter and Facebook [1]. For instance, only 3 out of 22 participants mentioned that data might be stored with a recipient (or the like). Thus, the idea of (not necessarily verbatim) copies based on retweets and other ways of interacting with a post seems to be less graspable. This is even worse as implicit copies can be a challenge to the user privacy as the user loses control over the content but might not even be aware of the existence of the copies.

While our data provides further insights into this being an issue that potentially affects user privacy, we do not have data to make recommendations on how to mitigate this risk. However, we argue that this is an important topic for the research community to study and want to highlight its necessity.

5. LIMITATIONS

The main limitation of this work is the limited sample size of the interview study. While we made sure to recruit participants from a wide spectrum of society, the data should not be interpreted as generalizable to the whole (internet) population but rather as trends. That said, we are confident that we cover the most relevant themes, which is supported by the fact that we reached saturation of themes after (max) 14 participants.

Furthermore, despite carefully selecting the scenarios to cover a wide range, the results are limited to the two tested contexts (plus two for expiration). As mentioned in the discussion, results might have been different had we tested other services (e.g., cloud storage), and thus, recommendations in this paper should be handled with care in these contexts. Since the selected scenarios cover different ends of the deletion spectrum, we argue that the major insights of this work are still (partially) applicable to online deletion overall.

6. CONCLUSION

In the present work, we explored users' understanding of online data deletion which is essential to maintaining user privacy and protecting their data. We identified two main views on how deletion works: UI-Based and Backend-Aware. We found that a large majority of participants were aware of a backend to the deletion process. Although participants' understanding of the backend processes of deletion varied in their complexity, explanations of online data deletion can build off of this understanding to explain the technical constraints of deletion in conceptual terms. Our results indicate that doing so could also have the potential to positively affect users' attitudes toward these constraints and be more accepting of certain retention periods.

Our results also provide insights into expiration preferences. We found that participants considered the usefulness of their online data to be very context-dependent, as opposed to time bound. Consequently, participants did not envision their online data having an expiration date that could be set on a chronological scale. Participants therefore favored control over the expiration of their data, such as moving data to a specific folder where they can manually set expiration dates.

A challenge raised by this work relates to user understanding of shared copies of online data for services where it is not well understood and can be problematic in terms of privacy. While the concept of a shared copy is clear for email (i.e., the recipient has a copy), it is not so clear in the social media contexts, where different ways of interacting with the data could lead to different copies (e.g., re-posts) or traces of it (e.g., comments referencing a post). Future work should explore these understandings, and how to best communicate to users this concept of shared copies in complex settings.

7. ACKNOWLEDGMENTS

First of all, we would like to thank our interview and focus group study participants for their valuable time. Furthermore, we are very grateful for the input provided by both the internal Google reviewers as well as the external reviewers whose input significantly helped to improve this work and this paper.

8. REFERENCES

[1] H. Almuhimedi, S. Wilson, B. Liu, N. Sadeh, and A. Acquisti. Tweets are forever: A large-scale quantitative analysis of deleted tweets. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 897–908, New York, NY, USA, 2013. ACM.

[2] O. Ayalon and E. Toch. Retrospective privacy: Managing longitudinal privacy in online social networks. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, pages 4:1–4:13, New York, NY, USA, 2013. ACM.

[3] L. Bauer, L. F. Cranor, S. Komanduri, M. L. Mazurek, M. K. Reiter, M. Sleeper, and B. Ur. The post anachronism: The temporal dimension of facebook privacy. In *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society*, WPES '13, pages 1–12, New York, NY, USA, 2013. ACM.

[4] M. Bishop, E. R. Butler, K. Butler, C. Gates, and S. Greenspan. Forgive and forget: Return to obscurity. In *Proceedings of the 2013 New Security Paradigms Workshop*, NSPW '13, pages 1–10, New York, NY, USA, 2013. ACM.

[5] L. J. Camp. Mental models of privacy and security. *IEEE Technology and Society Magazine*, 28(3):37–46, Fall 2009.

[6] A. De Luca, S. Das, M. Ortlieb, I. Ion, and B. Laurie. Expert and non-expert attitudes towards (secure) instant messaging. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 147–157, Denver, CO, 2016. USENIX Association.

[7] K. A. Ericsson and H. A. Simon. Verbal reports as data. *Psychological review*, 87(3):215, 1980.

[8] R. Kang, L. Dabbish, N. Fruchter, and S. 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)*, pages 39–52. USENIX Association Berkeley, CA, 2015.

[9] V. Mayer-Schönberger. *Delete: The virtue of forgetting in the digital age*. Princeton University Press, 2011.

[10] M. Mondal, J. Messias, S. Ghosh, K. P. Gummadi, and A. Kate. Forgetting in social media: Understanding and controlling longitudinal exposure of socially shared data. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 287–299, Denver, CO, 2016. USENIX Association.

[11] E. S. Poole, M. Chetty, R. E. Grinter, and W. K. Edwards.

More than meets the eye: Transforming the user experience of home network management. In *Proceedings of the 7th ACM Conference on Designing Interactive Systems*, DIS '08, pages 455–464, New York, NY, USA, 2008. ACM.

[12] K. M. Ramokapane, A. Rashid, and J. M. Such. "i feel stupid i can't delete...": A study of users' cloud deletion practices and coping strategies. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 241–256, Santa Clara, CA, 2017. USENIX Association.

[13] J. Rosen. The web means the end of forgetting, 2010.

[14] M. Sleeper, J. Cranshaw, P. G. Kelley, B. Ur, A. Acquisti, L. F. Cranor, and N. Sadeh. "i read my twitter the next morning and was astonished": A conversational perspective on twitter regrets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 3277–3286, New York, NY, USA, 2013. ACM.

[15] M. Tohidi, W. Buxton, R. Baecker, and A. Sellen. User sketches: A quick, inexpensive, and effective way to elicit more reflective user feedback. In *Proceedings of the 4th Nordic Conference on Human-computer Interaction: Changing Roles*, NordiCHI '06, pages 105–114, New York, NY, USA, 2006. ACM.

[16] Y. Wang, P. G. Leon, A. Acquisti, L. F. Cranor, A. Forget, and N. Sadeh. A field trial of privacy nudges for facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2367–2376, New York, NY, USA, 2014. ACM.

[17] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor. "i regretted the minute i pressed share": A qualitative study of regrets on facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, pages 10:1–10:16, New York, NY, USA, 2011. ACM.

[18] R. Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, pages 11:1–11:16, New York, NY, USA, 2010. ACM.

APPENDIX

The following two sections list the study instruments used for the deletion scenarios and the expiration exercise. Please note that: a) They have been slightly adapted (e.g., the scenarios script actually consists of 3 parts that we merged for the appendix); b) They are listed out of context.

A. DELETION SCENARIOS SCRIPT

1. Do you sometimes delete [emails/tweets/posts]?
   (a) If yes: Why?
   (b) If no: Why not?
2. Now let's imagine you just deleted an [email/tweet/post]. Please draw what happens after you press the "Delete" button. (instruction: hand participant pen and paper)
3. You just named a few things that occur when you delete an [email/tweet/post]. Please write them down as a list in the order in which they occur. Use "Press the delete button" as the first item on your list.
4. Imagine that you pressed the delete button now. When would the last item on your list take place?
5. After this last point on the list, is it possible for you to recover the deleted [email/tweet/post]?
   (a) If yes: Why?
   (b) If no: Why not?
6. Is it possible for the [service provider] to recover the deleted [email/tweet/post]?
   (a) If yes: Why? For what purpose is the data stored?
   (b) If no: Why not?

B. EXPIRATION GRAPH SCRIPT

1. Here is a card with an online context, and a type of personal data associated with that context written on it. (instruction: hand participant one of the cards in counterbalanced order)
2. Here is a screenshot of what this online context would look like. (instruction: hand participant screenshot, read description)
3. On a scale from 1-5, with 1 being the least sensitive and 5 being the most sensitive, how sensitive is this type of data to you? Please write your rating on the card.
4. Now we will be referring to this graph. (instruction: hand participant the expiration graph)
5. On the horizontal axis, please add different events which can occur in this scenario that could have an influence on the usefulness of this data. Usefulness refers to how useful it is for you that the service provider has this data.
6. The usefulness might change over time. Let me give you an example: It is most useful for your dentist to know the time of your appointment before it happens, and still quite useful on the day of the appointment. After the day of the appointment, it is perhaps less useful that your dentist has this information.
7. After the point when this data is no longer useful to you, what should happen to it, if anything? (instruction: if participant added an event with a usefulness rating of 0/1, refer to that point)

Programming Experience Might Not Help in Comprehending Obfuscated Source Code Efficiently

Norman Hänsch
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany
norman.haensch@fau.de

Andrea Schankin
Karlsruhe Institute of Technology
Karlsruhe, Germany
schankin@teco.edu

Mykolai Protsenko
Fraunhofer Institute for Applied and Integrated Security
Garching, Germany
mykolai.protsenko@aisec.fraunhofer.de

Felix Freiling
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany
felix.freiling@cs.fau.de

Zinaida Benenson
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany
zinaida.benenson@fau.de

ABSTRACT

Software obfuscation is a technique to protect programs from malicious reverse engineering by explicitly making them harder to understand. We investigate the effect of two specific source code obfuscation methods on the program comprehension efforts of 66 university students playing the role of attackers in a reverse engineering experiment by partially replicating experiments of Ceccatto et al. We confirm that the two obfuscation methods have a measurable negative effect on program comprehension in general but also show that this effect inversely correlates with the programming experience of attackers. So while the comprehension effectiveness of experienced programmers is generally higher than for inexperienced programmers, the comprehension gap between these groups narrows considerably if source code obfuscation is used. In extension of previous work, an investigation of the code analysis *behavior* of attackers reveals that there exist obfuscation techniques that significantly impede comprehension even if tool support exists to revert them, giving first supportive empirical evidence for the classical distinction between potent and resilient obfuscation techniques defined by Collberg et al. more than 20 years ago.

1. INTRODUCTION

In many developed economies, software is a major driver of innovation and industrial growth. To protect their intellectual property, prevent the creation of illegal copies of software and to avoid the unauthorized program flow changes that might benefit the attackers, software vendors employ various software protection techniques. Software protection is also a technique employed by cybercriminals to prevent malware analysis by security researchers.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

Software protection can be achieved in multiple ways. Historically, one of the most successful techniques is using specialized hardware, i.e., to disallow access to source or binary by moving it into an external tamper-proof execution compartment [1, 2]. A slightly weaker possibility to achieve software protection is to move only critical parts of software to a trusted processing environment such as a special processor mode [3] or a remote server [4]. While such trusted processing environments are much cheaper than specialized tamper-proof hardware compartments, both techniques incur a significant economical and organizational overhead.

A comparatively cheap alternative to additional specialized hardware is to assume that the attacker will eventually be able to access the code, but that the code is constructed in such a way that it cannot be easily reverse engineered. A central method to deter attackers in this context is *software obfuscation*, i.e., a software transformation that makes the program code harder to comprehend and to analyze. In contrast to many techniques offered in classical software engineering, software obfuscation is a security technique that aims at *inhibiting* software comprehension by attackers. It is the standard means to protect the bytecode of Android apps from analysis today, and it is applied in almost all malware samples spreading in the wild. Understanding the strength of software obfuscation is therefore key both (1) to raise the protection level for software vendors and (2) to help malware analysts to prioritize reverse engineering tasks.

In 2001 Barak et al. [5] showed that *perfect obfuscation* (meaning that a program does not expose more information than can be derived from its input/ouptut behavior) is impossible in general. In practice, most software protection techniques rely on the definition of *obfuscation transformation* provided by Collberg et al. [6], which means that the program's code is made somewhat more obscure by the application of the transformation without introducing a too high performance overhead.

Reverse engineering is always a combination of human ingenuity and tool support. This led Collberg et al. [6] to distinguish between resilience and potency of obfuscating transformations: *resilience* means the ability to withstand an *au-*

tomated deobfuscation attack, while *potency* refers to the grade of "obscurity" for the *human* reverse engineer added by the obfuscation. While formal complexity metrics can help approximate resilience and potency [7], the strength of obfuscation cannot be fully understood without analyzing its effect on program comprehension abilities of real users, a topic which we further study in this paper.

1.1 Related Work

Program comprehension is a mature field in software engineering, where qualitative and quantitative human factors methods have been used to study software comprehension and to evaluate tools [8, 9, 10, 11]. However, there is surprisingly little work on software comprehension in the context of software obfuscation. Ceccato et al. [12] pioneered the area by performing a series of five controlled experiments to measure the influence of code obfuscation on understanding decompiled Java source code. They studied two obfuscation methods, identifier renaming and opaque predicates, and showed that they have a measurable effect on the ability of humans to solve code comprehension and change tasks. In this work, we partially replicate their study and confirm their results. Using a similar experimental setup but with different programs, Viticchié et al. [13] analyzed the influence of the VarMerge obfuscation. Compared with clear code, VarMerge obfuscated code led to significant differences concerning time and efficiency of the attack, but not in correctness.

Although attacker modeling has been identified as one of the fundamental challenges in usable security research [14], user studies in secure programming has focused on the defenders so far. Oliveira et al. [15] showed that security is not a priority in programming tasks and needs additional cognitive effort. Acar et al. [16, 17] analyzed the influence of the documentation that programmers use when writing code. In an experiment with GitHub users, Acar et al. [18] found that correctly fulfilling security requirements is influenced by the years of programming experience, but not by professional status, e.g., student or working programmer. The latter is most relevant to our work as we investigate the influence of programming experience on reverse engineering skills.

1.2 Contributions

In this work, we measure the effect of source code obfuscation on program comprehension skills of human reverse engineers by means of a controlled experiment with 66 participants and advance the insight into the distinction between resilience and potency of obfuscating transformations as defined by Collberg et al. [6]. More specifically, our contributions are as follows:

- Using a slightly different study design, we replicate and validate the results by Ceccato et al. [12], i.e., we provide further experimental evidence that source code obfuscation makes program comprehension significantly harder.
- We provide original insight into the effect of two obfuscating transformations onto the *reverse engineering behavior*. We show that code analysis behavior differs significantly when trying to comprehend the results of an obfuscation method considered to be *potent* in comparison to a method that is considered to be *resilient*. We therefore provide first empirical evidence into the usefulness of these concepts that were defined in 1997 [6].
- To better understand the factors influencing the potency of obfuscation methods, we provide additional original insight into the impact of different programming experience levels on reverse engineering performance and behavior. We show that classical programming experience does not prepare well for the task of comprehending obfuscated code: While experienced participants were much more efficient than beginners when they worked on non-obfuscated code, the gap in efficiency narrowed significantly when given the obfuscated code. Specific obfuscation and debugging experience, however, appears to be helpful.

Overall, if software obfuscation is applied to protect malicious software, our insights may help to improve the education of malware analysis professionals. If obfuscation is used to protect legal software, then our insights may be helpful to evaluate the quality of protection.

1.3 Outlook

After providing background in Section 2, we state the research hypotheses in Section 3 and describe the experimental setup and methods in Section 4. Results are presented in Section 5. We discuss implications and limitations of our study in Sections 6 resp. 7, and conclude in Section 8.

2. BACKGROUND

We first provide background on the obfuscation techniques and code analysis. We further give details on the experimental setup of Ceccato et al. [12] upon which we build.

2.1 Obfuscation

The generally accepted view on obfuscation is based on the notion of program transformation making the code harder to analyze and to comprehend. Obfuscation can be applied at any level of abstraction, be it source code, byte-code or machine code. Here, we focus on source code obfuscation for two reasons: Firstly, source code obfuscation is still common in the context of Java since byte-code can be easily decompiled.[1] Secondly, source code obfuscation has been studied by Ceccato et al. [12], whose work we partially replicate.

One of the most widely used obfuscation techniques is *identifier renaming* where the names of classes, fields and methods, as well as of local variables are changed to meaningless character sequences. Since identifiers are usually carefully selected to reflect their semantic meaning, removal of this information complicates the process of code comprehension. *Name overloading* [19] extends identifier renaming by using the same names for multiple different entities. We use name overloading as the first obfuscation technique in our study and abbreviate it by NO.[2]

Name overloading does not change the structure of the code. The obfuscation technique of *opaque predicates* (abbreviated as OP) can be used to alter the program's execution flow. A predicate is called *opaque* if its outcome is known at obfuscation time but is hard to deduce by the reverse engineer [19].[3]

[1] While the new Android Runtime (ART) supports also the distribution of native code, using classical bytecode is still common because of backwards compatibility.
[2] Ceccato et al. also used NO but called it "identifier renaming" and used the abbreviation IR, see also footnote 6.
[3] Examples of true and false opaque predicates are $(x^2 + x)\ mod\ 2 = 0$ and $x^2 + 1 \leq 0$ respectively for any real value x.

Opaque predicates can be used to extend existing branches or to insert dead code. We use OP as second obfuscation technique in our experiments. Appendix A provides code examples to illustrate both obfuscation techniques.

2.2 Code Analysis and Eclipse

The process of understanding of the program's code and its key features is referred to as *code analysis*. *Static* code analysis does not involve actual execution of the program, whereas in *dynamic* analysis, code is at least partially executed. Usually, code analysis is supported by tools. For the purpose of Java source code analysis, the Eclipse IDE can be utilized. It can perform both static and dynamic analysis. In the following we briefly describe the capabilities of this tool, as it was used by participants in our study.

The Eclipse IDE supports static analysis by providing additional information for the source code, such as showing the class inheritance hierarchy or call graph, or highlighting cross-references. It can also automate standard modifications of the program performed by the analyst, such as renaming of variables, methods, and fields, or moving methods from one class to another. In this paper we refer to the latter operations as *advanced Eclipse commands*.

The Eclipse IDE also supports program execution in the *debugging* mode. Using this mode, the analyst can perform single-stepping, executing only one instruction at a time, set breakpoints, watch the variable values, and so on. This functionality can be useful to follow the execution of the code under analysis, in order to better understand the dependencies between code and external program's behavior, or to identify predicates suspicious of being opaque, namely those that always have the same value at runtime.

2.3 Obfuscation Studies by Ceccato et al.

Ceccato et al. [20, 21, 12] conducted a series of experiments using two programs and letting participants solve two code comprehension tasks and two code modification tasks for each program. The experiments varied in the type of obfuscation, the type of students (bachelor, master or PhD) and the universities in which they took place, while the experimental tasks remained the same. Since the results of all experiments are summarized in one single paper [12], we refer to this paper in the following.

One of the programs (called *Race* in the following) is an online game that lets two players conduct a car race. Another program, called *Chat*, lets people have public or private online conversations. The programs were given to the participants as source code decompiled from Java bytecode, as this is the usual way how the reverse engineers work on Java code. Depending on the experiment, the programs were provided in different variants: as clear code (unobfuscated), obfuscated with identifier renaming (which was in fact name overloading), or obfuscated with opaque predicates. General software metrics for both programs presented in Table 1 show that the programs are comparable in their complexity. Although Race has a higher number of methods and lines of codes (LOC) than Chat, it has a lower overall cyclomatic number [22] (roughly corresponding to the number of linearly independent paths in a function's code).

Ceccato et al. [12] conducted five experiments that cumulatively evaluated whether code obfuscation influences performance of reverse engineers: Do people solve code comprehension tasks slower and less correct on obfuscated code? If yes, which of the obfuscation methods (NO or OP) reduces the performance more severely? The code comprehension tasks from the study can be found in Table 2.[4] Overall, Ceccato et al. found statistically significant differences only for the obfuscation technique NO, supporting the belief that opaque predicates help to slow down automated analysis rather than performance of human reverse engineers.

	Race			Chat		
Metric	*Clear*	*NO*	*OP*	*Clear*	*NO*	*OP*
Classes	14	14	14	13	13	13
Methods	109	109	125	72	72	88
LOC	1215	1215	3783	1030	1030	3642
$\sum$ Cyclomatic	244	244	1131	253	253	1775

Table 1: Software metrics calculated for the different versions of the programs. Due to the nature of NO, the metrics are the same for NO and Clear.

Task	Description
Race: Box	In order to refuel the car has to enter the box. The box area is delimited by a red rectangle. What is the width of the box entrance (in pixel)?
Race: Laps	When the car crosses the start line, the number of laps is increased. Identify the section of code that increases the number of laps the car has completed (report the class name/s and line number/s).
Chat: Messages	Messages going from the client to the server use an integer as header to distinguish the type of the message. What is the value of the header for an outgoing public message sent by the client?
Chat: Users	When a new user joins, the list of the displayed "Online users" is updated. Identify the section of code that updates the list of users when a new user joins (report the class name/s and line number/s).

Table 2: Participants' tasks for the study of Ceccato et al. [12] and ours.

3. HYPOTHESES

We formulate research hypotheses that aim at answering the following research questions:

- Can we validate the results of Ceccato et al. concerning code *comprehension*?
- Does obfuscation influence the code analysis *behavior* of attackers?
- Does programming *experience* influence code comprehension and behavior of the attackers?

[4] Ceccato et al. also investigated code change tasks that we do not consider in our study. We present differences between their and our study in more detail in Section 4.

3.1 Code Comprehension Hypotheses

Considering the effect of code obfuscation on code comprehension, we evaluate the following hypothesis:

> Obfuscated code is more difficult to comprehend than clear code.

However, the term "obfuscated code" can be instantiated in many different ways. To evaluate such a hypothesis it would be necessary to investigate the effects of a "representative" set of obfuscation methods and it is not entirely clear what this could be. We therefore focus on the effects of the two obfuscation techniques NO and OP and conduct partial replication of prior work by Ceccato et al. [12].

Ceccato et al. found no significant difference between correctly comprehending clear code and code obfuscated with any of the two obfuscation techniques. For the efficiency the results differed. While no significant difference in the efficiency between working on clear and OP-obfuscated code was found, efficiency of working on NO-obfuscated code significantly decreased compared to working on clear code. Similarly, only working on NO-obfuscated code took significantly longer than working on clear code. Ceccato et al. [12] therefore rejected several of their hypotheses concerning OP-obfuscated code. However, since the number of participants in the various studies was quite small (10 to 22), we assume that some effects of the obfuscation methods might have been missed. Therefore, for code comprehension we formulate the same set of hypotheses as Ceccato et al. [12], where capitalized words set in italics indicate independent variables for the statistical analysis:

HC1$_{\mathbf{NO}}$ *NO-obfuscated* code is more difficult to comprehend than *Clear* code.

HC1$_{\mathbf{OP}}$ *OP-obfuscated* code is more difficult to comprehend than *Clear* code.

These hypotheses attempt to approximate the hypothesis on the general effect of obfuscation presented above.

Following the discussion on the *potency* of obfuscation methods [6], i.e., the differing grades of "obscurity" for the *human* reverse engineer added by the obfuscation, the next hypothesis aims at insights into the effects of conceptually different obfuscation techniques. Ceccato et al. found that understanding NO-obfuscated code is more difficult than understanding OP-obfuscated code. However, this difference was statistically significant in only one of two experiments that they conducted with this goal. We seek to validate their results with the following hypothesis:

HC2 *NO-obfuscated* code is more difficult to comprehend than *OP-obfuscated* code.

Ceccato et al. also report that participants with higher experience (measured by their study degree: bachelor, master or PhD student) performed slightly better on both, clear and obfuscated code (the results were not statistically significant). We therefore formulate the following hypothesis:

HC3 The higher the experience of attackers, the easier they comprehend *Clear* and *Obfuscated* code.

3.2 Code Analysis Behavior Hypotheses

Ceccato et al. did not investigate behavior of attackers in solving their tasks. However, they asked participants some questions about their analysis behavior in a post-experimental question, e.g., which percentage of the task time they spent reading the code, or how many program executions in debugging mode they used. They report some (mostly not statistically significant) differences in the answers for clear and obfuscated code. We take their investigation as an inspiration for looking at the *actual* attacker behavior.

In practice, the first step in code comprehension of obfuscated code is usually to identify the particular obfuscation technique and perform experiments with tools for automatic deobfuscation [23]. It is therefore to be expected that comprehension of obfuscated code results in different code analysis behavior from classical reverse engineering, namely that behavior attempts to first identify the obfuscation method or performs simple deobfuscation tasks. In general, we therefore evaluate the following hypothesis:

> Code obfuscation significantly changes code analysis behavior in comparison to analysis behavior for clear code.

Since code analysis behavior appears to target the obfuscation method first, we expect to find differences not only between clear code and obfuscated code in general, but also differences in the behavior between code obfuscated by different obfuscation techniques. We therefore explore the novel behavioral research question by evaluating the following hypotheses with regard to the *behavior* of the attackers for code comprehension tasks:

HB1$_{\mathbf{NO}}$ When analyzing *NO-obfuscated* code attackers behave differently than when analyzing *Clear* code.

HB1$_{\mathbf{OP}}$ When analyzing *OP-obfuscated* code attackers behave differently than when analyzing *Clear* code.

HB2 When analyzing *NO-obfuscated* code attackers behave differently than when analyzing *OP-obfuscated* code.

Since comprehending obfuscated code in practice seems to require additional expertise, we also formulate hypotheses concerning the influence of experience, as previously done for code comprehension:

HB3 Experienced attackers behave differently than beginners when analyzing *Clear* and *Obfuscated* code.

3.3 Measurements

We now describe how we measured code comprehension, behavior and experience.

3.3.1 Code Comprehension Measurements

We measure code comprehension in exactly the same way as proposed by Ceccato et al. [12]:

- *Correctness* (measured per program) is the number of correctly solved tasks: 0 if no task is solved correctly, 1 if precisely one task is solved correctly and 2 if both tasks are correct.
- *Time correct* (measured per program in minutes) is the time spent on average for correctly solving tasks for a program. It is computed as the sum of times spent on correctly solved tasks divided by the number of correctly solved tasks. If no answer was given correctly, the participant was taken out of the calculations.
- *Total time* (measured per program in minutes) shows how long a participant worked on the program, independently on the correctness of solutions. Although this is not a

code comprehension variable by itself, we use it to derive the notion of efficiency below.
- *Efficiency* (measured per program) is *Correctness* divided by *Total time*.

For the tasks that ask to point out a line number where a certain action is performed ("Race:Laps" and "Chat:Users" in Table 2), we evaluated the *Correctness* of the participants' answers in a different way than Ceccato et al. [12]. Whereas they accepted only one specific line number as correct answer, we have adopted a less restrictive interpretation that allowed the following solutions:

- The exact line according to Ceccato et al. [12].
- The line number of the corresponding function header, or the lines interval of the whole corresponding function.
- The exact line of the corresponding function's call site.

We think that all three answers provide a sufficient proof of the participant's understanding of the code functionality. In the hypotheses testing in the sequel, we consider Ceccato et al.'s evaluation for comparison. Two other tasks ("Race:Box" and "Chat:Messages") were evaluated exactly as by Ceccato et al.

3.3.2 Behavior Measurements

To record actions performed by participants during code analysis, we use the Eclipse plugin Fluorite [24] that creates an XML-log of all commands and events with the corresponding timestamps. This data allows us to reconstruct the reversing procedure of each participant with high precision. We extract the following information from the logs:

- The number of the *file open* operations, which correspond to either opening a new file or switching the focus to the already opened one.
- The number of executed *advanced commands* such as automatic identifier renaming, construction of call graphs and type hierarchies.[5]
- The number and the total time of program *executions*.
- The number of times and the total time of the program being in *debugging mode*.
- The total time of *code reading*, which is defined as the overall processing duration for the given program minus the execution and debugging time.

For each action, i.e., program execution, debugging, file open and advanced command, the start and the end timestamp relative to the begin of the program processing are used.

3.3.3 Experience Measurements

Cecatto et al. evaluated the experience of the participants based on whether they were bachelor, master or PhD students. They argued that this is a reliable measure since the authors were in charge of the participants' courses at the corresponding universities [12]. For our study we assume that the participants might have studied at different universities before. Moreover, the attended courses can greatly differ at our university due to different study programs. Further, Acar et al. [18] found that even differentiating between students and non-students showed no significant differences in their participants' skills. We therefore evaluated experience using a more general explorative approach.

[5]Advanced commands have the command ids starting with `org.eclipse.jdt.ui.edit.text.java`.

Individual differences in programming skills, programming experience or experience in dealing with obfuscated code may influence the performance of participants and their analysis behavior. *Experience* relates to the hypotheses HC3 and HB3 and is measured as follows:

- *Programming Experience* is measured on a scale from 1 to 4 using the following question in the pre-study questionnaire: *"How would you describe the quality and the type of the code you wrote so far?"* This question originates from Ceccato et al. [12] and has the following answer options:
  1. Few and small programs (e.g., course exercises)
  2. Many small programs
  3. Small programs and 1 or 2 big programs (e.g., thesis and projects)
  4. Big programs
- *Study-relevant Experience* refers to the experience and knowledge in code obfuscation, Java, the usage of Eclipse for software development, debugging software, the usage of Eclipse for debugging software. These factors are measured using questions *"Please indicate your experience with ..."* in the pre-study questionnaire on a 5-point Likert scale with values from 1 = *very low* to 5 = *very high*;
- *Comprehension Skills* are measured by considering the efficiency of a participant when working with *Clear* code.

4. METHOD

In this section we outline study materials and design, including ethical considerations, and describe recruitment and demographics of the participants. Finally, data analysis techniques are presented.

4.1 Study Materials

4.1.1 Code and Questionnaires

Ceccato et al. [12] provided us with original .jar-files for the clear code of the Chat and Race programs used in their studies. We obfuscated the source code of both programs (Chat and Race) either with name overloading (NO) or with opaque predicates (OP) using the SandMark tool [25] which was reportedly also used in previous work.[6] The resulting three .jar files were decompiled using JAD [26], leading to three source code versions of each program: two obfuscated versions (NO and OP) and the unobfuscated original version. These were used by the participants in our study.

We used the questionnaires by Ceccato et al. [12] that were slightly adapted for our study. For example, we did not ask the participants to estimate the number of code executions per task, since we could measure this in our setup. The questions asked in the survey, their order and under which circumstances they were presented to the participants can be found in Appendix B.

4.1.2 Technical Setup

The technical setup of our study was designed to be especially easy and efficient to replicate. We prepared virtual machines equipped with the Eclipse IDE for analyzing the

[6]While Ceccato et al. [12] claim to have investigated the effect of *identifier renaming* (IR) using SandMark, SandMark does not explicitly offer this obfuscation method. So while we were able to reproduce the obfuscated version of OP, we could not reproduce the code for IR. We therefore chose the "closest" obfuscation variant to identifier renaming provided by SandMark which was *name overloading* (NO).

programs and with the Firefox browser for filling out the online questionnaires. All questionnaires and the code comprehension tasks were combined into one online questionnaire that was developed with LimeSurvey[7]. Participants therefore did not have to change the medium they work on. This also ensured that the participants did not forget to answer the questions, as they could not proceed to the next task otherwise. We were also able to take more precise time measurements than the previous work [12], where the participants filled in the questionnaires on paper and wrote down start and end time of each task.

4.2 Participants

The participants were recruited at an engineering department of a German university. The recruiting materials (flyers, posters and emails) required the participants to have at least basic knowledge of Java and Eclipse.

In total 76 participants took part in our study (8 female). For the evaluation, data of 10 participants were excluded from the analysis because they indicated in the survey that they did not have enough time to successfully complete all tasks. This leaves a total of 66 participants. Most of them (44) were bachelor students, 20 master and 2 PhD students. Ages ranged from 18 to 31 with an average of 22 years.

Most participants were studying computer science (40), followed by computational engineering (4) and medical engineering (4). Furthermore, 16 participants (24.2%) stated that they already participated in a course related to software obfuscation, 7 participants stated that they already worked full-time as a programmer. Part-time working experience was reported by 16 participants.

Concerning previous coding experience, 34 participants (51.5%) stated that they already wrote one or two big programs. The two groups who either only worked on few small programs (19.7%) or on many small programs (21.2%) were almost equally represented. Participants with high experience in big programs made up 7.6% of the participants.

4.3 Study Design

4.3.1 Experimental Setup

Our experimental setup is slightly different from Ceccato et al. [12]. The main differences are summarized in Table 3. Whereas in their work, each participant attended two sessions on two different days in order to reduce the fatigue effects, we opted for having only one session per participant, because a simplified study design allowed us to recruit more participants and thus obtain more results for robust statistical analysis.

To reduce the fatigue effects in our study, we reduced the number of tasks on which each participant worked. For each program, the participants worked on the two comprehension tasks from the original study (Table 2). The two additional change tasks given by Ceccato et al. [12] were omitted.

Moreover, Ceccato et al. [12] used the *within subjects design* [27, 28] where each participant worked on all tasks for a particular study. For example,when they compared between clear code and OP, all 16 participants worked on clear and OP-obfuscated code. In the study where the influence of OP and NO were compared, all participants performed tasks on programs obfuscated with NO as well as with OP. This design is especially useful for small numbers of participants.

[7] https://www.limesurvey.org

	Ceccato et al. [12]	this paper
Sessions per participant	2	1
Number of tasks per program	4	2
Participants (Clear vs NO)	10 and 22[1]	31
Participants (Clear vs OP)	16	35
Participants (NO vs OP)	13 and 13[1]	66
Participants (total)	74	66

Table 3: Experimental setups by Ceccato et al. versus this work. Due to different study designs (*within subjects* [12] versus *between subjects* in this work), data of all our participants (66) could be used for comparison of NO- versus OP-obfuscated code. [1] Two separate studies were conducted.

We opted for the *between subjects design* when comparing the performance of participants working on NO-obfuscated code with the performance of different participants working on OP-obfuscated code. For robust statistical analysis, between subjects design needs a higher number of participants. However, we let all participants first work on the clear code, because we decided to assess their level of expertise in program understanding in this way (see Section 3.3.3). This measurement of expertise should therefore be free from fatigue effects. This study design also lets us compare performance on non-obfuscated code with performance on obfuscated code for each participant (i.e., *within subjects*).

4.3.2 Groups and Tasks

The overall study design is presented in Table 4. The participants were randomly assigned to one of the four experimental groups. Each participant first worked on the clear code of one program, and then on the code of the other program obfuscated with NO or OP. For each program, the participant had to solve two tasks that are presented earlier in Table 2. The tasks were presented in the randomized order.

4.3.3 Procedure and Ethics

The study received approval by the data protection office of the Friedrich-Alexander-Universität Erlangen-Nürnberg. Participants worked under anonymous IDs and were informed at the beginning of their session about data collected during the experiment. We also explained that our goal is not to test their individual performance, but to understand in general how people work on various code comprehension tasks.

We conducted 14 sessions with 7 participants per session on average. Each session lasted 90 minutes, but the participants could leave earlier. In particular, if participants found the tasks too demanding, they could quit and were nevertheless fully paid. They received a 10 EUR gift voucher for participation. On average they worked for 47 minutes.

Each session started with a short presentation by the same researcher using the standardized set of slides. First, the purpose of software obfuscation was introduced, then the procedure was explained. The screenshots of the two programs were included, to make the participants familiar with

Group	1st Program (clear code)	2nd Program (obfuscated)
1	Race: *Rnd*(Box,Laps)	NO(Chat): *Rnd*(Messages,Users)
2	Race: *Rnd*(Box,Laps)	OP(Chat): *Rnd*(Messages,Users)
3	Chat: *Rnd*(Messages,Users)	NO(Race): *Rnd*(Box,Laps)
4	Chat: *Rnd*(Messages,Users)	OP(Race): *Rnd*(Box,Laps)

Table 4: Groups and tasks. Each user first worked on the clear code of one program, and then on the NO- or OP-obfuscated code of another program. *Rnd* denotes the randomization of task order within each program.

the programs. One or two additional researchers (depending on the number of the session participants) were in the lab to ensure the smooth execution of the experiment.

After the presentation, the participants logged into the virtual machine with their anonymized participant ID. There, they opened Firefox and started filling out the online survey. After answering the pre-study questionnaire, they were shown a password that they entered to unzip the zip-file with the program code. By entering the password, Eclipse was automatically set up with the corresponding source code (unobfuscated for the first program) according to the group the participants belonged to. Also, the logging of all events and timings in Eclipse started.

Back in the online survey, a description of the the first program was shown. On the next page of the survey the first task was presented and the solution had to be filled in. When the first task was successfully completed, the survey asked the post-task questions. Next, the second task was presented in the survey. After finishing this task, participants where asked to close Eclipse. By doing so, a log-file with all events in Eclipse was sent to our server. Participants then filled out post-task questions again. Furthermore, the post-program questions were asked. Then the password for the second program was shown and the same procedure was repeated for the second (obfuscated) program.

4.4 Data Analysis

Statistical analysis was performed using SPSS [29]. For all tests, a significance level of $\alpha = 0.05$ was employed.

4.4.1 Effect of Code Obfuscation

To compare code comprehension and code analysis behavior for clear and for obfuscated code, we used Wilcoxon signed-rank tests (within subjects design). To compare both obfuscation methods with each other, we used Mann-Whitney U tests (between subjects design). Non-parametric tests were used because the assumption of normal distribution was violated for most variables (as indicated by Shapiro-Wilk and Kolmogorov-Smirnov tests).

4.4.2 Impact of Experience

Experience was assessed with three measures: *Programming Experience*, *Study-relevant Experience*, and *Comprehension Skills* (Section 3.3.3). We first analyzed the five questions of *Study-relevant Experience*. With a factor analysis, we extracted two factors with eigenvalues larger than 1 (Kaiser Guttman criterion). These two factors explained 82% of the variance in the data. Table 5 shows the factor loadings after varimax rotation. Factor 1 summarizes experience with obfuscated code and debugging and Factor 2 encompasses experience with Java and Eclipse. Individual experience levels

	Factor 1	Factor 2
Code obfuscation	.921	
Debugging software	.798	
Java		.773
Eclipse for software development		.940
Eclipse for debugging		.925

Table 5: Factor loadings after varimax rotation. Values below 0.4 are omitted.

	Progr. Exp.	Obfusc. Exp.	Java Exp.	Compr. Skill[1]
Obfus.Exp.	0.648**			
Java Exp.	0.466**	0.298*		
Compr. Skill[1]	0.323**	0.411**	0.097	
Compr. Skill[2]	0.264*	0.326**	0.140	0.945**

Table 6: Correlations between experience indicators; [1]our measurement, [2]strict measurement (Ceccato et al.); *p<.05, **p<.01.

were computed by averaging across the respective questions. In summary, we consider four indicators of experience:

- *Programming Experience*: quality and type of code written so far;
- *Obfuscation Experience*: experience with obfuscation and debugging;
- *Java Experience*: experience with Java and using Eclispe;
- *Comprehension Skills*: efficiency in working on clear code.

The four indicators were moderately correlated with each other (see Table 6), indicating that they can be integrated to measure individual levels of experience.

On the basis of the four experience indicators, we divided participants into experience groups using a data-driven approach. We ran a cluster analysis, which tries to identify homogeneous groups of cases, such that observations in the same group are as similar as possible, and observations in different groups are as different as possible. A k-means cluster analysis was performed, setting the parameter k to the value 2 to extract two groups of experience. The final groups, "Beginners" ($N = 21$) and "Experienced" ($N = 45$), differed significantly in all four indicators, all F's$(64) > 5.952$, p's< 0.018 (see Table 7).

To assess the moderating effect of experience on code comprehension and code analysis behavior, mixed-model Analyses of Variance (ANOVA) were run with Obfucation (Clear vs. Obfuscated Code) as within subjects factor and Ex-

	Beginners $N = 21$	Experienced $N = 45$
Programming Exp.	1.42 ± 0.60	2.96 ± 0.52
Obfuscation Exp.	1.55 ± 0.44	2.99 ± 0.73
Java Exp.	2.30 ± 0.60	3.13 ± 0.80
Compr. Skill	0.07 ± 0.06	0.18 ± 0.20

Table 7: Description of the experience groups (Mean±SD).

perience (Beginners vs. Experienced) between subjects factor.[8] Effects of obfuscation (irrespective of experience) are reflected by the main effect Obfuscation. Similarly, effects of experience (irrespective of the type of code) is reflected in the main effect Experience. Whether experience moderates the obfuscation effect (i.e., whether beginners and experts differ in working with obfuscated code) is reflected by the interaction between Obfuscation and Experience. If the interaction was significant, we run post hoc t-tests in order to compare beginners and experienced programmers when working with obfuscated code.

4.4.3 *Effect Sizes and Statistical Power*

To assess the practical meaning of the empirical results, we calculated effect sizes. For Wilcoxon signed-rank tests and Mann-Whitney U tests, we report r. For ANOVAs we report partial eta-squared (η_p^2). For unpaired t-tests, we report Cohen's d. For paired t-tests, we report Cohen's d_z, which corrects the effect size for correlations in a within-subjects design. However, both Cohen's d and η_p^2 can be greater than 1, making an intuitive interpretation difficult. Therefore, we also report ω^2, which ranges between 0 and 1. It can be interpreted as the percentage of variance in the data that is explained by the experimental manipulation. For interpretation, we followed the convention provided by Cohen [30].

Interpretation	Cohen's d & d_z	r	η_p^2 and ω^2
no effect	< 0.20	<0.10	<0.01
small effect	0.20-0.50	0.10-0.30	0.01-0.06
medium effect	0.50-0.80	0.30-0.50	0.06-0.14
large effect	> 0.80	>0.50	>0.14

Table 8: Interpretation of effect sizes.

We assume that effects indicate practical relevance if they are of at least medium size (Table 8). A power analysis showed that we were able to detect such an effect in the population with a probability of $\beta = 0.80$ in a within subjects design with a sample of $N = 35$ participants (i.e., running a Wilcoxon test) and in a between subjects design with a sample of $N = 134$ (i.e., running a Mann-Whitney U test). Referring to the actual number of participants (Table 3), most of our tests (apart from Clear vs OP) are underpowered, meaning that we might have missed some effects due to small sample size.

[8] Although the assumption of normal distribution has been violated for most variables, to our knowledge, there is no valid non-parametric equivalent to a two-way ANOVA implemented in our analysis tool SPSS. For example, the Kruskal-Wallis test can be used as non-parametric equivalent to the one-way ANOVA. However, as we are interested in the interaction between two factors, i.e. Obfuscation and Experience, the test is not valid in our case.

		Evaluation method *This paper*	*Ceccato et al.*
Race	Box	78.8% (52/66)	78.8% (52/66)
	Laps	78.8% (52/66)	54.5% (36/66)
Chat	Messages	57.6% (38/66)	57.6% (38/66)
	Users	31.8% (21/66)	18.2% (12/66)

Table 9: Task correctness rates when evaluating the results with our evaluation method versus with the stricter rules by Ceccato et al. (Section 3.3.1).

5. RESULTS

We present our results and, if applicable, compare them with the findings of Ceccato et al. [12]. We start with descriptive results (Section 5.1), and then analyze differences between clear and obfuscated code with regard to code comprehension and analysis behavior (Sections 5.2, 5.3 and 5.4). The results of these evaluations are summarized in Table 10. Finally, we assess the moderating effect of experience (Section 5.5 and Table 11).

5.1 Descriptive Results

Correctness results are presented in Table 9. Each of the 66 participants worked on four tasks, two with clear and two with obfuscated code. Using our less strict evaluation of all 264 solutions (see Section 3.3.1), 163 were rated correct and the remaining 101 were false. Using the more strict evaluation by Ceccato et al., our participants scored 138 correct and 126 false answers. In both cases, "Chat: Users" was the most difficult task, and "Race: Box" the easiest one.

The fastest participant took 21 minutes, the slowest finished after 90 minutes. For the Chat program 90.9% and for the Race program 95.5% of the participants agreed or strongly agreed that the descriptions of the application was clear.

5.2 Name Overloading ($HC1_{NO}$ & $HB1_{NO}$)

Tasks with clear and obfuscated code were solved with similar correctness, $T(31) = 91.50$, $p = 0.373$, $z = -0.892$, $r = -0.113$ (Table 10). To show the same level of correctness with obfuscated code, participants needed significantly longer, $T(31) = 384.00$, $p = 0.008$, $z = 2.665$, $r = 0.338$. This speed-accuracy trade-off was reflected in a significant effect on efficiency, $T(31) = 104.00$, $p = 0.014$, $z = -2.454$, $r = -0.312$. Time needed to correctly solve a task, i.e., a successful attack, was significantly longer for obfuscated code, $T(21) = 181.00$, $p = 0.023$, $z = 2.277$, $r = 0.351$.

Using stricter correctness by Ceccato et al. [12], we also found no difference concerning the correctness of code comprehension between clear and obfuscated code, $T(31) = 67.50$, $p = 0.648$, $z = -0.456$, $r = -0.058$. The effect of NO on efficiency was significant, $T(31) = 106.00$, $p = 0.046$, $z = -1.994$, $r = -0.253$. The time to correctly solve tasks showed no difference between the groups, $T(20) = 152.00$, $p = 0.079$, $z = 1.755$, $r = 0.277$. However, our sample size was not sufficient to detect effects of medium size (Section 4.4.3), such that we might have missed some effects.

	Descriptive Results						Parameter-free tests		
	Clear		NO		OP		*Clear*	*Clear*	*NO*
Measurement	*Median*	*IQR*	*Median*	*IQR*	*Median*	*IQR*	*vs NO*	*vs OP*	*vs OP*
Correctness	2.000	1.000	2.000	1.000	1.000	1.000	91.50	80.50	496.00
Efficiency	0.130	0.161	0.096	0.079	0.090	0.057	104.00*	125.00**	571.00
Total time	13.163	11.404	18.154	9.028	15.986	11.551	384.00**	456.00*	444.00
Time correct	5.758	5.595	8.888	3.817	6.167	6.702	181.00*	166.00	216.00
Strict correctness as measured by Ceccato et al.:									
Correctness	1.000	1.000	1.000	1.000	1.000	1.000	67.50	77.00	492.00
Efficiency	0.111	0.105	0.063	0.082	0.082	0.052	106.00*	161.00	570.00
Time correct	5.439	6.079	8.874	3.817	5.951	7.012	152.00	156.00	207.00
Number of:									
File open commands	13.000	19.000	30.000	30.000	19.500	19.250	429.00**	344.50	307.00**
Advanced commands	0.000	3.000	1.000	11.000	1.000	4.250	121.50**	126.00*	479.50
Program executions	1.000	3.000	3.000	5.000	2.000	2.000	216.50	376.00**	478.00
Debugging mode	0.000	1.000	0.000	3.000	2.000	5.000	105.00**	138.00**	560.00
Time spent on:									
Program executions	0.383	1.400	1.317	4.467	1.025	1.504	265.50*	388.50*	534.00
Debugging mode	0.000	0.450	0.000	10.117	3.000	11.292	120.00**	131.00*	496.00
Code reading	10.500	9.533	13.683	9.633	10.700	8.004	280.00	323.00	491.50

Table 10: Descriptive results and parameter-free statistics (Wilcoxon & Mann-Whitney-U tests) comparing clear and obfuscated code; all times are in minutes; *p<.05, **p<.01.

Reduced efficiency and increased total time may be due to changes in code analysis behavior. Participants opened files more frequently, $T(31) = 429.00$, $p < 0.001$, $z = 3.548$, $r = 0.451$, used more advanced commands, $T(31) = 121.50$, $p = 0.006$, $z = 2.771$, $r = 0.352$, and more often the debugging mode, $T(31) = 105.00$, $p = 0.001$, $z = 3.311$, $r = 0.420$. Overall they spent more time with program executions, $T(31) = 265.50$, $p = 0.022$, $z = 2.286$, $r = 0.290$, and debugging, $T(31) = 120.00$, $p = 0.001$, $z = 3.408$, $r = 0.433$. The observed effects were of medium size.

In summary, obfuscating source code with NO significantly reduced the efficiency of code comprehension ($HC1_{NO}$). Participants changed their code analysis behavior ($HB1_{NO}$), i.e., they opened files more frequently, used more advanced commands, and the debugging mode. The observed effects were of medium size, indicating their practical importance. The behavior of participants corresponds to what can be expected when dealing with NO since the inverse transformation to NO (rename identifier) is an advanced command in Eclipse. Other increases can be explained by additional effort to understand the meaning of individual identifiers.

5.3 Opaque predicates ($HC1_{OP}$ & $HB1_{OP}$)

If the source code was obfuscated with opaque predicates, we observed similar effects (Table 10). Participants needed significantly longer to understand the code, $T(35) = 456.00$, $p = 0.021$, $z = 2.309$, $r = 0.276$, in order to reach about the same level of correctness, $T(35) = 80.50$, $p = 0.198$, $z = -1.286$, $r = -0.154$. This is reflected in reduced efficiency, $T(35) = 125.00$, $p = 0.005$, $z = -2.778$, $r = -0.332$. Concerning the time needed to correctly solve a task, i.e., a successful attack, no difference between clear and obfuscated source code was found, $T(22) = 166.00$, $p = 0.200$, $z = 1.282$, $r = 0.193$.

Again, using the stricter measurements by Ceccato et al. [12], participants reached about the same performance in terms of correctness, $T(35) = 77.00$, $p = 0.980$, $z = 0.025$, $r = 0.004$, and were marginally less efficient, $T(35) = 161.00$, $p = 0.054$, $z = -1.926$, $r = -0.326$.

The impact of code obfuscation was also visible in code analysis behavior. Participants more often used advanced commands, $T(35) = 126.00$, $p = 0.018$, $z = 2.359$, $r = 0.282$, executed the program more frequently, $T(35) = 376.00$, $p = 0.003$, $z = 2.979$, $r = 0.356$, executed code longer in total $T(35) = 388.50$, $p = 0.020$, $z = 2.328$, $r = 0.278$, used the debugging mode more often, $T(35) = 138.00$, $p = 0.003$, $z = 2.923$, $r = 0.349$, and spent more time debugging, $T(35) = 131.00$, $p = 0.010$, $z = 2.580$, $r = 0.308$.

In summary, obfuscating source code with OP significantly reduced the efficiency of code comprehension ($HC1_{OP}$). Participants changed their analysis behavior ($HB1_{OP}$), i.e., they used more advanced commands, executed the program more frequently, and used the debugging mode more often. The observed effects were of medium size, indicating their practical importance. Compared to the changes with NO, the differences in behavior between Clear and OP appear to be more random which can be interpreted as an unguided search for understanding.

5.4 Comparison of Obfuscation Methods (HC2 & HB2)

The previous analyses showed that both obfuscation methods, name overloading and opaque predicates, significantly reduced code comprehension performance. To achieve a similar level of comprehension, participants changed their behavior of code analysis. A direct comparison between both obfuscation methods indicates that code comprehension was hindered similarly, i.e., we found no differences in correctness, total time or efficiency. Also, the effect on time needed

to correctly solve a task, i.e., time for a successful attack, was small and non-significant, $U(50) = 216.00$, $p = 0.066$, $z = -1.839$, $r = -0.260$ (Table 10).

Concerning behavior, participants opened files significantly more frequently when the source code was obfuscated with NO than with OP, $U(66) = 307.00$, $p = 0.002$, $z = -3.027$, $r = -0.373$. This effect is of medium size.

In summary, both obfuscation methods reduced efficiency of code comprehension and led to similar behavior of the participants in almost all aspects. The number of file openings being higher in NO could be due to the fact that the structure of the code is not changed by the transformation and thus many aspects of semantics remain. The main effort is to deduce useful meanings of identifiers using static and dynamic analysis techniques. We would have expected a significant difference in using advanced commands for NO than OP due to the expected higher use of the advanced command "rename identifier", but the usage of this particular primitive does not appear to be different in the data. We note, however, that our sample size was too small for a between subjects comparison, such that we might have missed some effects (Section 4.4.3).

5.5 Impact of Experience (HC3 & HB3)

Here we assess whether experience moderates code comprehension and code analysis behavior. We performed ANOVAs with Obfuscation (clear vs. obfuscated) as within subjects factor and Experience (beginners vs. experienced) as between subjects factors (see Section 4.4.2). As the main effect of Obfuscation replicates the results reported before, we only report the main effect of Experience and the interaction between Obfuscation and Experience here. Descriptive results and inferential statistics are presented in Table 11.

5.5.1 General Effect of Experience

The difference between beginners and experienced programmers, irrespective of the type of code, is reflected in the main effect of Experience. Beginners and experienced participants spent about the same time to solve the tasks (15.8 minutes vs. 17.6 minutes), $F(1, 64) < 1$. As experienced participants solved about 1.4 tasks whereas beginners solved only 0.8 task in this time correctly, $F(1, 64) = 13.907$, $p = 0.001$, $\eta_p^2 = 0.18$, $\omega^2 = 0.16$, their efficiency was significantly higher, $F(1, 64) = 8.008$, $p = 0.006$, $\eta_p^2 = 0.10$, $\omega^2 = 0.10$. The effects were of medium size, explaining 10% to 16% of the variability in the data. The same results were observed for the strict correctness of Ceccato et al. [12].

Beginners and experienced programmers showed different code analysis behaviors. Experienced participants executed advanced commands ten times more often than beginners, $F(1, 64) = 11.157$, $p < 0.001$, $\eta_p^2 = 0.15$, $\omega^2 = 0.13$, and used the debugging mode eight times more often, $F(1, 64) = 11.252$, $p = 0.001$, $\eta_p^2 = 0.15$, $\omega^2 = 0.13$. This was also visible in the overall time they spent in debugging mode, $F(1, 64) = 4.531$, $p = 0.037$, $\eta_p^2 = 0.07$, $\omega^2 = 0.05$. The latter effect was small, the other effects were of medium size.

In summary, experienced programmers solved 36% more tasks correctly in about the same time as beginners, which was reflected in a higher efficiency. To analyze the code, experienced participants used advanced commands and the debugging mode more often, which is consistent with the expected behavior of experienced reverse engineers.

5.5.2 Experience as Moderator of Comprehension

Solving tasks with obfuscated code requires more time to keep the level of correctness, i.e., efficiency is lower. Moreover, working on obfuscated code requires a change in code analysis behavior (see Sections 5.2 and 5.3). We are now interested in whether beginners and experienced programmers show similar or different changes. Statistically, this effect is reflected by the interaction between Obfuscation and Experience in the ANOVAs.

With regard to code comprehension, experience moderated the obfuscation effect on efficiency significantly, $F(1, 64) = 4.385$, $p = 0.040$, $\eta_p^2 = 0.05$, $\omega^2 = 0.05$. Beginners' efficiency did not significantly change when working on obfuscated code (0.06 tasks per minute) compared to clear code (0.07 tasks per minute), $t(20) < 1$. In contrast, experienced programmers were significantly more efficient with clear code (0.19 tasks per minute) than with obfuscated code (0.08 tasks per minute), $t(44) = 3.499$, $p = 0.011$, $d_z = 0.68$. When working with obfuscated code, their efficiency almost dropped to those of beginners (i.e., 0.08 vs. 0.07 tasks per minute). This difference between beginners and experienced programmers was statistically not significant, $t(64) = 1.387$, $p = 0.174$, $d = 0.37$. One may argue that the statistical power was not sufficient to detect the effect. Indeed, the small effect size $d < 0.50$ indicates that there is probably a small effect in the population. That is, programming experience may have an advantage for comprehending obfuscated code efficiently but this advantage is probably only minor (see Figure 1). This issue needs further investigation with a more appropriate sample size.

This drop in efficiency for experienced participants, $F(1, 64) = 1.609$, $p = 0.209$, $\eta_p^2 = 0.03$, $\omega^2 = 0.01$, might be due to an increase in total time in order to keep a similar level of correctness, $F(1, 64) < 1$. That is, experienced programmers invested more time to keep a high level of correctness, whereas beginners did not. However, the effects were of quite small size and the sample size was insufficient to detect these effects statistically. Also, we were not able to replicate them using the strict measurements by Ceccato et al. [12], although they point into the same direction.

When working on obfuscated code compared to clear code, experienced programmers changed their code analysis behavior, whereas beginners did not (Figure 1). This moderating effect of experience occurred for the usage of advanced commands, $F(1, 64) = 5.321$, $p = 0.024$, $\eta_p^2 = 0.08$, $\omega^2 = 0.06$, and the usage of the debugging mode, $F(1, 64) = 7.615$, $p = 0.008$, $\eta_p^2 = 0.11$, $\omega^2 = 0.09$. All effects were of medium size, explaining 6% to 9% of the variability in the data.

In summary, code comprehension and analysis behavior of *beginners* was not much impacted by obfuscated code. As expected, *experienced programmers* were more efficient when working with clear code. However, code obfuscation impeded code comprehension. The efficiency dropped by 57% to those of beginners. In our view, this is the most interesting result of our study. To keep the higher level of correctness compared to beginners, experienced programmers invested more time to solve the tasks. They changed their code analysis strategies, i.e., they used more often advanced commands and the debugging mode. Overall they spent more time with reading the source code. It appears

	Descriptive Results								ANOVA		
	Beginner				Experienced						
	Clear		Obfucated		Clear		Obfuscated				
Measurement	Mean	SD	Mean	SD	Mean	SD	Mean	SD	*Obf.*	*Exp.*	*Obf.* × *Exp.*
Correctness	0.90	0.77	0.86	0.85	1.53	0.66	1.27	0.72	1.377	13.907**	0.669
Efficiency	0.066	0.059	0.057	0.060	0.185	0.202	0.079	0.059	6.185*	8.008**	4.385*
Total Time	14.4	7.6	17.3	11.8	13.9	9.1	21.3	10.8	8.917**	0.815	1.609
Time correct	8.4	3.9	10.4	8.9	6.3	5.2	9.5	5.9	3.183	0.768	0.170
Strict correctness as measured by Ceccato et al.:											
Correctness	0.62	0.74	0.71	0.78	1.27	0.62	1.18	0.68	0.001	16.213**	0.604
Efficiency	0.049	0.061	0.048	0.057	0.151	0.188	0.075	0.059	3.061	8.145**	2.925
Time correct	10.0	7.7	7.6	2.8	6.3	5.4	9.1	5.9	0.015	0.329	2.743
Number of:											
File open comm.	18.7	18.1	23.0	20.5	17.4	16.7	30.2	17.2	6.719*	0.752	1.641
Advanced comm.	0.3	0.8	0.4	0.8	2.3	3.4	5.4	7.1	5.662*	11.157**	5.321*
Program exec.	0.8	0.9	2.6	2.6	2.5	4.0	4.5	6.7	6.286*	3.225	0.021
Debugging mode	0.1	0.3	0.4	0.9	0.9	1.9	3.0	3.4	13.368**	11.252**	7.615**
Time spent on:											
Program exec.	0.8	1.6	3.2	4.4	1.5	2.5	3.4	7.7	5.800*	0.169	0.081
Debugging mode	0.1	0.2	3.1	9.2	1.7	4.5	7.0	8.4	12.793**	4.531*	1.019
Code reading	14.0	8.7	12.1	8.0	11.0	8.1	13.8	9.1	0.064	0.157	2.110

Table 11: Descriptive results and inferential statistics comparing clear and obfuscated code for beginners and experienced programmers; all times are in minutes; *p<.05, **p<.01.

that classical programming experience does not help much in comprehending obfuscated source code.

5.5.3 Exploration: Areas of Experience

The previous analysis shows that no evidence was found that experienced programmers, who were much more efficient with clear code than beginners, differ from the latter when the source code was obfuscated. In the following we explore whether experience in a particular area may prevent from the drop in efficiency.

We correlated the level of experience (in one of the four areas of experience we had measured, see Sectionsec:analysis-experience) with the efficiency in working with obfuscated code (*Pearson* correlation). A positive correlation indicates that more experienced participants were able to keep a higher level of efficiency. Maybe not surprisingly, this was indeed the case if participants had *experience with obfuscated code and debugging* before, $r = 0.43$, $p < 0.001$. Neither programming experience in general, $r = 0.14$, $p = 0.271$, experience with Java and Eclipse, $r = 0.08$, $p = 0.544$, nor comprehension skills (measured as efficiency in working on clear code), $r = 0.16$, $p = 0.191$, did help. We conclude that reverse engineering needs special training in obfuscation techniques.

6. DISCUSSION

Before this study, we knew that obfuscation impeded program comprehension [12]. We were able to reproduce these findings for the same obfuscation methods, NO and OP, that were previously studied. We also obtained original results by studying the reverse engineering behavior. As might be expected, we found many significant differences in behavior between clear and obfuscated versions. These differences appeared to be more intentional for NO than OP. Participants appeared to have a clear strategy in countering NO but were still inhibited severely regarding efficiency. Given OP-obfuscated code, the analysis behavior appeared to be more random both in numbers of commands and time spent on different activities. With such a behavior, a decrease in efficiency is an understandable consequence.

Overall, the different behaviors for NO and OP are a first empirical support of the taxonomy of Collberg et al. [6] who distinguished obfuscating transformations regarding resilience and potency. For NO we found significant decreases in efficiency despite clear and understandable adaptions in behavior by participants. Such a strategic behavior change was not observable with OP. NO therefore appears to belong to the class of potent obfuscation techniques, increasing the "obscurity" for the human reverse engineers.

Furthermore, obfuscation seems to "reduce experience", i.e., the effect of software engineering experience on the success of program comprehension is much lower for obfuscated code than for unobfuscated code. This insight is important since it indicates that code comprehension in the realm of obfuscated software may be different from comprehension of traditional programs. We conjecture that comprehension strategies follow a two-step approach: in the first step the particular obfuscation method is identified; in a second step, an inverse transformation is attempted. Such a strategy can, however, only be applied if reverse engineers have (1) an understanding of different obfuscation techniques, and (2) the ability to inverse the obfuscation using ingenuity and/or tools. This is consistent with the findings of our experiment where understanding of obfuscation methods was more helpful than general programming experience.

7. LIMITATIONS

Because the analyzed programs and tasks could be of different difficulty, we used counterbalancing to mitigate this concern. We did not counterbalance clear and obfuscated code

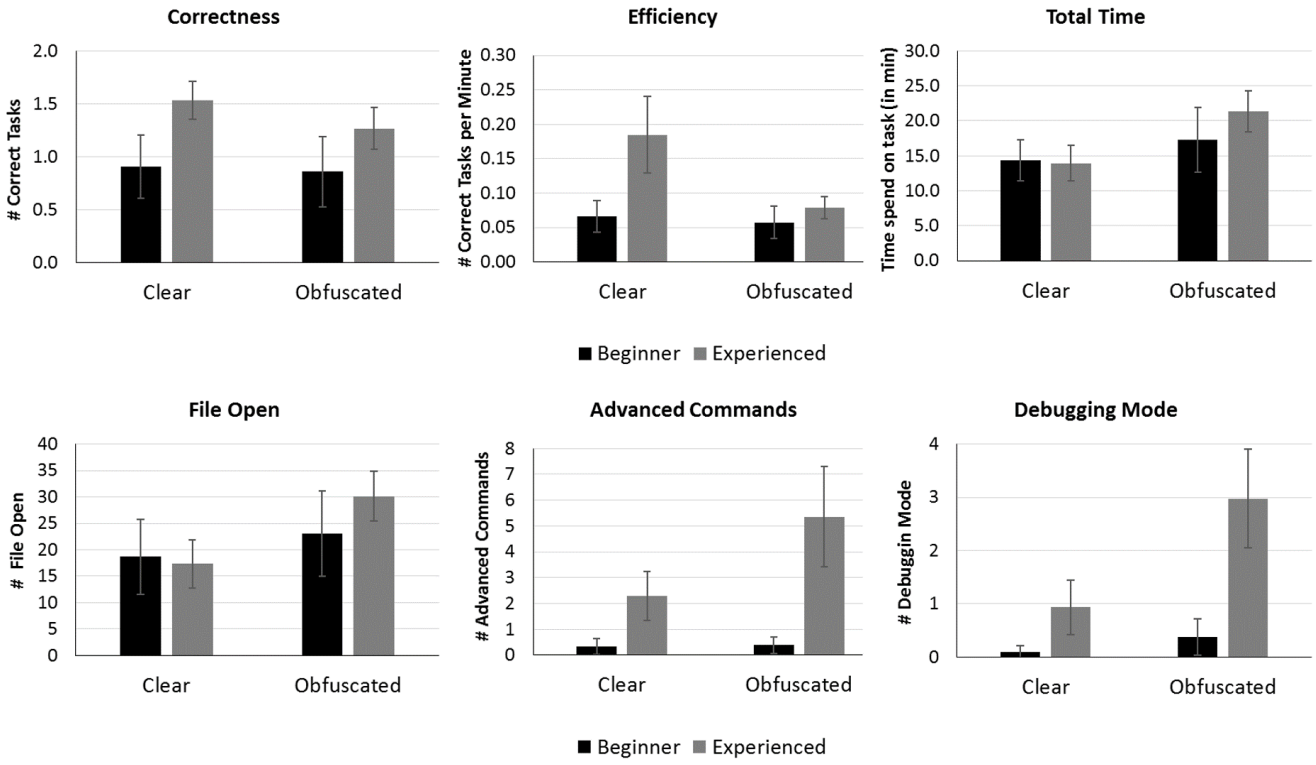


Figure 1: On clear code, experienced users were more efficient than beginners. This advantage of experience disappeared when the code was obfuscated. Experienced users invested more time to keep their level of correctness, but beginners did not. Beginners did not change their code analysis behavior when working on obfuscated code. Experienced users opened files more frequently and used advance commands and the debugging mode more often than when working on clear code. (Error bars indicate confidence intervals.)

tasks, as the participants worked on the clear code first. This was necessary for precise assessment of their comprehension skills. Therefore, learning effects may have positively influenced performance on obfuscated code, such that effects of obfuscation on code comprehension and behavior may actually be stronger than we found. In contrast, fatigue effects could have negatively influenced performance on obfuscated code. To counter this limitation, we analyzed only the data of participants who indicated that they had enough time to perform all tasks.

The sampling of experience was performed post hoc by placing participants into groups and not as a planned sampling based on experience. The representativeness of the sample (students) is limited, although the study by Acar et al. [18] provided evidence that experience may be a more important indicator of expertise than student status. A similar study with professionals using their own analysis equipment and a more realistic scenario (e.g., malware analysis) would be desirable, but would be hard to pursue given the scarcity of obfuscation analysis resources in the professional market.

8. CONCLUSIONS

In this work we measured effects of source code obfuscation on program comprehension skills of reverse engineers by means of a controlled experiment with 66 participants. We successfully replicated results by Ceccato et al. [12] that obfuscation techniques have a significantly negative effect on program comprehension. We also showed that the obfuscation methods NO and OP lead to significantly different analysis behavior. The differences provided insight into the relative strength of NO which withstood reverse engineering efforts, although it was clearly identifiable and the de-obfuscation tool (rename identifer) was available. This supports the distinction between resilience and potency of obfuscating transformations as defined by Collberg et al. [6] more than 20 years ago.

Future research should focus more specifically on the behavior of humans when facing obfuscated code. Do they follow a two-step approach as conjectured above? What if an unknown obfuscation technique or the combination of several is used? How do performance results change for professional malware analysts, or when deobfuscation tools are used?

Acknowledgment

This work was supported by the "Bavarian State Ministry of Education, Science and the Arts" as part of the FORSEC research association. We thank the anonymous reviewers for their helpful comments, and we are indebted to Mariano Ceccato, Brian Glass, Tilo Müller, Yan Zhuang and our shepherd Joseph Bonneau for their invaluable support.

9. REFERENCES

[1] J. Dyer, M. Lindemann, R. Perez, R. Sailer, L. van Doorn, S. W. Smith, and S. Weingart, "Building the IBM 4758 Secure Coprocessor," *IEEE Computer*, vol. 34, no. 10, pp. 57–66, 2001.

[2] U. Piazzalunga, P. Salvaneschi, F. Balducci, P. Jacomuzzi, and C. Moroncelli, "Security strength measurement for dongle-protected software," *IEEE Security & Privacy*, vol. 5, no. 6, pp. 32–40, 2007.

[3] T. Alves and D. Felton, "TrustZone: Integrated hardware and software security," ARM, Tech. Rep., Jul. 2004.

[4] O. Dvir, M. Herlihy, and N. Shavit, "Virtual leashing: Creating a computational foundation for software protection," *J. Parallel Distrib. Comput.*, vol. 66, no. 9, pp. 1233–1240, 2006. [Online]. Available: https://doi.org/10.1016/j.jpdc.2006.04.013

[5] B. Barak, O. Goldreich, R. Impagliazzo, S. Rudich, A. Sahai, S. P. Vadhan, and K. Yang, "On the (im)possibility of obfuscating programs," in *Proceedings of the 21st Annual International Cryptology Conference on Advances in Cryptology*, ser. CRYPTO '01. London, UK, UK: Springer-Verlag, 2001, pp. 1–18. [Online]. Available: http://dl.acm.org/citation.cfm?id=646766.704152

[6] C. Collberg, C. Thomborson, and D. Low, "A taxonomy of obfuscating transformations," Technical Report 148, Department of Computer Science, University of Auckland, Jul. 1997. [Online]. Available: http://citeseer.ist.psu.edu/collberg97taxonomy.html

[7] A. Capiluppi, P. Falcarin, and C. Boldyreff, "Code defactoring: Evaluating the effectiveness of java obfuscations," in *Reverse Engineering (WCRE), 2012 19th Working Conference on*, Oct 2012, pp. 71–80.

[8] M.-A. Storey, "Theories, tools and research methods in program comprehension: past, present and future," *Software Quality Journal*, vol. 14, no. 3, pp. 187–208, 2006.

[9] M. D. Penta, R. K. Stirewalt, and E. Kraemer, "Designing your next empirical study on program comprehension," in *Program Comprehension, 2007. ICPC'07. 15th IEEE International Conference on*. IEEE, 2007, pp. 281–285.

[10] B. Cornelissen, A. Zaidman, A. Van Deursen, L. Moonen, and R. Koschke, "A systematic survey of program comprehension through dynamic analysis," *Software Engineering, IEEE Transactions on*, vol. 35, no. 5, pp. 684–702, 2009.

[11] B. Dit, M. Revelle, M. Gethers, and D. Poshyvanyk, "Feature location in source code: a taxonomy and survey," *Journal of Software: Evolution and Process*, vol. 25, no. 1, pp. 53–95, 2013.

[12] M. Ceccato, M. Di Penta, P. Falcarin, F. Ricca, M. Torchiano, and P. Tonella, "A family of experiments to assess the effectiveness and efficiency of source code obfuscation techniques," *Empirical Software Engineering*, vol. 19, no. 4, pp. 1040–1074, 2014.

[13] A. Viticchié, L. Regano, M. Torchiano, C. Basile, M. Ceccato, P. Tonella, and R. Tiella, "Assessment of source code obfuscation techniques," in *Source Code Analysis and Manipulation (SCAM), 2016 IEEE 16th International Working Conference on*. IEEE, 2016, pp. 11–20.

[14] S. Garfinkel and H. R. Lipford, "Usable security: History, themes, and challenges," *Synthesis Lectures on Information Security, Privacy, and Trust*, vol. 5, no. 2, pp. 1–124, 2014.

[15] D. Oliveira, M. Rosenthal, N. Morin, K.-C. Yeh, J. Cappos, and Y. Zhuang, "It's the psychology stupid: how heuristics explain software vulnerabilities and how priming can illuminate developer's blind spots," in *Proceedings of the 30th Annual Computer Security Applications Conference*. ACM, 2014, pp. 296–305.

[16] Y. Acar, M. Backes, S. Fahl, D. Kim, M. L. Mazurek, and C. Stransky, "You get where you're looking for: The impact of information sources on code security," in *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE, 2016, pp. 289–305.

[17] Y. Acar, M. Backes, S. Fahl, S. Garfinkel, D. Kim, M. L. Mazurek, and C. Stransky, "Comparing the usability of cryptographic apis," in *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE, 2017, pp. 154–171.

[18] Y. Acar, C. Stransky, D. Wermke, M. L. Mazurek, and S. Fahl, "Security developer studies with github users: Exploring a convenience sample," in *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. Santa Clara, CA: USENIX Association, 2017, pp. 81–95. [Online]. Available: https://www.usenix.org/conference/soups2017/technical-sessions/presentation/acar

[19] C. Collberg and J. Nagra, *Surreptitious Software: Obfuscation, Watermarking, and Tamperproofing for Software Protection*, 1st ed. Addison-Wesley Professional, 2009.

[20] M. Ceccato, M. Di Penta, J. Nagra, P. Falcarin, F. Ricca, M. Torchiano, and P. Tonella, "Towards experimental evaluation of code obfuscation techniques," in *Proceedings of the 4th ACM workshop on Quality of protection*. ACM, 2008, pp. 39–46.

[21] M. Ceccato, M. D. Penta, J. Nagra, P. Falcarin, F. Ricca, M. Torchiano, and P. Tonella, "The effectiveness of source code obfuscation: an experimental assessment," in *Program Comprehension, 2009. ICPC'09. IEEE 17th International Conference on*. IEEE, 2009, pp. 178–187.

[22] T. J. McCabe, "A Complexity Measure," *IEEE Transactions on Software Engineering*, vol. SE-2, no. 4, pp. 308–320, Dec. 1976.

[23] B. Yadegari, B. Johannesmeyer, B. Whitely, and S. Debray, "A generic approach to automatic deobfuscation of executable code," in *2015 IEEE Symposium on Security and Privacy, SP 2015, San Jose, CA, USA, May 17-21, 2015*, 2015, pp. 674–691. [Online]. Available: https://doi.org/10.1109/SP.2015.47

[24] Y. Yoon and B. A. Myers, "Capturing and analyzing low-level events from the code editor," in *Proceedings of the 3rd ACM SIGPLAN Workshop on Evaluation and Usability of Programming Languages and Tools*, ser. PLATEAU '11. New York, NY, USA: ACM, 2011, pp. 25–30. [Online]. Available: http://doi.acm.org/10.1145/2089155.2089163

[25] C. Collberg, G. Myles, and A. Huntwork, "Sandmark–a tool for software protection research," *IEEE security & privacy*, no. 4, pp. 40–49, 2003.

[26] P. Kouznetsov, "Jad-the fast java decompiler," *URL: http://www. kpdus. com/jad. html*, 2006.

[27] A. Field and G. Hole, *How to design and report experiments*. Sage, 2002.

[28] G. Leroy, *Designing User Studies in Informatics*. Springer Science & Business Media, 2011.

[29] A. Field, *Discovering statistics using IBM SPSS statistics - 4th edition*. Sage, 2013.

[30] J. Cohen, *Statistical power analysis for the behavioral sciences . Hilsdale*. NJ: Lawrence Earlbaum Associates, 1988.

APPENDIX

A. EXAMPLES OF OBFUSCATED CODE

We illustrate the code obfuscation techniques by showing excerpts of code from our study in the Race program. Listing 1 shows the definition of the method `changeSpeed` from the file `MovingCarModel.java`, which changes the speed by a certain value (given as a parameter) depending on whether the car still has fuel (a value stored in the variable `gas`).

Listing 1: Clear code from Race `MovingCarModel.java`

```
1  public void changeSpeed(int i)
2  {
3      if(started)
4          if(gas == 0)
5          {
6              speed += i;
7              if(speed > maxSpeed / 10)
8                  speed = maxSpeed / 10;
9              else
10             if(speed < minSpeed / 10)
11                 speed = minSpeed / 10;
12         } else
13         {
14             speed += i;
15             if(speed > maxSpeed)
16                 speed = maxSpeed;
17             else
18             if(speed < minSpeed)
19                 speed = minSpeed;
20         }
21 }
```

Listing 2 shows the same code after applying *name overloading* where all identifiers have been renamed to some arbitrary values that have nothing to do with program semantics. The comparison between Listings 1 and 2 shows that apart from changing names of identifiers, the code stays structurally the same (e.g. lines 6 and 27 correspond).

Listing 3 shows the code from Listing 1 after introducing *opaque predicates*. In the given version of Sandmark, opaque predicates are generated using queries to a tree data structure which is manipulated in randomly looking ways. Predicates are then used to insert dead code that uses valid identifiers in random ways (see for example lines 45–47). To determine the truth value of a predicate (and therefore to eliminate dead code), an analyst has to first understand the way in which the tree was changed. The example shows that while all original code is maintained (e.g. line 6 corresponds to line 57), opaque predicates can be used to considerably complicate a program's control flow.

Listing 2: Code from Listing 1 obfuscated with NO

```
22 public void __m1(int i)
23 {
24     if(__f22)
25         if(__f19 == 0)
26         {
27             __f5 += i;
28             if(__f5 > __f6 / 10)
29                 __f5 = __f6 / 10;
30             else
31             if(__f5 < __f7 / 10)
32                 __f5 = __f7 / 10;
33         } else
34         {
35             __f5 += i;
36             if(__f5 > __f6)
37                 __f5 = __f6;
38             else
39             if(__f5 < __f7)
40                 __f5 = __f7;
41         }
42 }
```

Listing 3: Code from Listing 1 obfuscated with OP

```
43 public void changeSpeed(int i) {
44   if (Node.getI() != Node.getH()) {
45     lastFuel = (0L + time2) - (long) lap;
46     started = lastFuel == 0L;
47     Node.getF().setLeft(Node.getH().getLeft());
48   } else {
49     Node.getG().getLeft().swap(
50         Node.getG().getRight());
51     if (started)
52       if (Node.getI() == Node.getH()) {
53         if (gas == 0) {
54           if (Node.getF() == Node.getG()) {
55             Node.getF().setLeft(
56                 Node.getI().getRight());
57             speed += i;
58           } else {
59             [...]
60           }
61           if (Node.getI() != Node.getH()) {
62             lap = 1 + maxSpeed / status;
63             time += maxSpeed;
64             Node.getH().setLeft(
65                 Node.getH().getLeft());
66           } else {
67             Node.getF().getRight().swap(
68                 Node.getF().getRight());
69             if (speed > maxSpeed / 10) {
70               if (Node.getF() != Node.getG()) {
71                 Node.getF().getLeft().swap(
72                     Node.getH().getLeft());
73                 track = track;
74               } else {
75                 speed = maxSpeed / 10;
76                 Node.getF().setLeft(
77                     Node.getF().getLeft());
78               }
79   [...]
80 }
```

B. THE ONLINE SURVEY

In this section we present the full survey used in the study. This survey was slightly adapted from the materials of Ceccato et al. [12] to reflect the fact that we measured times of task completion and programming behavior, whereas Ceccato et al. asked their participants to note down their starting and finishing times, and to estimate the percentage of time they spent on reading and running the code, and executing the code in debugging mode.

B.1 Pre-Test Questions

At first the participants filled out a pre-test questionnaire in the online survey.

Q1. What is your position?

- ( ) Bachelor student
- ( ) Master student
- ( ) Diploma student
- ( ) PhD student
- ( ) Post Doc
- ( ) Professor
- ( ) Other:

Q2. What is your study subject? (this question was only displayed for participants who are either Bachelor, Master or Diploma student)

Q3. At which department are you working? (this question was only displayed for participants who are either PhD student, Post Doc or Professor)

Q4. How old are you?

Q5. How would you describe the quality and the type of the code you wrote so far?

- ( ) Few and small programs (e.g., course exercises)
- ( ) Many small programs
- ( ) Small programs and few (1 or 2) big programs (e.g., thesis and projects)
- ( ) Big programs

Q6. Have you ever worked as computer programmer?

- ( ) No
- ( ) Yes, part-time
- ( ) Yes, full-time

Q7. Do you have high or very high experience in any programming language(s)? If yes, please name them:

Q8. Did you participate in a Software Reverse Engineering or Hacking Lab Course?

- ( ) Yes
- ( ) No
- ( ) Other:

What is your experience/knowledge in... (5-point Likert scale from "very low" to "very high")

Q9. code obfuscation?

Q10. Java?

Q11. the usage of Eclipse for software development?

Q12. debugging software?

Q13. the usage of Eclipse for debugging software?

Before being able to work on the programs a password was displayed in the survey to gain access to the directory of the source files of the first program. This was done in order to prevent participants from analyzing the source code before the tasks to solve were presented.

B.2 Program Descriptions

After the participants got access to the the source files for a program, a short description about the program's general usage was displayed in the survey.

Race program

CarRace is a network game that allows two players run a car race.

The player that first completes the total number of laps wins the race. Use the arrow keys to control the car (your car is the green one). Keep "up" and "down" keys pressed to accelerate and brake. Press "right" and "left" arrows to turn right and left.

The car constantly consumes fuel, when the car runs out of fuel the speed drops. In order to avoid this case the players should stop at the box to refuel. The number of completed laps and the fuel level is displayed on the upper part of the window.

Chat program

ChatClient is a network application that allows people to have text based conversation through the network. Conversations can be public or private, depending on how they are initiated.

The application shows on the right a list of available rooms. When the application starts, the "default" room is accessed. It is a public room where all the users are participating. In order to access another room (e.g., Room 1) the name of the room must be clicked from the "Available Rooms" list, a new tab will be visualized. All the messages sent to a conversation within a room are received to all the users registered to that room.

A private conversation (only two users) can be initiated by clicking the name of a user from the "Online Users" list.

B.3 Questions after each task

After the completing the Pre-Test Questions the participants worked on the tasks as specified in the main part of the paper in Table 2. The tasks were presented in random order. After each task the following question had to be answered by all participants.

Q14. Did you have enough time to solve this task?

- ( ) Yes
- ( ) No

Only if the participant had enough time, the next questions were displayed and had to be answered using a 5-point Likert scale from "strongly agree" to "strongly disagree".

Q15. I had enough time to perform the tasks

Q16. The description of the task was perfectly clear to me

Q17. I experienced no difficulty in the identification of the segment of code relevant for the task

Q18. The debugging environment is useful to execute the task

Q19. I found the Eclipse Refactor facility useful for this task

Q20. For this task I spent a lot of time reading the code

Q21. For this task I spent a lot of time running the code

B.4 Questions after completing both tasks for a program

After both tasks and the corresponding questions about them were answered, questions regarding the programs itself were posed. Again, a 5-point Likert scale from "strongly agree" to "strongly disagree" had to be used.

Q22. The description of the application was clear

Q23. I experienced no difficulty in understanding the program

Q24. Running the code was useful to understand the code

The completion of these answers for the first program led to the password for getting access to the second program being displayed.

B.5 Post-Test Questions

After the questions about the second program were answered, some post-test questions had to be answered using a 5-point Likert scale from "strongly agree" to "strongly disagree".

Q25. I experienced no difficulty in using the development environment (Eclipse)

Q26. I experienced no difficulty in using the Eclipse debugger

The experiments were conducted over the course of two semesters. For the second semester we added two questions at the end of the survey in which the participants could indicate if they worked on similar code before. These questions were added after an additional analysis of the literature on code comprehension, where the so-called *domain experience* emerged as an additional performance factor [8].

Q27. Have you ever programmed any kind of program which was in your personal opinion similar to the chat program? If yes, please specify

Q28. Have you ever programmed any kind of program which was in your personal opinion similar to the race game?

"We make it a big deal in the company": Security Mindsets in Organizations that Develop Cryptographic Products

Julie M. Haney[1], Mary F. Theofanos[1], Yasemin Acar[2], Sandra Spickard Prettyman[3]

[1]National Institute of Standards and Technology
{julie.haney, mary.theofanos}@nist.gov

[2]Leibniz University Hannover
acar@sec.uni-hannover.de

[3]Culture Catalyst
sspretty50@icloud.com

ABSTRACT

Cryptography is an essential component of modern computing. Unfortunately, implementing cryptography correctly is a non-trivial undertaking. Past studies have supported this observation by revealing a multitude of errors and developer pitfalls in the cryptographic implementations of software products. However, the emphasis of these studies was on individual developers; there is an obvious gap in more thoroughly understanding cryptographic development practices of organizations. To address this gap, we conducted 21 in-depth interviews of highly experienced individuals representing organizations that include cryptography in their products. Our findings suggest a security mindset not seen in other research results, demonstrated by strong organizational security culture and the deep expertise of those performing cryptographic development. This mindset, in turn, guides the careful selection of cryptographic resources and informs formal, rigorous development and testing practices. The enhanced understanding of organizational practices encourages additional research initiatives to explore variations in those implementing cryptography, which can aid in transferring lessons learned from more security-mature organizations to the broader development community through educational opportunities, tools, and other mechanisms. The findings also support past studies that suggest that the usability of cryptographic resources may be deficient, and provide additional suggestions for making these resources more accessible and usable to developers of varying skill levels.

1. INTRODUCTION

In a dynamic, threat-laden, and interconnected digital environment, cryptography protects privacy, provides for anonymity, ensures the confidentiality and integrity of communications, and safeguards sensitive information. Given the need for cryptography, there is an abundance of cryptographic algorithm and library choices for developers wishing to integrate cryptography into their products and services. However, developers often lack the expertise to navigate these choices, resulting in the introduction of security vulnerabilities [27]. A 2016 industry survey that included over 300,000 code assessments found that 39% of those applications had cryptographic problems [72]. Implementing cryptography correctly is a non-trivial undertaking.

In 1997, security expert Bruce Schneier commented on the lack of cryptographic implementation rigor and expertise at that time, asserting, "You can't make systems secure by tacking on cryptography as an afterthought. You have to know what you are doing every step of the way, from conception to installation" [61]. Past studies have supported this observation by revealing a multitude of errors in the cryptographic implementations of software products (e.g., [17–19, 42]) and the pitfalls developers encounter when including cryptography within products (e.g,. [1, 2, 48]). This body of research suggests that developers have not progressed much in the past 20 years. However, as these studies have been largely focused on individual practices outside the professional work context or on the development of mobile apps, it is unclear if these shortcomings also apply to organizational development and testing, particularly among organizations for which security and cryptography are essential components. One exploratory survey examined high-level organizational practices in cryptographic development, but lacked rich insight into actual practices and motivators behind those [31]. Clearly, there is a gap in the literature in more thoroughly understanding organizational cryptographic development practices.

To address this gap, we performed a qualitative investigation into the processes and resources that organizations employ to ensure their cryptographic products are not fraught with errors and vulnerabilities. We define the scope of cryptographic products as those implementing cryptographic algorithms or using crypto (cryptography) to perform some function. We conducted 21 in-depth interviews involving participants representing organizations that develop either a security product that uses cryptography or a non-security product that heavily relies on cryptography. Unlike previous studies, our participants were professionals who were highly experienced in cryptographic development and testing, not computer science students or developers with little cryptographic experience.

The study aimed to answer the following research questions:

Q1 What are the cryptographic development and testing practices of organizations?

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

Q2 What challenges, if any, do organizations encounter while developing and testing these products?

Q3 What cryptographic resources do these organizations use, and what are their reasons for choosing these?

Our findings went deeper than uncovering practices, revealing a security mindset not noted in other research results. We discovered that some organizations believe they have achieved the expertise and rigor recommended by Schneier. Compared to developer populations studied in the past, the organizations in our investigation appear to have a stronger security culture and are more mature in their cryptography and security experience. The strong security culture we observed does not appear to be linked to company size or available resources. These security mindsets permeate the entire development process as they inform judicious selection of cryptographic resources and rigorous development practices.

Our work has several contributions. To our knowledge, this is the first in-depth study to explore cryptographic development practices and security mindsets in organizations from the viewpoint of those with extensive experience in the field. While some of the practices identified in our study may be considered known best practice within the security community, our paper is novel in that there are few research studies documenting occurrences of strong security culture and development in actual practice, and none within the cryptography context. Our study provides systematic, scientific validation to the anecdotal point often made by security experts that there is no magical, one-dimensional solution to cryptographic development. Rather, good crypto is the result of a concerted effort to build expertise and implement secure development practices. The enhanced understanding encourages additional research initiatives to explore variations in those implementing cryptography. This can aid in transferring lessons learned from more security-mature organizations to the broader development community through educational opportunities, tools, and other mechanisms. Our findings also support past studies that suggest that the usability of cryptographic resources may be deficient, and provide additional suggestions for making these resources more accessible and usable to developers of varying skill levels.

2. RELATED WORK

To provide context, this section begins with a brief overview of cryptographic standards and certifications frequently referenced in our interviews. We then underpin our assertion that cryptographic development is not a trivial undertaking by summarizing past research on crypto misuse and lack of crypto resource usability. We also present an overview of prior work on lack of security mindsets and secure development practices to serve as a contrast to the more security-conscious approaches of our study organizations.

2.1 Cryptographic Standards

Cryptographic algorithm standards are developed by consensus of community stakeholders (e.g. vendors, researchers, governments) to foster compatibility, interoperability, and minimum levels of security. These can be found in formal standards documents from organizations such as Institute of Electrical and Electronics Engineers (IEEE) [35], International Organization for Standardization (ISO) [36], and the U.S. National Institute of Standards and Technology (NIST) [52]. Likely due to the U.S. locations of most of the study organizations, the participants most often mentioned cryptographic requirements issued by NIST. As perhaps the best known government standard, the Federal Information Processing Standards Publication (FIPS) 140-2 "specifies the security requirements that will be satisfied by a cryptographic module utilized within a security system protecting sensitive but unclassified information" [49]. These requirements are mandatory for cryptographic products purchased by the U.S. Government, but also are used voluntarily outside the government. There are two certification programs associated with FIPS 140-2 [50, 51]. Under these programs, vendors may submit cryptographic algorithm and module implementations for validation testing to accredited testing laboratories.

2.2 Cryptographic Misuse

Numerous studies have highlighted the difficulty developers have in correctly implementing cryptography. In 2002, Gutmann observed that security bugs were often introduced by software developers who did not understand the implications of their choices [30]. Nguyen showed that even open-source implementations under public scrutiny have cryptographic flaws [53]. Lazar performed a systematic study of 269 cryptographic vulnerabilities in the Common Vulnerabilities and Exposures (CVE) database, noting that 17% of bugs were in cryptographic libraries and the remaining 83% were in individual applications, usually due to cryptographic library misuses [40]. Georgiev et al. discovered rampant misuse of Secure Sockets Layer (SSL) in security-critical applications due to poorly designed application programming interfaces (APIs) [25]. Fahl et al. analyzed 13,500 Android apps and found that 8% were susceptible to man-in-the-middle attacks [19]. Using static analysis, Egele et al. found similar issues, observing that 88% of over 11,000 examined Android apps contained a significant error in their use of a cryptographic API [18]. Li et al. analyzed 98 apps from the Apple App Store and found 64 (65.3%) that contained cryptographic misuse flaws [42].

2.3 Usability of Cryptographic Resources

Usability is often neglected in cryptographic resources such as standards and libraries, resulting in complex solutions that provide little assistance to developers in making secure choices [27, 48]. Several research groups attempted to remedy this by developing tools, e.g., OpenCCE [4], CryptoAssistant [23] and Crypto Misuse Analyzer [64], to guide developers in choosing and integrating appropriate cryptographic methods. Others proposed more usable cryptographic libraries. Forler et al. developed libadacrypt, a cryptographic library created to be "misuse-resistant" [20]. Bernstein et al. created the Networking and Cryptography library (NaCl) [8], a cross-platform cryptographic library designed to avoid errors found in widely used cryptographic libraries like OpenSSL [55]. Acar et al. conducted a usability study of cryptographic APIs that revealed that, in addition to usable interfaces, clear documentation with code samples and support for common cryptographic tasks were important in aiding developers [1].

2.4 Security Development and Mindset

There is much to learn from an examination of secure de-

velopment and testing practices since even the best implemented cryptography can be subverted by the flawed implementation of another system component. McGraw advocated for security to be integrated into all aspects of the software development lifecycle [43]. Several documents, for example the Microsoft Security Development Lifecyle [34] and the System Security Engineering Capability Maturity Model (SSE-CMM) [44], formally define secure development practices. More recently, the Open Web Application Security Project (OWASP) Secure Software Development Lifecycle project is working towards providing guidelines for web and application developers [54]. However, none of these resources specifically mentions considerations for cryptography.

Not surprisingly, the implementation of formal secure development processes is not an easy task. A 2016 Veracode survey of over 350 developers indicated that organizations are prevented from fully implementing a secure development process due to a variety of challenges, including security testing causing product timeline delays, complexity in supporting legacy security processes, security standards and policies varying across the organization, and developers not consistently following secure coding practices [73]. Kanniah and Mahrin found that a variety of organizational, technical, and human factors affected implementation of secure software development practices [38]. These factors included developer skill and expertise, communication among stakeholders, and collaboration between security experts and developers.

Failures in development and testing leading to security errors appear to reflect a deficiency in a security mindset. Schneier claimed that "Security requires a particular mindset. Security professionals – at least the good ones – see the world differently" [62]. A security mindset involves being able to think like an attacker, maintaining a commitment to secure practices, and perpetuating a strong security culture.

A need for a security mindset is revealed in several studies that explored reasons why developers make security errors. For example, Xie, Lipford, and Chu identified an absence of personal responsibility for security as well as a gap between developers' understanding of security and how to implement it [76]. Xiao et al. discovered that the failure to adopt secure development tools was heavily dependent on social environments and how tool information was communicated [75]. From a testing perspective, Potter and McGraw argued that security testing is commonly misunderstood and should be more risk-based, involving an understanding of a potential attacker's mentality [58]. Bonver and Cohen agreed, noting that security testers should work closely with architects and developers to identify potential vulnerabilities, taking into account how an attacker may exploit a system [9].

3. METHODOLOGY

Between January and June 2017, we conducted 21 interviews of individuals working in organizations that develop products that use cryptography. Following rigorous, commonly accepted qualitative research methods, we continued interviewing until we reached theoretical saturation, the point at which no new themes or ideas emerged from the data [45], exceeding the minimum of 12 interviews prescribed in qualitative research best practices [29].

Our research team was multidisciplinary and consisted of a computer scientist specializing in information security and human-computer interaction, a computer scientist specializing in usability, a mathematician with research experience in usable security and privacy, and a sociologist experienced in qualitative research. Having a diverse team may improve research quality "in terms of enabling sounder methodological design, increasing rigor, and encouraging richer conceptual analysis and interpretation" [6].

The study was approved by our institutional review board. Prior to the interviews, participants were informed of the purpose of the study and how their data would be used and protected. Interview data were collected and recorded without personal identifiers and not linked back to the participants or organizations. Interviews were assigned an identifier (e.g., C08) used for all associated data in the study.

3.1 Recruitment

To ensure that we could explore different perspectives within the cryptographic product space, our sampling frame consisted of individuals who had organizational experience designing, developing, or testing products that use cryptography or who were knowledgeable about and had played a key role in these activities (e.g., managers of teams that performed these tasks). We utilized a combination of purposeful and convenience sampling strategies, which are widely employed in exploratory qualitative research [56]. Purposeful sampling was used to select organizations of different sizes and participants who had knowledge and experience within this specialized topic area. This was combined with convenience sampling, where participants were sought based on ease of accessibility to the researchers and their willingness to participate in the study.

Nine individuals were recruited from prior researcher contacts. Additional participants were recruited from among vendors at the RSA conference [14], a large industry IT security conference that also hosts an exhibition floor with security-focused vendors. A list of 54 potential organizations was compiled after in-person researcher contact on the exhibition floor. After the conference, we identified organizations that provided organizational diversity in our sample and that were accessible to the researchers. We emailed 17 of them to invite participation in the study. Eleven organizations agreed to participate. One additional organization was recruited based on the recommendation of a participant.

3.2 Interviews

We collected data via semi-structured interviews. Interviews were conducted by two of the researchers and ranged from 30 to 64 minutes, lasting on average 44 minutes. We conducted 21 interviews with 1-3 participants per interview, 29 participants total. Five organizations opted to have more than one participant in the interview: three organizations had three participants and two organizations had two participants. Face-to-face interviews were conducted if feasible. Otherwise, participants were given the choice of a phone or video conference interview. Five interviews were conducted face-to-face, 10 by phone, and six via video. Interviews were audio recorded and transcribed by a third-party transcription service.

After the first nine interviews, we performed a preliminary analysis and chose to make minor revisions to the interview

protocol in accordance with the qualitative research practice of theoretical sampling. Theoretical sampling involves adjusting data collection while the study is in-progress to better explore themes as they arise [13].

The interviews began with demographic questions about the organization (e.g., size, products) and the individual participants (e.g., role within the organization, professional background). Subsequently, participants were asked to describe their organizations' development and testing practices and associated challenges for their cryptographic products. Questions then transitioned into exploring cryptographic resources used by the organizations and how the participants thought those resources might be improved, if at all. The complete interview protocol is included in Appendix A.

3.3 Analysis

We utilized both deductive and inductive coding practices. Initially we constructed an *a priori* code list based on our research questions and literature in the field to provide direction in the analysis. As we performed multiple rounds of coding, we also identified emergent codes in the data. This iterative, recursive process helped us identify additional codes and categories as we worked with the data until we reached saturation [26, 69].

Five interviews (almost 24%) were first coded individually, then discussed as a group to develop a codebook. Although there is debate on the amount of text to collectively sample in qualitative research, the amount of text we group coded far exceeds the minimum of 10% often cited as standard practice [33]. We calculated intercoder reliability on this subset of the data using the ReCal3 software as a tool to help us refine our codes [21, 22]. For the five interviews, we reached an average Krippendorf's Alpha score of .70, with a high of .78, which is considered within the fair to good bounds for exploratory research having rich data with many codes and a larger number of coders [11, 15, 39].

Beyond the agreement metric, and in line with the views of many qualitative research methodologists, we thought it was important to focus on how and why disagreements in coding arose and the insights gained from discussions about these [5, 63]. These discussions better allow researchers to refine coding frames and pursue alternative interpretations of the data. During analysis, we found that each coder brought a unique perspective that contributed to a more complete picture of the data. For example, two of our coders more often identified high-level, nuanced codes about emotions and personal values, which may be due to their many years of working in human-focused contexts. These interpretations were often missed by the other coders who had more of a technology-focused background.

After coding of the initial subset of data, the remaining 16 interviews were coded by two coders each. Once each pair completed their coding, they had a discussion about the data to address areas of divergence about their use and application of the codes. This discussion resulted in the coders being able to understand each other's perspectives and come to a final coding determination. New codes that were identified during these discussions were added to the codebook, with previously coded interviews then re-examined to account for additions. The final codebook is included in Appendix B.

During the coding phase, we also engaged in writing analytic memos to capture thoughts about emerging themes [13]. For example, one memo captured thoughts on cryptography complexity. Once coding was complete, we reorganized and reassembled the data, created coding arrays, discussed patterns and categories, drew models, discussed relationships in codes and data, and began to move from codes to themes [59]. The team met regularly to discuss our emergent ideas and refine our interpretations. This process allowed for the abstraction of ideas and the development of overarching themes, such as how an organization's maturity and security culture drive formal development practices.

3.4 Limitations

Our study has several limitations. First, interviews are subject to self-report bias in which participants tend to underreport behaviors they think may be viewed as less desirable by the researchers, and over-report behaviors deemed to be desirable [16]. Given that the researchers who conducted the interviews represented an institution known for its security expertise, this bias may have influenced participant responses. We also note that the answers to some of the interview questions reflected participants' *perceptions* of the security level of their products and the security mindset of their organizations, which may or may not reflect reality.

Since there is no prior research into what is representative of the cryptographic development community, our sample is not characteristic of all types of organizations in this space. Although we did strive for diversity in organization size, with a smaller sample size common in qualitative research, we cannot definitively identify differences due to this variable.

4. DEMOGRAPHICS

Table 1 provides an overview of the organizations and participants in our study. To protect confidentiality, product types and participant roles are generalized.

The organizations represented in our study were of different sizes, with six being very large (10,000 or more employees), six large (10,00 - 99,99 employees), three medium (100 - 999 employees), three small (10 - 99 employees), and three very small/micro (1 - 9 employees) [24, 32]. All organizations developed a security product that uses cryptography (e.g. end user security software, hardware security module) or a non-security product that heavily relies upon cryptography to protect it (e.g. Internet of Things devices, storage devices, operating systems). Customers of these products ranged widely and included consumers, other parts of the organization, and organizations and businesses in multiple sectors such as government, technology, health, finance, automotive, and retail. Of the 15 organizations that discussed how long their companies had been implementing cryptography in their products, 12 had 10 or more years experience, with six of those having at least 20 years experience. The remaining three were startup companies that had been doing cryptographic development since their inception.

The 29 participants were a highly experienced group with several having made major contributions to the cryptography field. All participants had technical careers spanning 10 or more years, with several having been in the field for 30+ years. At least one individual from each of the interviews either currently worked on cryptography and security as a major component of their jobs (19 participants), or had

Table 1: Interview Demographics

ID	Org Size	Reg	Prod Type	Participant(s)
C01	VL	U.S.	HW	Lead crypto architect
C02	VL	U.S.	COM	Lead cryptographer
C03	VL	U.S.	HW, SW	Systems architect
C04	VL	U.S.	HW	Crypto design reviewer
C05	VL	U.S.	HW	Crypto architect
C06	VS	U.S.	SW	Systems analyst
C07	VS	U.S.	COM	Founder & researcher
C08	VS	U.S.	IOT	Founder & developer
C09	VL	U.S.	IOT	Researcher
C10	L	U.S.	SW	Founder & engineering lead
C11	L	U.S.	HW, SW	Product manager
C12	S	U.S.	SW	1) CTO 2) Marketing engineer 3) Business manager
C13	S	U.S.	SW	1) Chief Evangelist 2) Strategy Officer
C14	M	U.S.	SW	1) Marketing lead 2) Developer 3) Quality assurance
C15	L	U.S.	SW	Principal engineer
C16	L	U.S.	SW	CISO
C17	M	Eur	SW	1) CTO 2) Security engineer
C18	L	Aus	SW	Crypto engineer
C19	L	U.S.	SW	CTO
C20	S	U.S.	COM	1) Founder & architect 2) Compliance lead 3) Marketing director
C21	M	Eur	SW	Crypto specialist

Org Size: VL=Very Large, M=Medium, S=Small, VS=Very Small/Micro. **Reg** (Region/location of participant): U.S.=United States, Eur=Europe, Aus=Australia. **Prod (Product) Type:** HW=Hardware, SW=Software, COM=Communications Security, IOT=Internet of Things.

worked on cryptography extensively in the past (3 participants). The other participants were marketing or product leads, but all had a technical background.

Most of the participants had learned cryptography "on-the-job" as opposed to having formal training in the field. Five had an education in mathematics, but only two of those had studied cryptography as part of their formal study. Three had an engineering education, one had a physics degree, and the rest were educated in a computer-related discipline.

Four out of the 29 total interview participants had enhanced their knowledge through involvement in cryptographic standards groups. A cryptography architect commented on the value of his involvement in IEEE cryptographic standards early in his career: *"That's where I got to commune with cryptographers for a couple of years, me on the engineering side, and them on the crypto side... You end up learning things as a result of that process" (C05).*

5. RESULTS

The interviews revealed an organizational security mindset seldom seen in other cryptographic development studies. Our results suggest that the security mindsets had their roots in organizational attributes and culture that lay the foundation for the selection and use of cryptographic resources and rigorous development and testing practices.

For this section, we report counts of interviews throughout, joining group interview participants' answers to account for their organization. Due to the semi-structured nature of the interviews, the counts do not indicate quantitative results, but are reported to give weight to certain themes that were mentioned across interviews.

5.1 Security Mindset Characteristics

The interviews revealed organizational and personal characteristics that demonstrated a strong security mindset. These characteristics included professional maturity gained through experience, a deep understanding of the complexity of cryptography, and evidence of a strong security culture.

5.1.1 Emphasis on Experience and Maturity

Bruce Schneier said, "Only experience, and the intuition born of experience, can help the cryptographer design secure systems and find flaws in existing systems" [61]. The study participants expressed the importance of this experience as they repeatedly highlighted their own and their organizations' substantial maturity with respect to developing secure products and working with cryptography.

Overall, the organizations placed great value on hiring and retaining experienced technical staff. As previously mentioned, the majority of participants had substantial individual experience with cryptographic products. They also tended to work with other seasoned individuals. One participant described his team: *"We have a couple of the same core people on our test team who've been here for 25 years. They've gotten very good" (C01).* A startup company had only a few employees, but they all were veteran security software developers: *"Everyone we have has a lot of experience. I think the most junior person has a master's degree... and 10 and a half, 11 years of experience" (C07).*

Eight interviews noted the importance of experience when doing secure development, especially with cryptography. One participant remarked that secure products are ultimately dependent on *"the people that are designing it having the necessary knowledge and experience and understanding the whole picture, not just the little microscopic piece they're working on" (C01).* An interviewee with a long cryptographic background emphasized that there is no substitute for experience as he recounted a story of how his former company had to hire three less-qualified, full-time people to replace him in the work he had been doing on a part-time basis. A company founder remarked about the high level of technical maturity needed to properly deal with the complexity of cryptography: *"The level of education somebody needs to attain to be effective at doing crypto is relatively high. So it's not like I can put somebody who's fresh out of school on something and expect good results" (C10).*

5.1.2 Recognition of Cryptography Complexity

Based on their own experiences, participants and their organizations were keenly aware that, even though developing secure cryptographic implementations may appear to be easy, it is deceptively difficult. Despite proven algorithms being available, *"the algorithms are fairly involved,*

and they're difficult to understand" (C20). Yet understanding the algorithm is just the first step. As described by an IoT researcher, translating the algorithm correctly into a product is *"not a trivial implementation" (C09).* One design reviewer remarked about the pervasiveness of cryptographic design errors he encountered over the course of his career: *"I think I reviewed about 2,500 products... I can only remember eight that did not have a problem that either... they had to fix, or... they had to change their marketing claims" (C04).*

Our interviews also suggest that building cryptographic systems appears to be more of an art than a science, requiring a careful balance between security, performance, and usability. One participant described these tensions:

> *"Crypto algorithms are already very highly optimized ... It's like balancing a supertanker on a 40,000-foot high razor blade, and if you make one small change you destroy the performance. If you make it the other way, you just destroy the security." (C04)*

Because of the complexity, our participants recognized that design and implementation errors can be rampant and require rigorous review and testing to answer the misleadingly simple question *"How do you know it's right?" (C08).* However, assessing cryptographic products can be challenging, requiring knowledge and experience to construct good tests. A cryptographic development architect described challenges his organization had in the past when they lacked maturity in cryptographic implementation and testing:

> *"We actually implemented a new symmetric encryption algorithm, and it passed all the tests... and it turned out that they did the algorithm completely wrong. That was because... they wrote a test which said, 'Generate some random data, encrypt it with the algorithm, decrypt it, and see if you get back the original data.' Well, yeah, it got back the original data, but the encrypted data was incorrect. It just was symmetric, so it did the same wrong thing encrypting and decrypting." (C01)*

The organizations also understood that cryptography is just one of many interdependent product components, with all of them having to be properly implemented to ensure security. This sentiment was echoed by one participant who commented, *"For us, the design of the overall architecture that uses the crypto algorithms is almost as important as the correctness of the underlying algorithms themselves" (C01).*

5.1.3 Security Culture

Each organization in our study appeared to have a strong security culture that was interdependent on the maturity and experience of its employees. A security culture is a subculture of an organization in which security becomes a natural aspect in the daily activities of every employee [60]. For development organizations, this involves having dedicated security people, spending money on security, making security a company core value, and offering secure products. For the studied organizations, the culture included a commitment to address security and the perpetuation of a security mindset to others in the organization.

Commitment to Security: The organizations we studied thought that having good product security and strong cryptographic implementations was a *"core value" (C07)*, *"the key to quality" (C09)*, and essential to company identity. As an example, a Chief Information Security Officer (CISO) of a large company remarked, *"In our company, we are developing and selling security to our customers. So we care about, basically, all three sides of the sort of security triangle [confidentiality, integrity, availability] in what we do." (C15).* A security engineer talked about his company's belief that security must be an important consideration even when faced with competing tensions such as time-to-market: *"Since we do a [security product], everybody feels that we need to add security and good crypto at every step, so it's not a big issue to find the right balance" (C17).* Another participant commented on how his company demonstrates its commitment to secure cryptography: *"We have some fairly large teams which concern themselves with cryptography and secure design methodology. All engineers get training on secure design and we make it a big deal in the company" (C05).*

Security culture is not just internally-motivated. External motivators, like gaining a larger market share or customer requirements, often necessitate strong attention to security. One participant commented on how customer expectations fostered a security culture that drove rigorous testing processes within his company:

> *"We serve the kinds of customers who rely on the stuff to work reliably and properly from the get-go, when they buy it. So it's not like... 'Maybe we'll update something later, if we find some problems.' That's not our philosophy, and that's not what our kind of customers expect... Part of that is also company culture." (C01)*

Although security culture is often thought of as a "top-down" phenomenon, it must be accepted by and acted upon at all levels. One participant, a CISO for a large company, commented on the importance of the security culture being pervasive throughout an organization:

> *"If there's senior executive support for a strong security program,... that helps tremendously. At the same time, if there is still a very strong feeling amongst a large number of developers that security, cryptography, and everything that's related to that is really a nuisance that should get out of the way and just to prevent them from writing more interesting features, it's definitely a concern." (C16)*

The interviews showed evidence that our participants are critical to the "bottom-up" support of organizational security culture. They serve as security champions and self-appointed educators, leading by example and projecting their values, personal philosophies, and commitment to security on the rest of the organization. Two of the participants explicitly embraced this role as a personal mission when they referred to themselves as "security evangelists." Another expressed his feeling of personal accountability to enact security in the products he supported: *"It's essentially a mark of my success or not, that I'm measured against, of whether those things remain secure or hacked" (C05).*

Perpetuating a Security Mindset: Just as the employees influence security culture, so does the culture influence employees by perpetuating a security mindset. Part of this perpetuation involves expert employees mentoring and supporting less-experienced personnel in their learning of secure programming methods and specialized security topics such

as cryptography, as discussed in ten interviews.

The interviews suggest that providing an opportunity for individuals to gain hands-on experience with real products is important in understanding the issues involved in developing a cryptographic product. However, given the distinct possibility that a novice will make errors in the implementation, precautions must be taken. Two participants suggested that mock training exercises conducted on a separate testing infrastructure may be valuable initial steps. Others discussed mentoring and peer review activities within their organizations. For example, one organization enacted *"parent programming" (C14)* for any code that uses cryptography. Another had a formal peer review process:

> *"We review everyone's code... When I write something down and it has a flaw in it, I'm told about it, which is good... I think what we do is we take smart people who care about doing good work, and we foster an environment where they're not afraid to receive constructive criticism, and they're not intimidated away from giving it." (C15)*

The influence of an organization's security mindset does not necessarily end when an individual leaves that organization. As evidenced by three participants who left large companies to start their own small businesses, there seemed to be a transfer of security culture and practices from previous employers. A small company founder described this transfer: *"I think some of it could be just kind of from my career background, what I learned were the best practices... I think we've just kind of learned them in the beginning and kind of kept that culture" (C08).*

Size Doesn't Matter: We found no significant differences in perceived security culture or overall security practices based on size. Obviously, larger organizations had more resources to dedicate to security and cryptographic development. However, smaller organizations understood that vulnerabilities in their products could do great harm to the company's reputation, and so were committed to security and made thoughtful decisions about how they developed and tested, even if on a smaller scale. For example, the founder of a micro business noted:

> *"Being a small company, we're trying to also gain credibility. And we don't want to just claim that we have the fastest [crypto implementation] in the world, we also want to make sure that it is built safely and validated... [W]e cannot afford for this thing not to work properly." (C07)*

5.2 Selection of Resources

Because of security mindsets, participants revealed a proclivity towards careful selection of resources to help them attain their goal of secure cryptographic implementations. In this section, we describe considerations taken when choosing and evaluating cryptographic resources.

5.2.1 Standards

In line with the popular quote "The nice thing about standards is that you have so many to choose from" by Andrew S. Tanenbaum [67], the interviews revealed that all the organizations used some type of cryptographic standards from organizations such as IEEE, ISO, American National Standards Institute (ANSI) [3], Internet Engineering Task Force (IETF) [37], and Payment Card Industry (PCI) [57]. All but one described using NIST standards or guidance documents, most commonly FIPS 140-2. The participants and their organizations were knowledgeable enough to understand and evaluate the appropriateness of cryptographic standards. However, not all organizations have the need to directly read the standards. For those that do, this requires maturity in the field given the standards' complexity. A Chief Technology Officer (CTO) at a small company expressed this challenge: *"A lot of standards are notoriously difficult to read. Unless you're an expert in the field, a lot of them don't make sense" (C12).* In another interview, a principal engineer commented on the difficulty in translating standards to products:

> *"I can tell you from my personal experience understanding the fundamentals of these things, still the standards were a challenge to use because they were very divorced from the implementation day-to-day details that I encounter while I'm trying to plug all the pieces together." (C15)*

Despite the complexity, when selected carefully and implemented correctly, standards were seen as beneficial for several reasons noted by our participants. Participants in eight out of 21 interviews commented that the community review of standards results in a more correct and secure solution. A CISO at a large software as a service company reflected on the value of public scrutiny: *"By relying on other standards that [were] vetted by multiple parties, I have much higher assurance that the underlying cryptography and design [are] done in an appropriate way" (C16).* A director of product management concurred with this: *"So the standards, because it's out there and everybody's looking at it and testing it, we depend on that as kind of a layer of security" (C14).*

Included in community review is the transparency of the standards process. One participant illustrated this observation using the popular AES standard as an example:

> *"It gave us a lot of comfort knowing how AES evolved ... being able to see all the steps, having that all happen out in the open and why and how it happened. Very helpful to us making the decision for what we're going to use and why we're going to use it." (C10)*

Participants in nine out of 21 interviews commented that the use of standards eliminates some of the difficulty of cryptographic development and testing by providing an authoritative foundation. The founder of a software company said his organization relies heavily on standards because *"inventing it on your own is just a different level of complexity that we knew enough to know we did not want to be involved in" (C10).* Standards also add confidence that a product will be *"interoperable with our customers, with our partners, even with our competitors" (C16).*

Finally, participants noted that standards-based products may elicit customer confidence. The director of product line management at a large credential management company spoke of the importance of customer trust in gaining market share, saying, *"If the standard is mature, then it means our product's going to be more easily accepted by customers" (C11).*

Standards meet the needs of most companies we interviewed; however, there are cases in which standards fail to address a specific need. In these situations, organizations may extend or modify standards. These extensions were viewed as adding rigor and security in addition to functionality, making the cryptographic implementations *"really ahead of any industry standard practices" (C05).*

Interestingly, three participants commented on distrust of government standards because of allegations that a U.S. government agency purposely weakened cryptographic algorithms [28]. For example, although one organization made extensive use of standards, they took special measures to exclude aspects of a government standard they felt were questionable and exercised extra rigor in their testing processes to account for potential vulnerabilities. In an extreme case, a consulting company's observation of customer distrust of government standards and their own frustration with the complexity of those led them to develop a new cryptographic primitive: *"I think it [the distrust] comes from . . . news reports or exposés that say this standard may not be as secure as we think. . . There is a lot of doubt out there . . . That's why there has to be additional options and alternatives" (C06).*

5.2.2 Certifications

Seventeen out of 21 organizations obtained at least one formal certification that the implementation of cryptographic algorithms in their products met standards specifications. Three additional organizations developed and tested to certification criteria without undergoing full certification. Eighteen organizations referenced FIPS 140-2 certification [49], five Common Criteria [41], three the Payment Card Industry Data Security Standard (PCI-DSS) [57] validation program, one the Underwriters Laboratories (UL) certification [70], and others pursued country-specific certifications.

The perception of the benefit provided by adhering to certification requirements was mixed among our participants. Among those who obtain certifications, only five organizations expressed that certifications establish additional confidence through independent testing. One of these remarked, *"You have a lot of assurance that everything's going to be tested and get that nice, kind of warm and fuzzy" (C08).* Six organizations noted that, even though they do not undergo the formal FIPS 140-2 certification process, they build to and test against the certification specifications to gain added assurance. A participant from a key and identity management software company stated, *"as a small company, I think it is actually extra important to make sure that we go through this battery of tests just to in a way reassure the people we're talking to that this is a robust product" (C12).*

For some participants, certifications are perceived as being more useful for meeting customer expectations than for bolstering security. Organizations most often obtain certifications because these are requirements of their customers in certain sectors (e.g. government, financial): *"for some areas, if you don't get the check-mark you don't get to play" (C11).*

Unfortunately, as noted in 12 of 21 interviews, certifications can be expensive in time and resources, making them prohibitive for smaller companies, especially those with products that run on multiple platforms and have frequent version updates. However, our interviews suggest that confidence in the cryptographic implementations may not be dependent on any certification, but rather on the rigorous development and testing practices these organizations undertake. The founder of a small company commented that FIPS 140-2 certification was too costly for them to pursue, but was confident that his product met the certification requirements: *"It's a place where we've done enough testing ourselves. I know it's fine. I know we would pass" (C10).* Another participant, whose company did undergo certifications, felt that the certifications provided little assurance beyond what was provided by their own internal processes:

> *"We always design and test ourselves to have confidence that [the product] meets all those requirements before we release it to the lab for their testing. So when they come back and say, 'It passed this,' we say, 'Well, okay, we expected that. Thank you.' So the surprise is if something fails, that we expected to pass. That doesn't happen very often." (C01)*

Four of our participants also remarked that certifications may not be a robust contributor to the security of a product. They expressed strong sentiments that certifications, especially FIPS 140-2, are more of a "checkbox" for customers *"without any additional benefit of security" (C02).* A CTO and long-time cryptographer added, *"FIPS 140 is. . . not focused on how to use crypto securely. It's focused on how to safely provide crypto functionality" (C19).*

In addition, seven out of 21 organizations commented that maintaining FIPS certification may, in some cases, weaken security by discouraging updates throughout the lifecycle of the product. Once a product undergoes a significant update (for example, fixing a security vulnerability), it may lose its certification. Organizations are then put in a difficult position: *"this ability to address vulnerabilities and to patch validated code is a real problem. It sends the wrong message if you do what you should do, which is patch it and live without [the certification]" (C19).*

5.2.3 Third-Party Implementations

The complexity of implementing algorithms from scratch and the expertise required to write those compelled two thirds of the organizations to follow industry best practices by not writing their own cryptographic code and instead using third-party cryptographic implementations, for example open source libraries such as OpenSSL or built-in operating system APIs. One participant used an analogy to explain why his organization used these resources: *"Not every person should be performing brain surgery on another person. I also don't believe every software engineer should really go write crypto code" (C07).*

The organizations selected these third-party implementations based on several factors. First, four mentioned an implicit trust of the resource based on the reputation of the vendor or general community acceptance. In describing why his organization chose a set of cryptographic libraries, one participant said, *"You pretty much trust those libraries because they are widely used, and you can run test vectors against them easily" (C20).* A third of the organizations said that they have more confidence in formally vetted, certified implementations. A participant from a small company that works on IoT cryptography commented, *"If a vendor has submitted a library through FIPS 140-2 certification, versus a code that was up on GitHub for example,. . . I would*

be more inclined to trust the one that has gone through the FIPS validation" (C08).

Despite benefits of using third-party implementations, the organizations respected that the use of these external resources can still be difficult for less-knowledgeable individuals. A developer provided an example:

> *"[Crypto libraries] in general don't provide enough to be able to use them correctly out of the box... But there's many out there that think that they can just use AES, and I included it and I'm using it. But I'm not using it correctly, and then I'm leaving myself open to attack." (C14)*

This point illustrates that there is an important distinction between a cryptographic algorithm being certified or deemed "secure" and the proper use of the algorithm by developers. Similarly, third-party libraries have faced their share of security vulnerabilities due to implementation errors of otherwise sound cryptographic algorithms. Some of these vulnerabilities have had far-reaching impacts, for example the Heartbleed vulnerability in OpenSSL [71]. A security architect commented that third-party implementations are *"much more likely to contain implementation bugs and vulnerabilities" (C20).* Therefore, some organizations attempted to vet third-party implementations by doing their own vulnerability checks. One participant noted that his organization monitors the vulnerability databases for security issues with the libraries they use. Another commented on the extra security checks his company performs: *"We validate outside third-party libraries and software as they come in and confirm that they are bug-free and up to the latest standard or perform the right risk assessments" (C16).*

Finally, organizations may augment third-party implementations with their own internally developed modifications to avoid potential errors or address gaps in the resources. For example, to prevent developers from making errors while using the Windows CryptoAPI, a vulnerability assessment software company had *"written libraries on top of that to present a prettier facade in front of it because it's a fairly difficult library to use the way it's delivered" (C15).*

5.2.4 Academic and Research Resources

Our interviews revealed differing views on the value of academic research resources. Participants in three out of 21 interviews said that they have referenced academic resources (e.g., attended academic conferences or read (attack) research papers) to either better understand cryptography or to keep up with advances in the field. However, other participants passionately voiced their lack of confidence in the relevance of cryptographic research to their organizations' real-world industry challenges. One participant commented, *"People out in academia are famous for claiming there are holes in this kind of stuff where they don't actually exist, because they don't configure things according to the recommendations" (C01).* A cryptography architect asserted that his company's testing methods for cryptographic implementations were more state-of-the-art than those described in the academic world: *"No, we don't reference academic papers. They're not where we are in understanding the test problem... So there's a six-year gap between the... methods that we developed being identified in academia" (C05).*

5.3 Development and Testing Rigor

Our interviews revealed that the organizations translated their commitment to security and their expertise into rigorous development and testing practices. Interestingly, when participants spoke about how they test the cryptographic components, they saw secure software development as the foundation to providing cryptographic functionality.

5.3.1 Formal Processes

Of our 21 interviews, 20 reported that their companies employed formal development and testing strategies (those that are structured and standardized within the company) to ensure that their products, including the cryptographic components, were secure, while one participant said that they contracted developers to do that for them.

The development and testing practices were often the result of an evolutionary process spanning many years, as noted by 10 interviews. A director of quality assurance remarked, *"Part of [the role of] our test lead is now to verify that we're at the appropriate levels from a security standpoint... so it's a big focus for us now, whereas in the past, it was kind of a side item" (C14).* A company founder described his organization's introduction of more robust techniques over time:

> *"And it was an evolution, honestly. It started to where we didn't have hardly anything and new tools came on the market, as well as we had more time to focus on it. That allowed us to kind of improve code incrementally over time." (C10)*

Twelve interviews revealed a strong security mindset when they described developing and testing their products' cryptographic components based on risk. They would carefully build threat models to protect against strong adversaries, perform penetration testing, and would monitor current vulnerabilities to ensure their products were secure against those. A cryptographic design reviewer commented on this risk-based approach: *"All we can do is try to build good adversary models and then try to determine if our systems can stand up to those adversaries" (C04).*

Another discussed how his company's practices ensured that security had been considered throughout development:

> *"We have architects that do security reviews, that do threat modeling. And it's not just about the crypto, but more in general, how do you use the product. Who gets to do what? What are the risks? How do we mitigate those risks?... And one of the items for the engineering gate release is making sure... we mitigated anything that needed to be mitigated." (C11)*

Generally, the organizations' development and testing processes adhered to the principles in Figure 1. First, security specifications, architectures, and designs would be developed and reviewed. These would then be programmed against. Internal or external code reviews would be subsequently be conducted. During testing, tests would be run using internally developed test vectors or those provided with cryptographic resources. Static analysis tools and testing tools were also generally used. This development and testing process would be iterated whenever functionality changed, and performed in an expedited fashion if updates were being made due to a discovered security vulnerability.

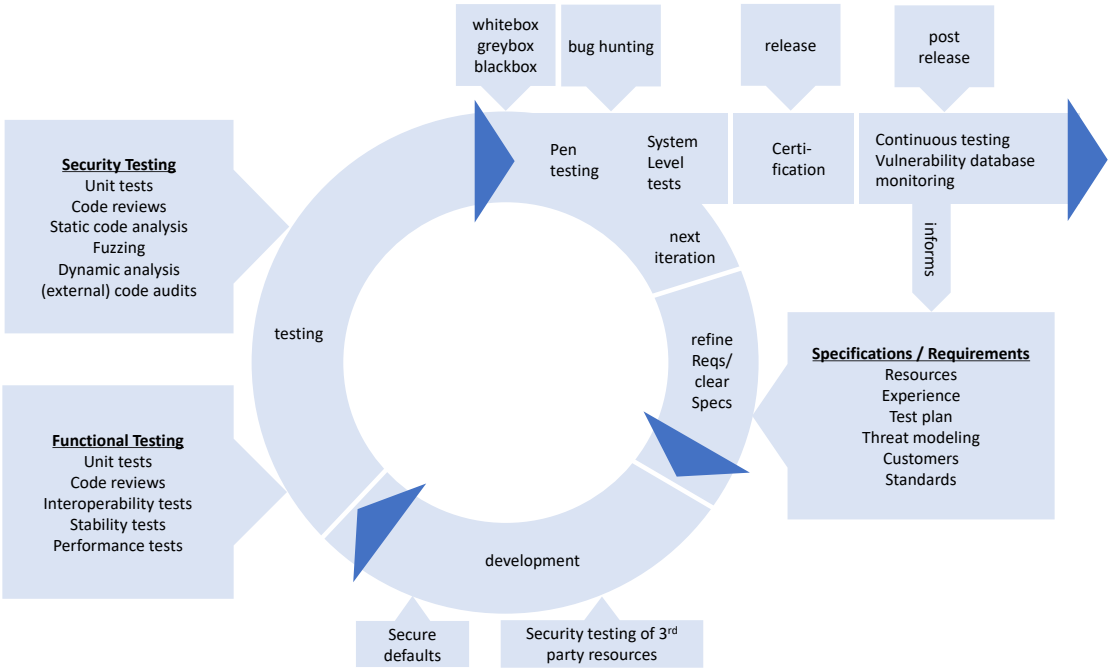


Figure 1: Security Development Lifecycle [34] adapted to include development and testing strategies as extracted from the interviews. Not all processes apply to all interviewed organizations.

5.3.2 *Development*

As mentioned above, the development phase for the organizations followed typical, commonly accepted development practices such as requirements and risk analysis and programming. We specifically highlight two practices that were mentioned most often in the interviews.

Participants in nine interviews spoke about design reviews, which were critical for finding potential errors in cryptographic components early in the process. Those that do cryptographic review must be highly skilled and able to piece together others' thought processes:

> *"A lot of the review is really just archeology. It's delving down into what they're producing, trying to understand both the explicit and the implicit assumptions, and then identifying where there are conflicts that lead to attacks." (C04)*

Nine organizations also mentioned the importance of doing code reviews to look for security and functional errors and vulnerabilities. A participant from a security software company remarked, *"We have a mandatory and systematic code review. Each line of code and each comment of code needs to be reviewed by usually at least two peers" (C17).*

5.3.3 *Testing*

Figure 2 shows the types of testing mentioned in the interviews. In 16 interviews, automated testing was discussed, often as being integrated with manual testing. The CTO of a company that produces security software discussed this integration of automated and manual testing: *"We have automatic tests, unit tests, integration tests, functional tests . . . , code analysis. . . . We have additional, manual tests being done. . . on top of the automated tests" (C17).*

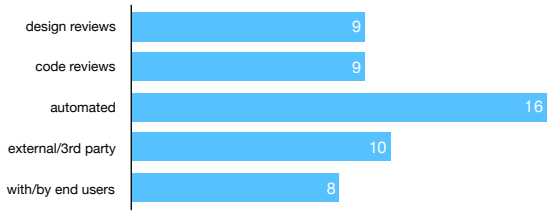


Figure 2: Development and testing practices explicitly mentioned for secure cryptography.

Ten interviews mentioned the use of third-party testing to improve the security of their cryptographic products. For example, organizations used bug-hunting services, or external testing such as blackbox, greybox, or whitebox penetration tests to increase the chances that bugs would be found in a controlled environment, and prior to product release. One participant described the benefit of his company using a bug-finding platform:

> *"You have some complete geniuses there that have found things that we never would have found. . . I feel like you're way more trustworthy if you are actually upfront about this stuff and you are actively soliciting people to attack you and paying them for their effort. You get a much higher confidence that some random person attacking won't just find something easy." (C10)*

Third-party review can also be used to gain trust with customers as expressed by a product manager: *"When customers ask us. . . 'Can you prove to me that it's done securely?' we can point to another organization that's independent to show" (C11).*

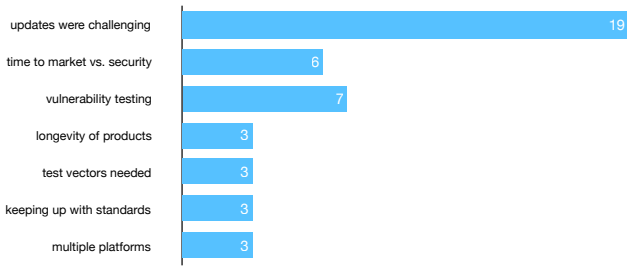


Figure 3: Challenges in development and testing.

End-user testing was not deemed as important for some organizations as they mostly develop products that become components in other products. However, this type of testing is critical for those producing products that will be directly used by businesses and consumers, and was discussed in eight interviews. These interviews mentioned formal beta-testing, continuous feedback, employing a user testing service, or recruiting convenience samples to test the product. One participant described the importance of user testing for his organization's consumer security software:

> *"We'd bring in our friends and family and sit them down and watch them. And it was eye-opening. It caused us to change a lot of what we did because, essentially, they didn't get the concept... We also used usertesting.com, which has been very great. You can show 10 people something and you have a pretty good idea." (C10)*

Another company takes advantage of beta testing to identify potential issues in their product: *"We have a very extensive beta program, and we have a very active customer base... so we have no lack of feedback" (C15).*

5.3.4 Challenges

All 21 of our interviews mentioned challenges to development and testing (Figure 3). We focus here on challenges directly related to cryptography, excluding challenges that have already been discussed, e.g., cryptographic complexity.

One challenge mentioned in six interviews was the tension between getting a product to market and taking the time to do robust security development and testing. For example, a participant from a company that spends years securely developing its cryptographic hardware modules observed that in most of the hardware/software industry, *"The design focus is simply not to do a solid job which will last a long time cryptographically.... They cannot care because they have to get the products out in a timely fashion" (C05).*

Testing for vulnerabilities in cryptographic implementations was another challenge mentioned in seven interviews, with three expressing concern for adequate testing of side-channel attacks. A participant who integrates cryptography into IoT devices said, *"especially in the embedded world, what the test vectors don't address I think is side-channel attacks, [which] could be really detrimental to the embedded device" (C08).*

Another challenge, as mentioned in three interviews, was the longevity of cryptographic products in customer spaces. An IoT researcher commented, *"Many devices are going to be deployed for 20 years. So, maybe the crypto in 20 years is no longer secure" (C09).* Another participant expressed the difficulty of having to maintain legacy cryptographic algorithms in a product: *"Old things become weak and you shouldn't use them anymore, and you need to add new ones ... However, customers are not so cooperative... It may take 10 years before everybody stops using something" (C01).*

Four interviews discussed challenges in having to troubleshoot or update third-party cryptographic implementations when vulnerabilities or errors are found. A principal engineer at a security software company described, *"when we're using any third-party library...and you happen to do something, and it fails, it's really hard to figure out what went wrong" (C15).*

Other cryptography-related challenges included the need for more test vectors (3 interviews), keeping up with changes in cryptographic standards (3), and having to use cryptographic libraries on multiple platforms (3).

In spite of these challenges, organizations in our study reported confidence in their processes and the resulting security of their cryptographic products (16 interviews). For example, a senior systems analyst at a small company described his confidence in the cryptographic algorithm they had developed: *"I don't make any bold claims, but at the same time, we looked at our encryption algorithm and we considered it quantum-proof" (C06).* In another interview, a company founder stated, *"I think we are definitely in the higher echelon for going above and beyond" (C10).*

6. IMPLICATIONS

6.1 Expanding Research Contexts

Our results suggested that the organizations believe they have a mature workforce, appreciate the complexity of crypto, possess a strong security culture, effectively use cryptographic resources, and practice secure development. However, the various studies mentioned in Related Work (e.g., [1, 2, 17–19, 42, 48]) indicate that there are many poor cryptographic implementations "in the wild," and developers typically lack a fundamental understanding of cryptography.

Where, then, is the disconnect between our findings and past research? It is possible that our self-report data are merely perceptions and do not accurately reflect the security mindsets of the organizations. Participants may be overconfident or may have overstated their organizations' security practices because of observer bias. We also recognize the value of future research to verify organizational claims by examining vulnerability databases, for example Common Vulnerabilities and Exposures (CVE) [68], to enumerate security issues in their products. Additionally, knowledge of an organization's development maturity level (e.g. Capability Maturity Model [12]) could be used as a comparison point.

Alternatively, this population is likely quite distinctive from previously studied developer populations and contexts, which may explain some of the differences in our findings. First, our participants exhibited more maturity in security and cryptography, with all having more than 10 years of security development experience. Second, organizational culture and constructs were an important driver of security mindsets within our study, while previous studies often involved

independent application developers, many of whom were not full-time developers or had not received formal education or organizational training in programming, cryptography or security. Third, there was a marked difference in the types of cryptographic resources used. Other studies (e.g., [2]) indicated that developers are reliant on information gleaned from search engines and Stack Overflow, which was only mentioned once in our interviews. Instead, the organizations in our study turned to more authoritative resources such as cryptographic standards and certification specifications. Finally, many of the studies that identified cryptographic vulnerabilities examined mobile apps, presumably because these were easy to obtain from public application stores, and open source projects. Our participants were developing more complex, expensive software and hardware products. Lack of security in these products had greater consequences, with the potential of harming the company's reputation or resulting in loss of sales.

These differences suggest that perhaps the security research community is not capturing a comprehensive picture of the cryptographic development environment. This demonstrates the need for the research community to diversify their study populations and contexts, and consider mechanisms to bridge the gap between more security-mature and less-skilled developers who implement cryptography in their products.

6.2 Support for Other Populations

The evidence of the criticality of organizational security culture and collaboration during development and testing raises the question of whether it is even possible for "solo" developers, such as the application developers with little cryptography experience or peer support represented in previous studies, to be truly successful in this area. How then can the research community explore ways to facilitate the transfer of strong security practices observed in some organizations to others with less support and experience?

As previously stated, unlike the population in our study, many developers sampled in past studies rely on online communities such as Stack Overflow [65] when implementing cryptography. But there is little evidence that these communities provide the level of support necessary for secure development. Future research may involve further assessing the value of current online communities for cryptographic development and exploring alternatives as means by which security mindsets can be created and perpetuated.

The findings also reveal that mentoring and peer review are critical to perpetuating security mindsets within organizations. Past studies have sought to understand the effectiveness of software development mentoring, peer review, and pair programming (e.g., [7, 47, 74]). However, more work needs to be done to determine whether mentoring for cryptographic development requires different tactics and how to best support this outside of organizational constructs.

6.3 Cryptographic Resource Usability

Whereas the bulk of responsibility in producing secure cryptographic products lies with the organizations themselves, our results imply that cryptographic resource providers can also do more to contribute to developer confidence. The most mentioned complaint our participants had with standards and certification guidance was the complexity of the language. This underlies a need for standards organizations to work towards a common language between cryptography experts who write the standards and developers and engineers who use them. Although standards documents may not be the appropriate place for large amounts of explanatory text, supplementary guidance that contains more instruction, cautions against common errors, and provides example implementations may prove to be valuable. Just as research has been done on language for security alerts and warnings (e.g. [10, 66]), it would also be helpful for researchers to explore the efficacy of language that explains cryptography concepts to non-experts.

Additionally, developers could benefit from more explanations of motivation, in other words, the "why" behind cryptographic choices. One participant echoed this recommendation as an important way to move beyond the "checkbox" mentality of standards: *"So you're thinking big picture, 'I'm doing this for this a reason,' because otherwise, you just get in the cookbook approach of, 'Do I meet this? Yes, yes, yes. Check, check, check' " (C19).*

Of course, many developers have no need to look at the standards directly since they use third-party implementations. However, third-party implementations may also be difficult to interpret and use securely if one lacks basic knowledge of cryptography. Our study results support past research calling for increased usability of cryptographic APIs. Similar to the work of Montandon et al. that proposed a new platform for providing API code examples [46], we also recommend investigating new approaches to community vetting of sample code since developers often copy flawed code snippets from forums such as Stack Overflow [2].

Notably, the participants spoke of a security problem with FIPS 140-2: updating certified software for security would break its certification, so companies relying on the certification had to decide between withholding an update or having to undergo recertification. This insight into an instance where reliance on a certification can decrease security underlines the need to closely and continuously involve cryptographic experts in the certification process. Their experience as users of the certification can help evolve the process and shape it to be more resilient, usable, and secure.

7. CONCLUSION

Our study offers new insight into the cryptographic development and testing practices of a previously unstudied population of organizations and participants who were highly experienced in cryptographic development. Our results suggest that organizational security mindsets are based on maturity and a strong security culture, which in turn guide selection of cryptographic resources and inform rigorous development and testing practices. Based on these observations, we see opportunities for development organizations, cryptographic resource providers, and security researchers to facilitate an environment conducive to building the expertise required to correctly and securely implement cryptography.

Disclaimer

Certain commercial companies or products are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the companies or products identified are necessarily the

best available for the purpose.

Acknowledgements

The authors would like to thank the following individuals who offered insightful comments that helped improve the quality of this paper: the anonymous reviewers, Curt Barker, Sascha Fahl, Simson Garfinkel, Michelle Mazurek, Matthew Smith, Brian Stanton, and Jeff Voas.

8. REFERENCES

[1] Y. Acar, M. Backes, S. Fahl, S. Garfinkel, D. Kim, M. Mazurek, and C. Stransky. Comparing the usability of cryptographic APIs. In *Proceedings of the 38th IEEE Symposium on Security and Privacy*, 2017.

[2] Y. Acar, M. Backes, S. Fahl, D. Kim, M. L. Mazurek, and C. Stransky. You get where you're looking for: The impact of information sources on code security. In *Proceedings of the 37th IEEE Symposium on Security and Privacy*, pages 289–305, May 2016.

[3] American National Standards Institute. ANSI. `https://www.ansi.org/`, 2018.

[4] S. Arzt, S. Nadi, K. Ali, E. Bodden, S. Erdweg, and M. Mezini. Towards secure integration of cryptographic software. In *Proceedings of the 2015 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software (Onward! '15)*, pages 1–13, 2015.

[5] R. S. Barbour. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *British Medical Journal*, 322(7294):1115–1117, 2001.

[6] C. A. Barry, N. Britten, N. Barber, C. Bradley, and F. Stevenson. Using reflexivity to optimize teamwork in qualitative research. *Qualitative Health Research*, 9(1):26–44, 1999.

[7] A. Begel and B. Simon. Novice software developers, all over again. In *Proceedings of the Fourth International Workshop on Computing Education Research*, pages 3–14, 2008.

[8] D. J. Bernstein, T. Lange, and P. Schwabe. The security impact of a new cryptographic library. In *Proceedings of the International Conference on Cryptology and Information Security (LatinCrypt '12)*, pages 159–176, 2012.

[9] E. Bonver and M. Cohen. Developing and retaining a security testing mindset. *IEEE Security & Privacy*, 6(5), 2008.

[10] C. Bravo-Lillo, L. F. Cranor, J. Downs, and S. Komanduri. Bridging the gap in computer security warnings: A mental model approach. *IEEE Security & Privacy*, 9(2):18–26, 2011.

[11] H. Brenner and U. Kliebsch. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, pages 199–202, Mar. 1996.

[12] CMMI Institute. What is capability maturity model integration (CMMI)? `http://cmmiinstitute.com/capability-maturity-model-integration`, 2018.

[13] J. Corbin and A. Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, Thousand Oaks, CA, 4th edition, 2015.

[14] Dell Incorporated. RSA Conference: Where the world talks security. `https://www.rsaconference.com/`, 2018.

[15] K. DeSwert. Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha. Technical report, Center for Politics and Communication, 2012.

[16] S. I. Donaldson and E. J. Grant-Vallone. Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology*, 17(2):245–260, 2002.

[17] T. Duong and J. Rizzo. Cryptography in the web: The case of cryptographic design flaws in ASP.NET. In *Proceedings of the 31st IEEE Symposium on Security and Privacy*, pages 481–489, May 2011.

[18] M. Egele, D. Brumley, Y. Fratantonio, and C. Kruegel. An empirical study of cryptographic misuse in Android applications. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security (CCS '13)*, pages 73–84, 2013.

[19] S. Fahl, M. Harbach, T. Muders, L. Baumgartner, B. Freisleben, and M. Smith. Why Eve and Mallory love Android: An analysis of Android SSL (in)security. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS '12)*, pages 50–61, 2012.

[20] C. Forler, S. Lucks, and J. Wenzel. Designing the API for a cryptographic library: A misuse-resistant application programming interface. In *Proceedings of the 17th Ada-Europe International Conference on Reliable Software Technologies (Ada-Europe '12)*, pages 75–88, 2012.

[21] D. Freelon. Recal3: Reliability for 3+ coders. `http://dfreelon.org/utils/recalfront/recal3/`.

[22] D. G. Freelon. ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science*, 5(1):20–33, 2010.

[23] R. Garcia, J. Thorpe, and M. Martin. Crypto-Assistant: Towards facilitating developer's encryption of sensitive data. In *Proceedings of the 12th Annual International Conference on Privacy, Security and Trust (PST '14)*, pages 342–346, 2014.

[24] Gartner. IT glossary: Small and midsize business (SMB). `http://www.gartner.com/it-glossary/smbs-small-and-midsize-businesses/`, 2017.

[25] M. Georgiev, S. Iyengar, S. Jana, R. Anubhai, D. Boneh, and V. Shmatikov. The most dangerous code in the world: validating SSL certificates in non-browser software. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, pages 38–49, Oct. 2012.

[26] B. G. Glaser and A. L. Strauss. *The Discovery of Grounded theory: Strategies for Qualitative Research*. Transaction Publishers, 2009.

[27] M. Green and M. Smith. Developers are not the enemy! The need for usable security APIs. *IEEE Security & Privacy*, 14(5):40–46, Sept. 2016.

[28] L. Greenemeier. NSA efforts to evade encryption technology damaged U.S. cryptography standard. `https://www.scientificamerican.com/article/nsa-nist-encryption-scandal/`, Sept. 2013.

[29] G. Guest, A. Bunce, and L. Johnson. How many

interviews are enough? An experiment with data saturation and variability. *Field Methods*, 18(1):59–82, 2006.

[30] P. Gutmann. Lessons learned in implementing and deploying crypto software. In *Proceedings of the 2002 Usenix Security Symposium*, pages 315–325, 2002.

[31] J. M. Haney, S. L. Garfinkel, and M. F. Theofanos. Organizational practices in cryptographic development and testing. In *Proceedings of the 5th IEEE Conference on Communications and Network Security (CNS '17)*, Oct. 2017.

[32] B. Headd. The role of microbusinesses in the economy. `https://www.sba.gov/sites/default/files/`, Feb. 2015.

[33] R. Hodson. *Analyzing Documentary Accounts (No. 128)*. Sage Publications, 1999.

[34] M. Howard and S. Lipner. *The security development lifecycle Vol. 8*. Microsoft Press, Redmond, WA, 2006.

[35] Institute of Electrical and Electronics Engineers. IEEE Standards Association. `http://standards.ieee.org/`, 2018.

[36] International Organization for Standardization. Standards. `https://www.iso.org/standards.html`, 2018.

[37] Internet Engineering Task Force. IETF. `https://www.ietf.org/`, 2018.

[38] S. L. Kanniah and M. N. Mahrin. A review on factors influencing implementation of secure software development practices. *World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, 10(8):3022, 2016.

[39] K. Krippendorff. *Content Analysis: An Introduction to its Methodology*. Sage, 2004.

[40] D. Lazar, H. Chen, X. Wang, and N. Zeldovich. Why does cryptographic software fail?: A case study and open problems. In *Proceedings of the 5th Asia-Pacific Workshop on Systems (APSys '14)*, pages 7:1–7:7, 2014.

[41] D. Leaman. National Institute of Standards and Technology: NVLAP Common Criteria testing. NISTHB 150-20. `https://www.nist.gov/sites/default/files/documents/nvlap/NIST-HB-150-20-2014.pdf`, 2014.

[42] Y. Li, Y. Zhang, J. Li, and D. Gu. iCryptoTracer: Dynamic analysis on misuse of cryptography functions in iOS applications. In *Proceedings of the International Conference on Network and System Security*, pages 349–362, Oct. 2014.

[43] G. McGraw. Software security. *IEEE Security & Privacy*, 2(2):80–83, 2004.

[44] C. G. Menk. System security engineering capability maturity model and evaluations: Partners within the assurance framework. `https://csrc.nist.gov/csrc/media/publications/conference-paper/1996/10/22/proceedings-of-the-19th-nissc-1996/documents/paper010/cmmtpep.pdf`, 1996.

[45] S. Merriam and E. Tisdell. *Qualitative Research: A Guide to Design and Implementation*. John Wiley & Sons, San Francisco, CA, 4 edition, 2016.

[46] J. E. Montandon, H. Borges, D. Felix, and M. T. Valente. Documenting APIs with examples: Lessons learned with the APIMiner platform. In *Proceedings of the 20th IEEE Working Conference on Reverse Engineering (WCRE)*, pages 401–408, 2013.

[47] M. M. Müller. Two controlled experiments concerning the comparison of pair programming to peer review. *Journal of Systems and Software*, 78(2):166–179, 2005.

[48] S. Nadi, S. Krüger, M. Mezini, and E. Bodden. Jumping through hoops: Why do Java developers struggle with cryptography APIs? In *Proceedings of the 38th International Conference on Software Engineering (ICSE '16)*, pages 935–946, 2016.

[49] National Institute of Standards and Technology. FIPS Pub 140-2: Security requirements for cryptographic modules. `http://csrc.nist.gov/publications/fips/fips140-2/fips1402.pdf`, 2001.

[50] National Institute of Standards and Technology. Cryptographic module validation program. `https://csrc.nist.gov/projects/cryptographic-module-validation-program`, 2016.

[51] National Institute of Standards and Technology. Cryptographic algorithm validation program (CAVP). `="http://csrc.nist.gov/groups/STM/cavp/`, 2017.

[52] National Institute of Standards and Technology. Cryptographic standards and guidelines. `https://csrc.nist.gov/Projects/Cryptographic-Standards-and-Guidelines`, 2018.

[53] P. Nguyen. Can we trust cryptographic software? Cryptographic flaws in GNU privacy guard v1. 2.3. In *Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques*, pages 555–570, 2004.

[54] Open Web Application Security Project. OWASP secure software development lifecycle project. `https://www.owasp.org`, 2017.

[55] OpenSSL Software Foundation. OpenSSL cryptography and SSL/TLS toolkit. `https://www.openssl.org/`, 2018.

[56] M. Q. Patton. *Qualitative research*. John Wiley & Sons, San Francisco, CA, 2005.

[57] PCI Security Standards Council. PCI Security. `https://www.pcisecuritystandards.org/pci_security/`, 2018.

[58] B. Potter and G. McGraw. Software security testing. *IEEE Security & Privacy*, 2(5):81–85, 2004.

[59] J. Saldaña. *The Coding Manual for Qualitative Researchers*. Sage, 3 edition, 2015.

[60] T. Schlienger and S. Teufel. Information security culture-From analysis to change. *South African Computer Journal*, (31):46–52, 2003.

[61] B. Schneier. Why cryptography is harder than it looks. `https://www.schneier.com/essays/archives/1997/01/why_cryptography_is.html`, 1997.

[62] B. Schneier. The security mindset. `https://www.schneier.com/blog/archives/2008/03/`, 2008.

[63] D. Seltzer-Kelly, S. J. Westwood, and D. M. Peña-Guzman. A methodological self-study of quantitizing: Negotiating meaning and revealing multiplicity. *Journal of Mixed Methods Research*, 6(4):258–274, 2012.

[64] S. Shuai, D. Guowei, G. Tao, Y. Tianchang, and

S. Chenjie. Modelling analysis and auto-detection of cryptographic misuse in Android applications. In *Proceedings of the 12th IEEE International Conference on Dependable, Autonomic, and Secure Computing*, pages 75–80, Aug 2014.

[65] Stack Exchange. Stack overflow. `https://stackoverflow.com/`, 2018.

[66] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor. Crying wolf: An empirical study of SSL warning effectiveness. In *USENIX Security Symposium*, pages 399–416, 2009.

[67] A. S. Tanenbaum. *Computer Networks: 2Nd Edition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[68] The MITRE Corporation. Common vulnerabilities and exposures (CVE). `https://cve.mitre.org/`, 2018.

[69] D. R. Thomas. A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2):237–246, June 2006.

[70] Underwriters Laboratories (UL). Certification. `https://services.ul.com/categories/certification/`, 2018.

[71] US-CERT. OpenSSL Heartbleed vulnerability (CVE-2014-0160). `https://www.us-cert.gov/ncas/alerts/TA14-098A`, 2014.

[72] Veracode. The state of software security. `https://www.veracode.com/sites/default/files/Resources/Reports/state-of-software-security-volume-7-veracode-report.pdf`, 2016.

[73] Veracode. Veracode secure development survey: Developers respond to application security trends. `https://info.veracode.com/report-veracode-developer-survey.html`, 2016.

[74] L. Williams, R. R. Kessler, W. Cunningham, and R. Jeffries. Strengthening the case for pair programming. *IEEE Software*, 17(4):19–25, 2000.

[75] S. Xiao and E. Witschey, J.and Murphy-Hill. Social influences on secure development tool adoption: Why security tools spread. In *Proceedings of the 17th ACM conference on Computer supported Cooperative Work and Social Computing*, CSCW '14, pages 1095–1106, Feb. 2014.

[76] J. Xie and B. Lipford, H. R.and Chu. Why do programmers make security errors? In *Proceedings of the 2011 IEEE Symposium on Visual Languages and Human-Centric Computing*, pages 161–164, Sept. 2011.

APPENDIX

A. INTERVIEW QUESTIONS

1. Can you tell me about your organization - what it does, what it produces?
2. What is your role within your organization with respect to cryptographic products?
3. How did you get into this field?
   (a) At what point and why did you become concerned with cryptography and secure development?
   (b) In which field(s) is your formal education?
4. Do you work in a unit or department that is part of a larger organization?

   If yes : What is the size of the unit or department?
   (a) What is the size of your overall organization?
5. Can you tell me about the kinds of products your organization develops, and specifically those that use cryptography?
6. Who are the typical customers for your products that use cryptography?
7. How long has your organization been working on products that use cryptography?
8. Is cryptography your organization's primary business focus, or is it an enabler within your products?
9. For your products that use cryptography, what processes or techniques , if any, does your organization use to minimize bugs and errors in code during the development process?
   (a) Why does your organization choose to use these methods? [only use if participant has difficulty coming up with response:] for example, industry standard, customer demand, robustness and quality
10. What processes or techniques does your organization use to test and validate the cryptography component in your products?
    (a) Why does your organization choose to use these methods? [only use if participant has difficulty coming up with response:] for example, industry standard, customer demand, robustness and quality
    (b) What kind of end-user testing, if any, does your organization do to prevent customers from misconfiguring or misusing the cryptography component in your products?
11. Does your organization do any certifications or third-party testing?
    (a) What reasons led you to decide to use certifications or third-party testing?
    (b) How do you establish confidence in the results of the certifications or third-party testing?
    (c) What are the challenges or issues your organization has experienced with certifications or third-party testing, if any?
12. What, if any, are your organization's biggest challenges with respect to developing and testing cryptography within your products?
    (a) How do you think these challenges can be overcome, if at all?
    (b) Has your organization experienced a tension between secure development and testing and getting a product to market? If so, how has that impacted your organization's processes?
13. Do your customers have specific requirements regarding development and testing? If so, what are those requirements?
14. How do updates impact your development and testing processes, if at all? (time-sensitive vs. deprecation)
15. What resources do you use to help you develop and test the cryptography component of your products? [only use if participant has difficulty coming up with response:] for example, standards, industry specifications, books, academic papers, standard libraries, APIs
    (a) What are the reasons your organization chooses to use those particular resources?

    If the participant does NOT use standards: What are the reasons that your organization does not use standards?
16. [If the participant uses standards:] What kinds of standards do you use?
    (a) What is the role of standards in your organization's development and testing processes?
    (b) What do you see as the value or benefit of using these standards, if any?
17. How could standards or other cryptographic resources be improved to be more useful?
    (a) How could NIST standards and guidance be improved to be more useful?
18. Is there anything else you'd like to add about the topics we've discussed?

B. CODEBOOK

Participant Demographics

Organization Demographics

Organization Characteristics

- Team Interactions
- Security Culture
- Maturity
- Talent/Hiring

Development and Testing Practices

- Formal
- Informal
- Risk-based
- Practices
  - Automated
  - Human/Manual
  - External/3rd party testing
  - End-user testing
- Reasons/Philosophy
- Evolution
- Confidence
- Challenges
- Updates

Certifications/Compliance Programs

- Which ones (identify)
- Problems and Challenges
- Reasons
- Confidence
- Improvements

Resources

- Standards
- Government
- Industry/3rd Party
- Internally developed
- Research
- Gaps

Security

- Vulnerabilities and Errors
- Usability and Complexity
- Relationship and Tensions

Security Education

- Customers
- Developers/Engineers

Emotions

- Positive
- Negative

Influences

Customers

Evolution of Security Field

Complaints

Participant Values and Perceptions

Trust

A Comparative Usability Study of Key Management in Secure Email

Scott Ruoti
University of Tennessee
ruoti@utk.edu

Jeff Andersen
Brigham Young University
andersen@isrl.byu.edu

Tyler Monson
Brigham Young University
monson@isrl.byu.edu

Daniel Zappala
Brigham Young University
zappala@cs.byu.edu

Kent Seamons
Brigham Young University
seamons@cs.byu.edu

ABSTRACT

We conducted a user study that compares three secure email tools that share a common user interface and differ only by key management scheme: passwords, public key directory (PKD), and identity-based encryption (IBE). Our work is the first comparative (i.e., A/B) usability evaluation of three different key management schemes and utilizes a standard quantitative metric for cross-system comparisons. We also share qualitative feedback from participants that provides valuable insights into user attitudes regarding each key management approach and secure email generally. The study serves as a model for future secure email research with A/B studies, standard metrics, and the two-person study methodology.

1. INTRODUCTION

The cryptography needed to deploy secure email is well studied and has been available for years, and a number of secure email systems have been deployed and promoted recently, including ProtonMail, Tutanota, Mailvelope, Virtru, Voltage, Encipher.it, etc. While some of these systems have millions of users, the vast majority of email users still do not use secure email [21]. The lack of adoption of secure email is often attributed to the significant gap between what the technology can offer and the ability of users to successfully use the technology to encrypt their emails.

Beginning with Whitten and Tygar [36], secure email usability studies have shown that key management is a significant hurdle for users. More recent usability studies (e.g., [1, 2, 23]) show signs that progress toward greater usability is being made, but limitations in each study make it difficult to draw conclusions regarding the impact key management has on secure email usability, other than the need for automation. We previously conducted studies [23, 24, 26, 27] that directly compared key management schemes from different families, but the systems implementing the various key management schemes were wildly different, introducing a significant confounding factor. Bai et al. [2] compared two key management schemes, but their study explored user mental models and trust, not usability generally.

Additionally, even though public key directories have recently received significant attention [19, 29], it is unclear how their usability compares to other key management schemes. Lerner et al. [18] studied a public key directory system but didn't use a standard metric, making it difficult to directly compare their results to past work. Atwater et al. [1] simulated a public key directory, but permitted a user to send an email to a recipient who had not yet generated a key pair. Normally, when a user attempts to send an email to a recipient who has not yet generated a key pair, they must wait until the user does so and uploads their public key to the key directory. Because this affected numerous participants in their study, it is unclear how this issue impacted their results.

Our work was motivated by the desire to build on these earlier studies and reduce the number of confounding factors in order to increase our confidence in the resulting usability measurements. In this paper, we describe a user study comparing three key management schemes, taken from different families, to better understand how key management impacts the usability of secure email during initial setup and first use of the system. Using the MessageGuard research platform, we built three secure email tools which differ only in the key management scheme they implement (passwords, public key directory [PKD], and identity-based encryption [a]), reducing potential confounding factors in the study. In our study design we used a standard metric, allowing comparison to results from past studies. Finally, we replicated our earlier paired participant study setup [23], allowing us to evaluate grass roots adoptability.

In total, 47 pairs of participants completed our study. All three systems received favorable ratings from users, with server-derived public keys being considered the most usable, followed by user-generated public keys, and finally shared secrets. Each system performed better than similar (i.e., same key management) systems previously studied in the literature. Users also provided valuable qualitative feedback helping identify pros and cons of each key management scheme.

The contributions of this paper include:

1. **First A/B evaluation of key management using standard metrics.** Our study was able to confirm Atwater et al.'s [1] findings that public key directories

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

are usable. Additionally, we find evidence that the secure email design principles we identified in previous work [24] generalize beyond server-derived public keys.

2. **The MessageGuard platform.** To enable this work, we built MessageGuard, a research platform for building secure email and other end-to-end encryption prototypes. MessageGuard significantly simplifies the effort required to work in this space and provides a means whereby research results may be shared and replicated. MessageGuard has a pluggable architecture, making it easy to build prototype variants for use in A/B testing.

3. **Lessons learned and recommendations.** Our study elicits user attitudes regarding the three key management schemes we evaluate, including security and usability trade-offs identified by participants. For example, even after understanding that the user-generated public key scheme protects against a stronger threat model than server-derived public keys, many users indicate that they do not need that level of security and prefer server-derived public keys because they can immediately send email without waiting for the recipient to generate a public/private keypair. Based on our findings, we give recommendations for future work.

2. BACKGROUND

In this section, we first describe several key management schemes commonly used with end-to-end email encryption. Next, we provide a chronological review of usable secure email research.

2.1 Key Management

We study three families of key management schemes used in end-to-end encryption of email: shared secrets, user-generated public keys, and server-derived public keys. Each has different methods for creating, sharing, and linking cryptographic keys to email addresses. We describe each scheme briefly; a more complete treatment can be found in [9].

2.1.1 Shared Secrets

Users can encrypt their emails using symmetric keys derived from a secret shared between pairs of users. Most commonly, these secrets are in the form of simple passwords, which are more readily communicated and remembered by users than cryptographically secure random values. The security of this key management scheme is dependent on users' ability to satisfy the following requirements when they create and share passwords: (1) choose a unique password for each user they will communicate with, (2) choose passwords that will resist a brute-force attack, (3) communicate passwords over a secure channel, and (4) safely store passwords.

2.1.2 User-generated Public Keys

Before sending or receiving encrypted email, users must first generate a cryptographic key pair. A user's private key should never be shared with any other party and must be safely stored by the user. The user's public key, with relevant metadata, is then distributed to other users in a number of ways, such as sending the key directly to other users, posting the key to a personal website, or uploading the key to a key directory.

There are numerous ways to verify the authenticity of a public key (i.e., the binding of a public key to an email address), some of which include:

1. **Manual validation.** Users can directly communicate with each other and directly share their public key or compare key fingerprints[1]. Users are expected to know each other personally and thus be able to confirm the identity of those they are communicating with.

2. **Web of trust.** Users can have their public key signed by one or more other users, who are expected to only sign public keys that they have verified using manual key validation. When retrieving a public key, users check to see if it has been signed by a user they trust to have validated it properly. Users may choose to transitively trust public keys that are trusted by users they trust, forming a web of trust.

3. **Hierarchical validation.** Users can have their public key signed by an authoritative signer (e.g., a certificate authority), which will only sign a public key after verifying that the user who submitted it owns the associated email address. When retrieving a public key, its signature is validated to ensure that it was properly signed by an authoritative signer. This method of key validation is most commonly associated with S/MIME [8, 13].

4. **Public key directory.** Users can submit their public keys to a trusted key directory. This directory will only accept and disseminate public keys for which it has verified that the user who submitted the key owns the associated email address. Due to its trusted nature, keys retrieved from the directory are assumed to be authentic. The behavior of the key directory can be audited through the use of certificate transparency [29] or a CONIKS-like ledger [19].

Manual verification and the web of trust are commonly associated with PGP [12], though any of the above can be used with PGP.

The security of these schemes depends on the ability of users to protect their private keys, obtain necessary public keys, and faithfully validate these public keys. If users lose access to their private keys (e.g., disk failure with no backup), they will be unable to access their encrypted email.

2.1.3 Server-derived Public Keys

In this scheme, a user's public key is generated for them by a server they trust, which may also store their private key (called key escrow). This alleviates the problems associated with a user losing their private key, and is often used in corporate environments. A variant of this scheme is identity-based encryption (IBE) [31]. With IBE, a user's public key is generated mathematically based on their e-mail address and public parameters provided by an IBE key server. A user's private key is also generated by the IBE key server, which will only release that key to the user after the user verifies ownership of the associated email address. In any situation

[1] A public key's fingerprint is typically derived from a cryptographic hash of the public key.

when a user cannot trust a server with their private key (e.g., an activist in an oppressive regime, or a journalist that needs to protect sources) key escrow should not be used.

2.2 Usable Secure Email

Whitten and Tygar [36] conducted the first formal user study of a secure email system (PGP 5 with manual key validation), uncovering serious usability issues with key management and users' understanding of the underlying public key cryptography. The results of their study took the security community by surprise and helped shape modern usable security research.

Garfinkel and Miller [13] created a secure email system using S/MIME (hierarchical key validation) and demonstrated that automating key management provides significant usability gains. However, their study also revealed that the tool "was a little too transparent," leading to confusion and mistakes.

We previously created Private WebMail (Pwm) [27], a secure email system that tightly integrates with Gmail and uses identity-based encryption (IBE) to provide key management that is entirely transparent to users. User studies of Pwm demonstrate that it was viewed very positively by users, and significantly outperformed competing secure email systems.

Atwater et al. [1] compared the usability and trustworthiness of automatic versus manual encryption, finding that there were no significant differences between the two approaches. As part of this study, Atwater et al. developed two email clients—one integrated with Gmail and one standalone—both of which simulated the user experience of using a public key directory.

We also developed a novel two-person methodology [23] for studying the usability and grassroots adoptability of secure email. In particular, this study involved recruiting pairs of recipients (e.g., friends, spouses), who would then be responsible for sending secure email among themselves. Compared to single-participant studies, this methodology revealed differences between the experience of initiating others and being initiated by others into using secure email. Our study compared systems using three different families of key management: shared password, public key directory, and IBE; unfortunately, confounding factors in this study make it difficult to draw any conclusion on how key management affects secure email's usability.

Bai et al. explored user attitudes toward different models for obtaining a recipient's public key in PGP [2]. In their study, they built two PGP-based secure email systems, one that used manual key validation and one that used a public key directory. Users were provided with instructions on how to use each tool and given several tasks to complete. The results of this study showed that, overall, individuals recognized the security benefits of manual key validation, but preferred the public key directory and considered it to have sufficient security. While this study gathered data on user attitudes regarding two key management schemes, it did not evaluate their usability.

More recently, we further refined our Pwm system [24], identifying four design principles that increase the usability, correct behavior, and understanding of secure email: (1) having informative and personalized initiation messages that guide users through installing the secure email software and give them confidence that the email they received is not malicious; (2) adding an artificial delay during encryption to build trust in the system and show users who their message is being encrypted for; (3) incorporating inline, context-aware tutorials to assist users as they are sending and receiving their first encrypted emails; and (4) using a visually distinctive interface to clearly demarcate which content is encrypted/to-be-encrypted and helping users avoid accidentally sending sensitive information in the clear.

	Whitten and Tygar [36]	Garfinkel and Miller [13]	Ruoti et al. [27]	Atwater et al. [1]	Ruoti et al. [23, 26]	Bai et al. [2]	Ruoti et al. [24]	Lerner et al. [18]	This work
Comparative			✓		✓	✓	✓		✓
A/B Study						✓			✓
Standard metric			✓	✓	✓		✓		✓
Two-person					✓				✓

Table 1: Comparison of Usable Secure Email Research

Lerner et al. [18] built Confidante, a secure email tool that leverages Keybase, a public key directory, for key management. A user study of Confidante with lawyers and journalists demonstrated that these users could quickly and correctly use the system.

The significant differences between this earlier work and our current work are summarized in Table 1.

3. SYSTEMS

To limit confounding factors in our study, it was necessary to build several secure email tools that differed only in how key management was handled. To accomplish this we substantially modified our Private WebMail 2.0 (Pwm 2.0) system [24], leaving its UI unchanged, but otherwise completely rewriting its codebase to add support for a pluggable key management subsystem. This allowed us to rapidly develop three secure email prototypes that only differed in how they handled key management, while keeping the remaining system components consistent. We call this pluggable version of Pwm 2.0 *MessageGuard*.

We choose to extend Pwm 2.0 for several reasons. First, it is an existing system with established favorable reviews, saving us a significant amount of development time and helping avoid the possibility of designing a new secure email tool that was viewed unfavorably by users. Second, it had the highest usability score [24] of any secure email systems evaluated using the System Usability Metric (SUS) [6]. Third, this allowed us to test whether the secure email design principles proposed by Ruoti et al. and implemented in Pwm 2.0 (see Section 2.2) generalize beyond IBE-based systems.

In addition to adding a pluggable key management system to MessageGuard, we also added several other features to MessageGuard in order to allow other researchers to use it as a research platform for building end-to-end encryption

prototypes. First, MessageGuard supports a wide range of non-email sites (e.g., Facebook, Twitter, Blogger), automatically scanning these pages for user-editable content and allowing users to encrypt this content end-to-end. Second, the page scanning functionality is pluggable, allowing researchers to create finely-tuned, per-site end-to-end encryption plugins. Finally, MessageGuard includes pluggable user interface, encryption, and content packaging subsystems.

There are three key benefits to using MessageGuard as a research platform:

1. *Accelerates the creation of content-based encryption prototypes.* MessageGuard provides a fully functional content-based encryption system, including user interfaces, messaging protocols, and key management schemes. The modular design of MessageGuard allows researchers to easily modify only the portions of the system they wish to experiment with, while the remaining portions continue operating as intended. This simplifies development and allows researchers to focus on their areas of expertise—either usability or security.
2. *Provides a platform for sharing research results.* Researchers who create prototypes using MessageGuard can share their specialized interfaces, protocols, or key management schemes as one or more patches, allowing researchers to leverage and replicate each other's work. Additionally, research can be merged into MessageGuard's code base, allowing the community to benefit from these advances and reducing fragmentation of efforts.
3. *Simplifies the comparison of competing designs.* MessageGuard can be used to rapidly develop prototypes for use in A/B testing. Two prototypes built using MessageGuard will only differ in the areas that have been modified by researchers. This helps limit the confounding factors that have proven problematic in past comparisons of content-based encryption systems.

The source code for MessageGuard is available at `https://bitbucket.org/account/user/isrlemail/projects/MES`.

In the remainder of this section, we give a brief overview of MessageGuard. Additional details are available in Appendix A–C, and a complete description can be found in a technical report [25]. Next, we describe the workflow for the three secure email variants that we created using MessageGuard. We chose well-known instances of each key management scheme and explain the rationale for that choice: passwords, public key directory (PKD), and IBE. Other alternatives and hybrids of these approaches are possible.

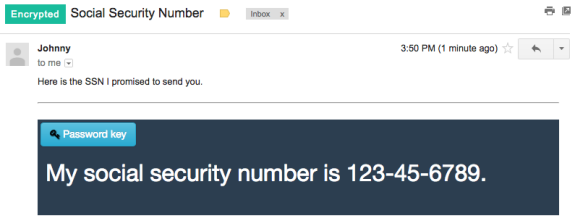
These systems can be downloaded and are available for testing at `https://{pgp,ibe,passwords}.messageguard.io`

3.1 MessageGuard

MessageGuard tightly integrates with existing web applications, in this case Gmail, using *security overlays*. Security overlays function by replacing portions of Gmail's interface with secure interfaces that are inaccessible to Gmail. Users then interact with these secure overlays to create and read encrypted email (Figure 1a and Figure 1b).

Figure 2 shows the MessageGuard architecture:

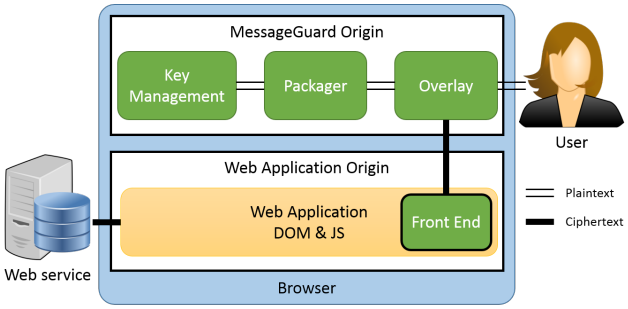
(a) Composition Overlay



(b) Read Overlay

Figure 1: MessageGuard Overlays

- The **front end** scans for encrypted payloads and data entry interfaces and replaces these items with a secure overlay. The front end is the only component that runs outside of MessageGuard's protected origin, and it can only communicate with overlays using the `window.postMessage` API. The overlay always encrypts user data before transmitting it to the front end component and sanitizes any data it receives from the front end. In addition, the front end also displays tutorials that instruct new users how to use MessageGuard. These are all context-sensitive, appearing as the user performs a given task for the first time.
- **overlays** use `iframes` and the browser's same-origin policy to keep plaintext from being exposed to the email server and its application. A read overlay displays sensitive information to the user, and a compose overlay allows users to encrypt sensitive information before sending it to the website. Overlays have a distinctive, dark color scheme that stands out from most websites, allowing users to easily identify secure overlays from insecure website interfaces.
- The **packager** encrypts/decrypts user data and encodes the encrypted data to make it suitable for transmission through web applications. The packager uses standard cryptographic primitives and techniques to encrypt/decrypt data (e.g., AES-GCM). Ciphertext is packaged with all information, save the key material, necessary for recipients of the message to decrypt it.
- The **key management** component enables a variety of key management schemes to be configured, without changing other aspects of MessageGuard such as the read or compose overlays.



A user's sensitive data is only accessible within the MessageGuard origin.

Figure 2: MessageGuard Architecture

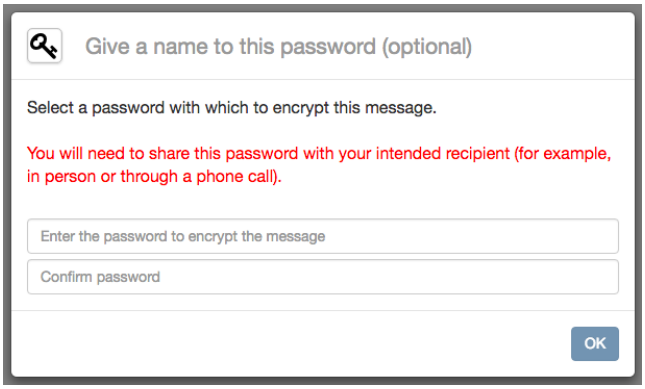


Figure 3: Dialog for Entering a New Password with Which to Encrypt Email.

3.2 Passwords

We choose to evaluate passwords as they are a scheme that should be familiar to users. The workflow for our password system is as follows:

1. The user visits the MessageGuard website. They are prompted to download the system.
2. After installation, the system is immediately ready for use.
3. When the user attempts to send an encrypted email, they are informed that they need to create a password for encrypting the email (see Figure 3). After creating the password, the user can send their encrypted email.
4. The user must communicate to the recipient the password used to encrypt the email message. This should happen over an out-of-band (i.e., non-email) channel.

3.3 Public Key Directory (PKD)

We choose to evaluate public key directories because they have received significant attention lately [2, 18, 19, 29]. The workflow for our public key directory system is as follows:

1. The user visits the MessageGuard website. They are instructed to create an account with their email address.[2] Their address is verified by having the user click a link in an email sent to them. They are then able to download the system.
2. After installation, the user is told that the system will generate a key pair for them. The public key is automatically uploaded to the key directory, as the user is already authenticated to the key directory from the previous step.
3. The user attempts to send an encrypted email but is informed that the recipient hasn't yet installed the system.[3] They are then prompted to send their recipient an email inviting them to install the system. This email message is auto-generated by MessageGuard, with the system able to add a custom introduction message if desired.
4. Once the recipient has installed the system, which generates and publishes their public key, they inform the sender that they are ready to proceed. The sender can now send their encrypted email.

3.4 Identity-based Encryption (IBE)

We choose to evaluate IBE because it is the key management scheme that has been shown to be most usable in past studies, providing a good baseline for this work. The workflow for our IBE-based system is as follows:

1. The user visits the MessageGuard website. They are instructed to create an account with their email address.[2] Their address is then verified by having the user click a link in an email sent to them. They are then able to download the system.
2. After installation, the user is informed that the system will retrieve their IBE key from the key server. This happens automatically because the user is already authenticated to the key server from the previous step.
3. The user can send encrypted email to any address.
4. The recipient, upon receiving the encrypted email, is prompted to visit the MessageGuard website and create an account. After their address is verified and their private key is downloaded from the key server, they can read the encrypted message.

4. METHODOLOGY

We conducted a within-subjects, IRB-approved lab study wherein pairs of participants used three secure email systems to communicate sensitive information to each other (study materials are found in Appendix D). Our study methodology is patterned after our previous paired participant methodology [23], allowing us to examine usability in the context of

[2]We chose to require a MessageGuard account in order to prevent a compromised email provider from being able to transparently upload (PKD) or download (IBE) cryptographic keys from the MessageGuard key server, which would be possible if these operations were only protected by email-based authentication.

[3]The recipient must install the system and use it to upload a public key before the sender can encrypt email for the recipient.

two novice users, without potential bias or other behaviors introduced by direct involvement with a study coordinator.

The study ran for two and a half weeks—beginning Monday, May 23, 2016, and ending Tuesday, June 7, 2016. In total, 55 pairs of participants (110 total participants) took the study. Due to various reasons discussed later in this section, we excluded results from eight participant pairs. For the remainder of this paper, we refer exclusively to the remaining 47 pairs (94 participants).

4.1 Study Setup

Participants took 50–60 minutes to complete the study, and each participant was compensated $15 USD in cash. Participants were required to be accompanied by a friend, who served as their counterpart for the study, and were instructed to use their own Gmail accounts.[4]

When participants arrived, they were given a consent form to sign, detailing the study and their rights as participants. Participants were informed that they would be in separate rooms during the study and would need to use email to share some sensitive information with each other. They were told that they were free to communicate with each other however they normally would, with the caveat that the sensitive information they were provided must be transmitted over email. Additionally, participants were informed that they could browse the Internet, use their phones, or engage in other similar activities while waiting for email from their friend. This was done to provide a more natural setting for the participants, and to avoid frustration if participants had to wait for an extended period of time while their friend figured out an encrypted email system. Finally, participants were told that a study coordinator would be with them at all times and could answer questions related to the study but were not allowed to provide instructions on how to use any of the systems being tested.

4.2 Study Tasks

Using a coin flip, one participant was randomly assigned as Participant A and the other as Participant B (referred to as "Johnny" and "Jane", respectively, throughout the paper). The participants were then led to separate rooms to begin the study. The participants were then guided through the study by following a Qualtrics survey, which included both instructions and then questions regarding their experience.

After answering demographic questions, participants were asked to complete a multi-stage task three times, once for each of the secure email systems being tested. The order in which the participants used the systems was randomized. To complete this task, participants were asked to role-play a scenario about completing taxes. Johnny was told that his friend, Jane, had graduated in accounting and was going to help Johnny prepare his taxes. To do so, Johnny needed to send her his social security number and his last year's tax PIN. Johnny was told that because this information was sensitive, he should encrypt it using a secure email system he could download at a URL we gave him. Jane was told that she would receive some information regarding taxes from Johnny but was not informed that the information would be encrypted.

The tasks they were asked to perform were:

1. Johnny would encrypt and send his SSN and last year's tax PIN to Jane.
2. Jane would decrypt this information, then reply to Johnny with a confirmation code and this year's tax PIN. The reply was required to be encrypted.
3. After Johnny received this information, he would inform Jane that he had received the necessary information, and then the task would end. This confirmation step is added to ensure that Johnny could decrypt Jane's message. We did not require the confirmation message to be encrypted.

During each stage, participants were provided with worksheets containing instructions regarding the task and space for participants to record the sensitive information they received. These instructions did not include directions on how to use any of the systems. Both participants were provided with the information they would send (e.g., SSN and PIN), but were told to treat this information as they would their own sensitive information. Participants completed the same tasks for each of the three systems being tested.

Immediately upon completing the tasks for a given secure email system, participants were asked several questions related to their experience with that system. First, participants completed the ten questions from the System Usability Scale (SUS) [6, 7]. Multiple studies have shown that SUS is a good indicator of perceived usability [34], is consistent across populations [28], and has been used in the past to rate secure email systems [1, 23, 24, 27]. Next, participants were asked to describe what they liked about each system, what they would change, and why they would change it.

After completing the tasks and questions for all three secure email systems, participants were asked to select which of the email systems they had used was their favorite, and to describe why they liked this system. Participants were next asked to rate the following statements using a five-point Likert scale (Strongly Disagree–Strongly Agree): "I want to be able to encrypt my email," and "I would encrypt email frequently."

Finally, the survey told participants that MessageGuard could be enhanced with a master password, which they would be required to enter before MessageGuard would function. This would help protect their sensitive messages from other individuals who might also use the same computer. After reading the description about adding a master password to MessageGuard, users were asked to describe whether they would want this feature and why they felt that way.

4.3 Post-Study Interview

After completing the survey, participants were interviewed by their respective study coordinators. The coordinators asked participants about their general impressions of the study and the secure email systems they had used. Furthermore, the coordinators were instructed to note when the participants struggled or had other interesting events occur, and

[4]Using their own accounts increases ecological validity, but has privacy implications. To help mitigate these concerns we have destroyed the screen recordings for this study. Though not used, we did prepare study accounts for any participants who were not comfortable using their own account.

during the post-study interview the coordinators reviewed and further explored these events with the participants.

To assess whether participants understood the security provided by each secure email system, coordinators questioned participants regarding what an attacker would need to do to read their encrypted messages. Coordinators would continue probing participants' answers until they were confident whether or not the user correctly understood the security model of each system.

After describing their perceived security models, participants were then read short descriptions detailing the actual security models of each system. Participants were encouraged to ask questions if they wanted further clarification for any of the described models. After hearing these descriptions, participants were then asked to indicate whether their opinions regarding any of the systems had changed. Participants were also asked whether they would change their answer regarding their favorite system on the survey.

Upon completion of the post-study interview, participants were brought together for a final post-study interview. First, participants were asked to share their opinions on doing a study with a friend, as opposed to a traditional study. Second, participants were asked to describe their ideal secure email system. While participants are not system designers, we hoped that this question might elicit responses that participants had not yet felt comfortable sharing.

4.4 Quality Control

We excluded responses from eight pairs of participants.[5] First, three pairs were removed because the secure email tools became inoperative during the study, making it impossible for participants to complete the study.[6] Second, two pairs were removed because the participants did not speak or read English well enough to understand the study instructions and study coordinators. Third, we removed three participant pairs that were not paying attention to the study survey and filled in nonsense answers.

4.5 Demographics

We recruited Gmail users for our study at a local university, as well as through Craigslist. We distributed posters across campus to avoid biasing our participants toward any particular major. Participants were evenly split between male and female: male (47; 50%), female (47; 50%). Participants skewed young: 18 to 24 years old (75; 80%), 25 to 34 years old (18; 19%), 35 to 44 years old (1; 1%). Most participants were college students: high school graduates (1; 1%), undergraduate students (71; 76%), college graduates (15; 16%), graduate students (7; 7%). Participants were enrolled in a variety of technical and non-technical majors.

4.6 Limitations

Our study involved each user sending email to one other user. This approach was helpful in understanding the basic usability of the systems tested, but it might not reveal all the usability issues that would occur in other communication models, such as a user sending email to multiple individuals. Future work could examine other usage scenarios.

Our study also has several common limitations. First, our population is not representative of all groups, and future research could broaden the population (e.g., non-students, non-Gmail users). While we did use Craigslist to try and gather a more diverse population, these efforts were largely unsuccessful. Second, our study was a short-term study, and future research should look at these issues in a longer-term longitudinal study. Third, since our study was run in a trusted lab environment, participants may not have behaved the same as they would in the real world [20, 33].

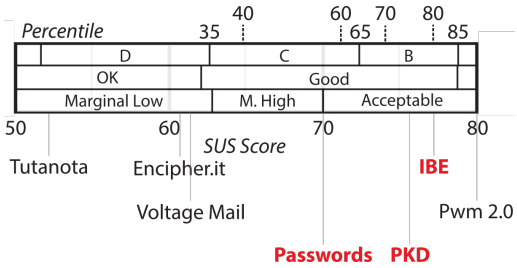
[5]When we excluded a participant's results, we also excluded their partner's results.

[6]These errors were not related to the usability of the system. For example, in one case, the Chrome Webstore went down, making it impossible for users to download the necessary extensions.

	Count	Mean	Standard Deviation	Confidence Interval ($\alpha = 0.05$)	Range	Percentile
Passwords	94	70.0	15.0	±3.0	67.0–73.0	56%
PKD	94	75.7	14.9	±3.0	72.7–78.7	76%
IBE	94	77.3	13.5	±2.7	74.6–80.0	81%

Percentiles are calculated by looking up the SUS score in a table [30]. When a SUS score is not in the table we estimate the percentile based on the available data.

Table 2: SUS Scores



The red items are systems evaluated in our study. The black items are systems evaluated in previous work that share key management schemes with the systems we tested: Encipher.it uses passwords, Tutanoa uses a public key directory, Pwm 2.0 and Voltage Mail use IBE.

Figure 4: Adjective-based Interpretation of SUS Scores

5. RESULTS

This section contains the quantitative results from our study: the SUS score for each system, task completion times, mistakes made by participants, participant understanding of each system's security model, rankings for the favorite system, and several other minor results. For brevity, we refer to the three variants tested as Passwords, PKD (public key directory), and IBE (identity-based encryption). The data for this study can be downloaded at `https://isrl.byu.edu/data/soups2018/`.

In several situations, we performed multiple statistical comparisons on the same data. In these cases, we use the Bonferroni correction to adjust our α value appropriately. Where a correction is not needed, we used the standard value $\alpha = 0.05$.

5.1 System Usability Scale

The System Usability Scale (SUS) score for each system is listed in Table 2. To give context to these scores, we leverage the work of several researchers that correlated SUS scores with more intuitive descriptions of usability [3, 4, 30, 34]. The descriptions are presented in Figure 4.

Passwords' score of 70.0 is rated as having "Good" usability, receives a "C" grade, and reaches the 56$^{\text{th}}$ percentile. PKD's SUS score of 75.7 is rated as having "Good" usability, receives a "B" grade, and falls in the 76$^{\text{th}}$ percentile of systems tested with SUS. IBE's score of 77.3 is also rated as having "Good" usability, receives a "B+" grade, and is in the 81$^{\text{st}}$ percentile.

A one-way repeated measures ANOVA comparing the effect of system on SUS scores revealed a statistically significant omnibus ($F(2, 186) = 13.43$, $p < .001$). The difference between Passwords' and PKD's scores are statistically significant (Tukey's HSD test—$p < 0.01$) as is the difference between Passwords' and IBE's SUS scores (Tukey's HSD test—$p < 0.01$). In both cases, the differences in means represent a significant improvement (20 and 25 percentile difference, respectively). In contrast, the difference between PKD's and IBE's SUS scores are not statistically significant. We also tested to see whether there was a difference between the SUS score ratings of Johnny and Jane, but the difference was not statistically significant (two-tailed student t-test, matched pairs—$p = 0.29$, $\alpha = 0.0125$).

Next, we compared the SUS scores for our variants against SUS scores of publicly available systems that used the same key management schemes. In each case our secure email variants outperformed these publicly available systems. We compared Encipher.it [27] against our Password variant, which scored 8.75 points higher (~25 percentile difference), Tutanota [23] against our our PKD variant, which scored 23.5 points higher (~60 percentile difference), and Voltage Mail against our IBE variant [27], which scored 14.64 points higher (~45 percentile difference).

Finally, we explored whether the order in which systems were tested had an effect on their SUS scores, finding three orderings with a non-negligible effect size: (1) Passwords scored 9.5 points higher when tested immediately after PKD, (2) PKD scores 9.5 points lower when it is tested after Passwords, (3) IBE scores 14.1 points lower when the system ordering is Passwords->IBE->PKD. All three of these differences are statistically significant (two-tailed student t-test, equal variance—$p < 0.001$, $p = 0.002$, $p < 0.001$, respectively, $\alpha = 0.0125$).

5.2 Time

We recorded the time it took each participant to finish the assigned task with each system. For timing purposes the tasks were split into two stages. The first stage started when Johnny visited the MessageGuard website and ended when he had successfully sent an encrypted email with his SSN and last year's tax PIN. The second stage started when Jane received her first encrypted email and ended when she had decrypted it, replied with the appropriate information, and received the confirmation email from Johnny. It is possible for stage one and two to overlap; if Johnny first sends an encrypted message without the required information, this will start the timer for stage two without stopping the timer for stage one. We took this approach because stage one is clearly not finished, but Jane is also able to start making progress on completing stage two.

	Stage	Count	Mean	Standard Deviation	Confidence Interval ($\alpha = 0.05$)	Range
Passwords	1	46	3:31	1:25	±0:25	03:06–03:56
	2	44	6:54	3:34	±1:03	05:51–07:57
	1 + 2	43	10:22	4:00	±1:12	09:10–11:34
PKD	1	47	8:02	3:06	±0:53	07:09–08:55
	2	45	3:24	1:28	±0:26	02:58–03:50
	1 + 2	45	11:33	3:53	±1:08	10:25–12:41
IBE	1	46	3:30	1:30	±0:26	03:04–03:56
	2	44	5:58	2:36	±0:46	05:12–06:44
	1 + 2	43	9:30	3:50	±1:09	08:21–10:39

Table 3: Time Taken to Complete Task (min:sec)

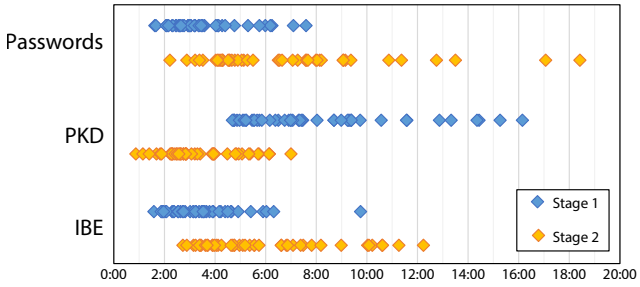


Figure 5: Individual Participant Task Completion Times

Timings were calculated using the video recordings of each participant's screen. We had missing or corrupted video in four cases. Task completion time data from the remaining recordings is given in Table 3 and Figure 5.

A two-way repeated measures ANOVA comparing the effect of system and stage on stage completion time fails to find a statistically significant overall difference between systems, but does reveal a statistically significant interaction effect (System—$F(2, 82) = 2.60$, $p = .08$, Stage—$F(1, 41) = 1.936$, $p < .0.17$, Interaction—$F(2, 82) = 82.52$, $p < .001$). By design, PKD shifts a significant portion of user effort from Stage 2 to Stage 1—Jane installs PKD in Stage 1 instead of Stage 2—resulting in a statistically significant difference in stage completion times (Tukey's HSD test—in all cases $p < 0.001$) with a large effect size (Stage 1—+4:30, Stage 2—−3:00). The difference between Passwords and IBE was not statistically significant for either Stage 1 or Stage 2.

We also explored whether system ordering had an effect on task completion times. As shown in Table 4, if a system was the first system tested, its task took considerably longer to complete than if it was not the first system tested. This difference is statistically significant for all three systems (two-tailed student t-test, equal variance—in all cases $p < 0.001$, $\alpha = 0.016$).

5.3 Mistakes

We define mistakes to be instances when users send sensitive information in normal email when it should have been

	Stage	Count	Mean When First	Mean When Not First	Effect Size
Passwords	1	46	4:50	3:01	−1:49 (−38%)
	2	44	9:49	5:48	−4:01 (−41%)
	Both	43	14:48	8:50	−5:58 (−40%)
PKD	1	47	9:36	7:13	−2:23 (−25%)
	2	45	4:20	2:54	−1:26 (−33%)
	Both	45	13:56	10:14	−3:42 (−27%)
IBE	1	46	4:47	2:49	−1:58 (−41%)
	2	44	8:01	4:48	−3:13 (−40%)
	Both	43	12:55	7:39	−5:16 (−41%)

Table 4: Time Taken to Complete Task as a Function of Whether it Was Tested First (min:sec)

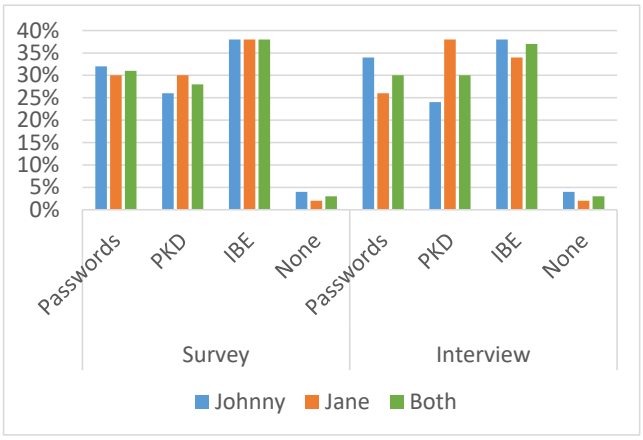


Figure 6: Participants' Favorite System

encrypted. For Passwords, a user is also considered to have made a mistake if they send the encryption password in a plaintext email.[7]

In Passwords, all mistakes were a result of users sending their password in plaintext email (Johnny–[9; 19%], Jane–[1; 2%]). For five of these mistakes (5; 11%), Johnny first sent the password over cellular text messaging, but for various reasons Jane never got this message. When Jane received her encrypted email, she didn't yet have the password and would email Johnny requesting the password, which he sent to her using email. Additionally, in four cases Johnny used Google Chat to send their password, giving Google access to both the secure email and the password used to encrypt it. Still, we chose not to include this as a mistake as it is not as egregious as sending the password over email.

In PKD and IBE there were a low number of mistakes, and each was made by Johnny (PKD–[n = 1; 2%], IBE–[2; 4%]). In all three cases, the participant transmitted the sensitive information in the unencrypted greeting[8] of the encrypted message. This happened in spite of the fact that two of these participants watched the compose tutorial, which warned them that text in that field would not be encrypted.[9]

5.4 Understanding

In the post-study interview we asked participants to identify what an attacker would need to do to read their encrypted email. The goal of this question was to evaluate whether participants understood the security model of each system they had tested. Study coordinators asked follow-up questions until they were confident that they could judge whether the participant had a correct understanding.

[7]Mistakes could conceivably also include revealing PKD or IBE private keys, but neither of our systems allowed users to make this mistake.

[8]The MessageGuard front end provides an unencrypted greeting field, which senders can populate with text readable by recipients who have not installed MessageGuard, aiding in the onboarding process.

[9]This problem could potentially be addressed by making users explicitly enable unencrypted greetings, instead of displaying it as a default field.

In five cases (Johnny–2, Jane–3), the study session ran late and participants had to leave without completing the post-study interview. As such, percentages in this Subsection are calculated off a different total number of participants (Johnny–45, Jane–44, Both–89).

Few participants had a correct understanding of PKD's (Johnny–[2; 4%], Jane–[2; 5%], Both–[4; 4%]) and IBE's (Johnny–[2; 4%], Jane–[3; 7%], Both–[5; 6%]) security models. Generally, participants believed that if an attacker could gain access to a user's email then they could decrypt that user's messages. Only a handful of participants recognized that signing up for an account was meaningful. During the interviews, most participants indicated they saw no difference in the security of IBE and PKD.

In strong contrast, nearly all participants had a clear understanding of how password-based encryption protected their emails (Johnny–[41; 91%], Jane–[41; 93%], Both–[82; 92%]).

5.5 Favorite System

At the end of the study survey, participants were asked to indicate their favorite system, and why. Later, during the post-study interview, participants were given descriptions of each system's security model and were invited to ask further clarifying questions as needed. After hearing these descriptions, participants were allowed to update which system they felt was their favorite. Participants' preferences, both pre- and post-survey, are summarized in Figure 6.

Overall, participants were split on which system they preferred (During Survey—PKD–[26; 28%], IBE–[36; 38%], Passwords–[29; 31%]; After Interview—PKD–[29; 31%], IBE–[34; 36%], Passwords–[28; 30%]). While IBE was a slight favorite, the difference was not statistically significant (Chi-squared test—Survey–$\chi^2[2, N = 282] = 2.56, p = 0.28$, Interview–$\chi^2[2, N = 282] = 1.01, p = 0.60$). Of the three participants who did not select a favorite system (3; 3%), two indicated that they liked all three systems equally, and the third participant indicated that he disliked all three systems because he erroneously believed that the systems caused his encrypted email to not be stored by Gmail.

Approximately a sixth of participants (15; 16%) changed their favorite system after better understanding the security

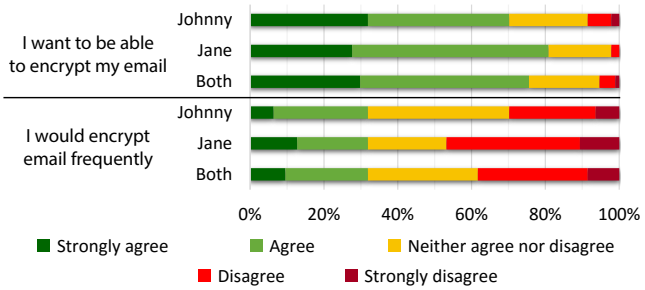


Figure 7: Participant Opinions Regarding Secure Email

models of each system: one from Passwords to PKD, two from passwords to IBE, four from PKD to IBE, six from IBE to PKD, and two from IBE to Passwords. In total, Passwords lost one vote, PKD gained three votes, and IBE lost two votes.

5.6 Other Results

We also recorded how often participants used various features in MessageGuard. We noted that Johnny frequently watched both the compose and read tutorials (Compose–[41; 87%], Read–[38; 81%]). Jane similarly watched the read tutorial (43; 91%), with a slightly lower rate of watching the compose tutorial (6 out of 10 participants; 60%).[10] We found that Johnny was likely to include a plaintext greeting with his encrypted email (33; 70%). When Jane did send a new encrypted message, she included an unencrypted greeting a little under half of the time (4 of 10 participants; 40%).[11]

We noted that Johnny used a variety of methods to transmit the password used to encrypt his email, overall preferring phone-based communication channels (cellular text messaging–23, phone call–11, email–9, Google Chat–4, in person–2, Facebook Chat–1).[12] In three cases (phone call–2, email–1) Johnny did not transmit the password, but merely gave clues to Jane that were sufficient for her to figure it out.

At the end of the survey, participants were asked whether they wanted to be able to encrypt their email and whether they would frequently do so. Participant responses to these questions are summarized in Figure 7. Overall, participants were in strong agreement that email encryption is something they want (want–[71; 76%], unsure–[18; 19%], don't want–[5; 5%]). Still, participants were split on how often they would use secure email, with the plurality going to infrequent use (frequent use–[30; 32%], unsure–[28; 30%], infrequent use–[36; 39%]). This is in line with previous results regarding desired secure email usage [24].

6. QUALITATIVE RESULTS

In this section we discuss participants' qualitative feedback and observations from the study coordinators. We refer to participants using a unique identifier R[1–47][A,B], where A refers to the Johnny role and B refers to the Jane role.

[10] Jane only saw the compose tutorial if she started a new email chain.

[11] Encrypted replies do not contain plaintext greetings.

[12] These usage numbers do not sum to 47 as Johnny sometimes used multiple methods to communicate the password.

6.1 Passwords

Participants gave Passwords a lower SUS score than both PKD and IBE, but overall indicated it was quite usable. Even though users rated Passwords as usable, a substantial number indicated they preferred PKD and IBE due to these systems not requiring a password to encrypt email.

Communicating the password to the recipient was the main problem with password-based encryption. As already discussed, many participants shared their password over plaintext email. In some cases, they recognized this didn't seem secure, but still proceeded. Some participants questioned the security of using out-of-band channels to send the password.

> *"We also communicated the password through a text message. I'm not sure what that does for the security of the system if we are using an outside and unprotected means of communication in order to make it work."* [R24B]

Many participants also felt that communicating a password out-of-band negated the need to use secure email, as they could just communicate the sensitive information over the out-of-band channel. R39B indicated,

> *"It was way lame that I had to call him because I might as well have just given him the info that way. . . . If I'm gonna communicate with them through email, it's because I want to do it through email, not through a phone call."*

Several participants noted it would be annoying to manage separate passwords while communicating securely with multiple people. In this regard, R9A expressed,

> *"I may want to use [Passwords] often in sending regular messages to many people. If I had to share a password each time, it may make the process cumbersome."*

Participants had several suggestions to improve Passwords. First, participants proposed allowing only a single password to protect an email thread. Users could reuse passwords to encrypt replies, but many participants became confused and created new passwords, necessitating more password exchanges. Second, some participants felt that it would be helpful to have a built-in password complexity meter or random password generator when creating passwords.

> *"If you don't have a random password generator, then people will just end up using familiar passwords, which is actually more of a problem than if there were no passwords at all."* [R18B]

Unlike PKD and IBE, the security model for the Passwords system was well-understood by participants. Understanding the security model of passwords helped users trust the system's security.

> *"It was nice to be able to create a password that only myself and the sender know. It felt more secure. . . . "* [R3A]

6.2 PKD

In general, participants described the PKD system as fast and easy-to-use. The most common complaint about PKD

was that recipients needed to install PKD before they could be sent encrypted messages. As stated by R1A, *"It's not great that sending someone an encrypted email means you have to ask them to download an extension."* Additionally, some participants felt they were less likely to install the system if they didn't already have an encrypted message.

> *"I am more motivated (i.e., I can more readily see the need) to install the app if the encrypted message is already sitting there in my inbox. Also, the fewer emails I have to send/receive the better."* [R9B]

The most significant issue we discovered with our PKD system was that very few participants understood its security model (4; 4%), with most participants assuming an attacker only needed access to the user's email account to read their encrypted email. After explaining PKD's security model to participants, they felt much more confident in its security. Particularly, participants liked that it did not rely on any third parties. For example, after hearing about PKD's security model R47B enthusiastically changed her favorite system from Passwords to PKD and stated,

> *"Just because it had to be from your computer, it seems like, if they were to get the [encrypted contents], it'd be a little bit harder for them to get [the plaintext contents]."*

Participants' interest in PKD was tempered by the risk of losing all their encrypted email if something were to happen to the private key stored on their computer.

> *"I guess, depending on what you're doing, [PKD] could be helpful, but it could also be very frustrating . . . if you changed systems or something like that, it could be frustrating to realize that you couldn't decrypt previously sent messages."* [R18A]

6.3 IBE

Similar to previous studies [23, 24, 27], participants found IBE to be extremely usable. Task completion times show that IBE was faster than the other two systems.

Prior implementations of IBE relied on automatic email authentication to deliver private keys [24, 27]. Our implementation has users create a username and password on the key server for authenticating a request to retrieve a private key.[13] This prevents the email provider from being able to access the user's private key. This added security can impact usability. While most users did not mind setting up an account, several participants disliked this aspect.

> *"As a general comment, I think the password one was my favorite, since you didn't have to create an account for MessageGuard."* [R3B]

As with PKD, participants had a poor understanding of IBE's security model. Nearly all participants thought PKD and IBE had poor security, incorrectly believing that anyone who broke into their Gmail account could read all encrypted emails. After receiving instructions on IBE's security model, some participants who initially preferred IBE switched their preference to PKD; most remained with IBE, stating it had adequate security. Additionally, these participants felt that the ability to send an IBE-encrypted message to a recipient without waiting for them to first install MessageGuard trumped the security drawbacks of IBE.

[13] Our PKD system also required users to create an account.

6.4 User Attitudes

We asked participants if they would be interested in MessageGuard including a master password. With a master password, MessageGuard would not encrypt or decrypt email until this password was entered. Moreover, cryptographic keys would be encrypted using the master password before being stored to disk.[14] Overall, participants were interested in this feature (Johnny–[33; 70%], Jane–[35; 74%], Both–[72; 77%]). Participants felt this would provide an important security property when multiple users shared a single computer. The participants not interested in a master password indicated they had sole access to their computer, and a master would add a hassle for no real security gain.

Participants also expressed a strong desire to better understand how the secure email systems worked. They felt this would help them verify the system was properly protecting their data. Additionally, several participants stated they would not feel comfortable using a "random" tool from the Internet. Instead, they looked for tools that were verified by security experts or were distributed and endorsed by a well-known brand (e.g., Google).

7. DISCUSSION

We discuss lessons learned, usability and security trade-offs, and validation of prior work.

7.1 Lessons Learned

It is unclear whether the mistake of sending the password via email represents users' lack of understanding regarding the security of email [22], a lack of concern for the safety of their sensitive information during the role play, an artifact of taking the study in a trusted environment [33], or a mixture of the three.

With so much of PKD's key management automated (e.g., key generation, uploading and retrieval of public keys), it is likely participants had insufficient contextual clues showing the system's security model. While reducing the automation of the system could improve understanding, these changes would likely come at an unacceptably high usability cost [23, 26, 32, 36]. Future work should examine ways the system could conform to users' existing mental models.

During the user study, several participant pairs encountered an edge case for IBE—Jane had multiple email address aliases, and the message was encrypted for a different alias than Jane used when she set up her MessageGuard account. This resulted in Jane being unable to decrypt Johnny's message. This was especially confusing for Johnny and Jane because they had no indication of what they needed to do to resolve the issue. MessageGuard's design anonymizes the identity of the recipients, so the system could not inform Jane which email alias she needed to register with her MessageGuard

[14] The master password differs from the MessageGuard account password in that the former is used only locally to protect access to cryptographic keys stored on the local device, whereas the latter is used to protect against an adversary uploading (PKD) or downloading (IBE) cryptographic keys to/from the MessageGuard key server. Users could choose to use the same password for both use cases.

account in order to read the message. The difficulty of handling email aliases is not limited to IBE. It affects PKD as well. It is unclear how best to solve this problem, and this is an area for future work.

7.2 Usability and Security Trade-offs

Hiding cryptographic details increases usability, but inhibits understanding of a system's security model.[15] For example, both IBE and PKD hid key management from the user, leading to high usability scores. However, post-study interviews revealed participants did not understand the security model of either system. In contrast, the Passwords system required users to manually manage their keys (using passwords). This led to lower usability scores for Passwords, but nearly all users understood its security model.

Tools relying on third-party key servers sacrifice security but significantly reduce the burden of adopting the system. For example, evaluations of PKD systems using manual key exchange have consistently found these systems to be unusable [26, 32, 36]. Our PKD system significantly improved its usability at the expense of trusting a third-party by employing a public key directory. Similarly, IBE fully trusts its third-party server with private keys, making it trivial to send any recipient an encrypted message. Even though participants recognized the lower security of IBE, many indicated that it had "good enough" security for their needs.

7.3 Validation of Prior Research

Our results demonstrate that the design principles we identified in previous work [24, 27] generalize beyond IBE, and are also applicable to PKD and password-based systems. Many favorable participant responses demonstrated the importance of tight-integration; context-sensitive, inline tutorials; and unencrypted greetings (R7A, R9A, R26B, respectively):

> *"I really like the integration into Gmail, so that I can safely send information without having to use an entirely new system."*

> *"The tutorial was very helpful. I also found the icons to be helpful in using the tool. I was surprised at how easily the program integrated into my e-mail. There was never any confusion as to what I needed to do or as to what was going on."*

> *"I like . . . that the subject/top of the email are not encrypted to help others realize that this is not spam."*

We also gathered further evidence showing paired-participant usability studies [23] are helpful in assessing the usability of secure email systems. Both the quantitative and qualitative data revealed strong differences between Johnny and Jane, indicating that there is value in gathering information for both roles. When asked, participants indicated they enjoyed working with a friend and felt it was more natural than working with a study coordinator. This was especially true for our Passwords system, where they indicated calling their friend was natural, but not something they would feel comfortable doing with a coordinator.

[15] Understanding a system's security model is important as it allows users to understand what actions are safe and what put them at risk.

8. CONCLUSION

The paper compared the usability of three different key management approaches to secure email: passwords, public key directory, and IBE. The systems were built using state-of-the-art design principles for usable, secure email [1, 2, 24, 27] and were evaluated using standard metrics and a paired-participant study methodology [23]. This evaluation was the first A/B evaluation of key management schemes in which participants were allowed to self-discover how the system worked. It is also the largest secure email study to date (94 participants), which is twice as large as previous studies [23].

Our research demonstrates that each key management approach has the potential to be successfully used in secure email. Additionally, participants' qualitative feedback provides valuable insights into the usability trade-offs of each key management approach, as well as several general principles of usable, secure email. Finally, our work provides evidence that validates prior work on the design principles [24] used in our systems as well as the study methodology [23].

While our results are very positive, they are focused on helping users begin using secure email. Further research is needed regarding how secure email systems, including MessageGuard, perform when used on a day-to-day basis. Based on our experience, we make the following recommendations for this future research:

- The public key directory scheme requires that users store and backup their private keys securely and reliably. They also need to transfer them between devices. Future work should explore users' ability to do so, as this could be a potential usability impediment that would also greatly reduce security.
- Future work needs to examine how to design encrypted email systems that support key email functionality, including spam filtering and search.
- Given the promising results for the various key management schemes in a laboratory setting, the next step is to design and conduct longitudinal studies to see if the results hold over an extended period in real-world scenarios.
- Participants in our study struggled to understand the threat model of the public key directory and IBE schemes. This is problematic inasmuch as users overestimate the security of the system and send sensitive data they would not if they properly understood the system's threat model. Future work should examine how tutorials can be constructed to address this issue. Particular care should be taken to validate that tutorials will not be ignored by users when completing secure email tasks.
- Future email studies should compare features of interest using A/B tests, standard metrics, and a two-person methodology to increase the confidence in results from these studies and also help situate new results clearly within the existing body of work.

Acknowledgment

We thank the anonymous reviewers and our shepherd, Marian Harbach, for their suggestions that helped improve the paper. This work was supported in part by the National Science Foundation under Grant No. CNS-1528022.

References

[1] E. Atwater, C. Bocovich, U. Hengartner, E. Lank, and I. Goldberg. Leading Johnny to water: designing for usability and trust. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 69–88, Montreal, Canada. USENIX Association, 2015.

[2] W. Bai, M. Namara, Y. Qian, P. G. Kelley, M. L. Mazurek, and D. Kim. An inconvenient trust: user attitudes toward security and usability tradeoffs for key-directory encryption systems. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 113–130, Denver, CO. USENIX Association, 2016.

[3] A. Bangor, P. Kortum, and J. Miller. An empirical evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction*, 24(6):574–594, 2008.

[4] A. Bangor, P. Kortum, and J. Miller. Determining what individual SUS scores mean: adding an adjective rating scale. *Journal of Usability Studies*, 4(3):114–123, 2009.

[5] C. Bravo-Lillo, L. Cranor, J. Downs, S. Komanduri, S. Schechter, and M. Sleeper. Operating system framed in case of mistaken identity: measuring the success of web-based spoofing attacks on OS password-entry dialogs. In *Nineteenth ACM SIGSAC Conference on Computer and Communications Security (CCS 2012)*, pages 365–377, Raleigh, NC. ACM, 2012.

[6] J. Brooke. SUS—a quick and dirty usability scale. In *Usability Evaluation in Industry*. CRC Press, Boca Raton, FL, 1996.

[7] J. Brooke. SUS: a retrospective. *Journal of Usability Studies*, 8(2):29–40, 2013.

[8] R. Chandramouli, S. L. Garfinkel, S. J. Nightingale, and S. W. Rose. Trustworthy email. *Special Publication (NIST SP) 800-177*, 2016.

[9] J. Clark, P. C. van Oorschot, S. Ruoti, K. Seamons, and D. Zappala. Securing Email. *ArXiv e-prints*, Apr. 2018. arXiv: `1804.07706 [cs.CR]`.

[10] R. Dhamija and J. D. Tygar. The battle against phishing: Dynamic Security Skins. In *First Symposium on Usable Privacy and Security (SOUPS 2005)*, pages 77–88, Pittsburgh, PA. ACM, 2005.

[11] S. Fahl, M. Harbach, T. Muders, and M. Smith. Confidentiality as a service–usable security for the cloud. In *Eleventh International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom 2012)*, pages 153–162, Liverpool, England. IEEE Computer Society, 2012.

[12] S. Garfinkel. *PGP: Pretty Good Privacy*. O'Reilly Media, Inc., Sebastopol, CA, 1995.

[13] S. L. Garfinkel and R. C. Miller. Johnny 2: a user test of key continuity management with S/MIME and Outlook Express. In *First Symposium on Usable Privacy and Security (SOUPS 2005)*, pages 13–24, Pittsburgh, PA. ACM, 2005.

[14] W. He, D. Akhawe, S. Jain, E. Shi, and D. Song. ShadowCrypt: encrypted web applications for everyone. In *Twenty-First ACM SIGSAC Conference on Computer and Communications Security (CCS 2014)*, pages 1028–1039, Scottsdale, AZ. ACM, 2014.

[15] C. Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Seventeenth New Security Paradigms Workshop (NSPW 2009)*, pages 133–144, Oxford, England. ACM, 2009.

[16] T. W. v. d. Horst and K. E. Seamons. Encrypted email based upon trusted overlays, 2009. US Patent 8,521,821.

[17] B. Lau, S. Chung, C. Song, Y. Jang, W. Lee, and A. Boldyreva. Mimesis aegis: a mimicry privacy shield—a system's approach to data privacy on public cloud. In *Twenty-Third USENIX Security Symposium (USENIX Security 2014)*, pages 33–48, San Diego, CA. USENIX Association, 2014.

[18] A. Lerner, E. Zeng, and F. Roesner. Confidante: usable encrypted email: a case study with lawyers and journalists. In *Security and Privacy (EuroS&P), 2017 IEEE European Symposium on*, pages 385–400. IEEE, 2017.

[19] M. S. Melara, A. Blankstein, J. Bonneau, M. J. Freedman, and E. W. Felten. CONIKS: a privacy-preserving consistent key service for secure end-to-end communication. In *Twenty-Fourth USENIX Security Symposium (USENIX Security 2015)*, pages 383–398, Washington, D.C. USENIX Association, 2015.

[20] S. Milgram and E. V. d. Haag. *Obedience to Authority*. Ziff–Davis Publishing Company, New York, NY, 1978.

[21] H. Orman. *Encrypted Email: The History and Technology of Message Privacy*. Springer, 2015.

[22] K. Renaud, M. Volkamer, and A. Renkema-Padmos. Why doesn't Jane protect her privacy? In *Fourteenth Privacy Enhancing Technologies Symposium (PETS 2014)*, pages 244–262, Philadelphia, PA. Springer, 2014.

[23] S. Ruoti, J. Andersen, S. Heidbrink, M. O'Neill, E. Vaziripour, J. Wu, D. Zappala, and K. Seamons. "We're on the same page": a usability study of secure email using pairs of novice users. In *Thirty-Fourth ACM Conference on Human Factors and Computing Systems (CHI 2016)*, pages 4298–4308, San Jose, CA. ACM, 2016.

[24] S. Ruoti, J. Andersen, T. Hendershot, D. Zappala, and K. Seamons. Private Webmail 2.0: simple and easy-to-use secure email. In *Twenty-Ninth ACM User Interface Software and Technology Symposium (UIST 2016)*, Tokyo, Japan. ACM, 2016.

[25] S. Ruoti, J. Andersen, T. Monson, D. Zappala, and K. Seamons. MessageGuard: a browser-based platform for usable, content-based encryption research, 2016. arXiv preprint arXiv:1510.08943.

[26] S. Ruoti, J. Andersen, D. Zappala, and K. Seamons. Why Johnny still, still can't encrypt: evaluating the usability of a modern PGP client, 2015. arXiv preprint arXiv:1510.08555.

[27] S. Ruoti, N. Kim, B. Burgon, T. Van Der Horst, and K. Seamons. Confused Johnny: when automatic encryption leads to confusion and mistakes. In *Ninth Symposium on Usable Privacy and Security (SOUPS 2013)*, Newcastle, United Kingdom. ACM, 2013.

[28] S. Ruoti, B. Roberts, and K. Seamons. Authentication melee: a usability analysis of seven web authentication systems. In *Twenty-fourth International Conference on World Wide Web (WWW 2015)*, pages 916–926, Florence, Italy. ACM, 2015.

[29] M. D. Ryan. Enhanced certificate transparency and end-to-end encrypted mail. In *Twenty-Second Network and Distributed System Security Symposium (NDSS 2014)*, San Diego, CA. The Internet Society, 2014.

[30] J. Sauro. *A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices*. Measuring Usability LLC, Denver, CO, 2011.

[31] A. Shamir. Identity-based cryptosystems and signature schemes. In *Fourteenth International Cryptology Conference (Crypto 1984)*, pages 47–53, Santa Barbara, CA. Springer, 1984.

[32] S. Sheng, L. Broderick, C. Koranda, and J. Hyland. Why Johnny still can't encrypt: evaluating the usability of email encryption software. In *Poster Session at the Symposium On Usable Privacy and Security*, Pittsburgh, PA, 2006.

[33] A. Sotirakopoulos, K. Hawkey, and K. Beznosov. "I did it because I trusted you": challenges with the study environment biasing participant behaviours. In *Usable Security Experiment Reports Workshop at the Symposium On Usable Privacy and Security*, Redmond, WA, 2010.

[34] T. S. Tullis and J. N. Stetson. A comparison of questionnaires for assessing website usability. In *Usability Professional Association Conference*, pages 1–12, Minneapolis, MN. Usability Professionals Association, 2004.

[35] N. Unger, S. Dechand, J. Bonneau, S. Fahl, H. Perl, I. Goldberg, and M. Smith. SoK: secure messaging. In *Thirty-Sixth IEEE Symposium on Security and Privacy (S&P 2015)*, pages 232–249, San Jose, CA. IEEE Computer Society, 2015.

[36] A. Whitten and J. Tygar. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In *Eighth USENIX Security Symposium (USENIX Security 1999)*, pages 14–28, Washington, D.C. USENIX Association, 1999.

APPENDIX

A. MESSAGEGUARD'S DESIGN GOALS

In this section, we give the threat model that motivates our work. Next, we describe how to implement security overlays in order to enhance existing web applications with content-based encryption. Finally, we discuss our goals for MessageGuard, that are necessary to support research of content-based encryption in a usable, secure, and extensible manner.

A.1 Threat Model

In content-based encryption, sensitive content is only accessible to the author of that data and the intended recipient. In contrast to transport-level encryption (e.g., TLS), which only protects data during transit, content-based encryption protects data both during transit and while it is at rest. In our threat model, we consider web applications, middleboxes (e.g. CDNs), and the content they serve to be within the control of the adversary. The adversary wins if she is able to use these resources to access the user's encrypted data. While it is true that most websites are not malicious, in order to support ubiquitous, content-based encryption, it is necessary to protect against cases where websites are actively trying to steal user content. Users' computers, operating systems, software, and content-based encryption software[16] are all considered part of the trusted computing base in our threat model.

Our threat model is concerned with ensuring the confidentiality and integrity of encrypted data, but does allow for the leakage of meta-data necessary for the encrypted data to be transmitted and/or stored by the underlying web application. For example, in order to transmit an encrypted email message, the webmail system must have access to the unencrypted email addresses of the message's recipient. Additionally, the webmail provider will be able to inspect the encrypted package and gain learn basic information about the encrypted package (e.g., approximate length of message, number of recipients).[17]

While our threat model is necessarily strict to support the wide range of web applications that researchers may wish to investigate, we note that research prototypes built using the MessageGuard platform are free to adopt a weaker threat model that may be more appropriate for that particular research.

A.2 Security Overlays

There are several approaches for implementing overlays: `iframes` [16, 27], the `ShadowDOM` [14], user script engines such as Greasemonkey [11], and the operating system's accessibility framework [17]. Based on our analysis of each of these approaches, `iframes` are the implementation strategy best suited to work across all operating systems and browsers (including mobile). Additionally, `iframe`-based security overlays have security and usability that are greater than or equal to that of other approaches. As such, we designed MessageGuard using security overlays based on `iframes`.

Relying on `iframes` largely restricts MessageGuard to supporting only web applications deployed in the browser. Still the browser is an ideal location for studying content-based encryption: (1) There are a large number of high-usage browser-based web applications (e.g., webmail, Google Docs). (2) Traditional desktop and mobile application development increasingly mimics web development, allowing lessons learned in browser-based research to also apply to these other platforms. (3) There is already a substantial amount of research into adding content-based encryption to web applications, both academic (e.g., [1, 11, 14, 27]) and professional (e.g., Virtru, Mailvelope, Encipher.it).

[16] This includes the software's website and any web services the software relies upon (e.g., a key server).

[17] This type of leakage also occurs in HTTPS.

A.3 Platform Goals

We examined the existing work on content-based encryption (e.g., [13, 32, 35, 36]) in order to establish a set of design goals for MessageGuard. These goals are centered around enabling a researcher to investigate usable, content-based encryption.

A.3.1 Secure

MessageGuard should secure users' sensitive content from web applications and network adversaries.

MessageGuard should protect data in its overlays from being accessed by the web application. Sensitive data that is being created or consumed using MessageGuard should be inaccessible to the underlying web application. A corollary to this rule is that no entities that observe the transmission of data encrypted by MessageGuard should be able to decipher that data unless they are the intended recipients.

MessageGuard's interfaces should be clearly distinguishable from the web application's interfaces. In addition to protecting content-based messages from websites, it is important that systems clearly delineate which interfaces belong to the website and which belong to the content-based encryption software. This helps users to feel assured that their data is being protected and assists them in avoiding mistakes [24, 27]. Additionally, visual indicators should be included that can help protect against an adversary that attempts to social engineer a user into believing they are entering text into a secure interface when in reality they are entering text directly into the adversary's interface [5, 10].

A.3.2 Usable

MessageGuard should provide a usable base for future research efforts.

MessageGuard should be approachable to novice users. Easy-to-use systems are more likely to be adopted by the public at large [35]. Furthermore, complicated systems foster user errors, decreasing system security [27, 36]. While some systems need to expose users to complex security choices, basic functionality (e.g., sending or receiving an encrypted email) should be approachable for new users. At a minimum this includes building intuitive interfaces, providing integrated, context-sensitive tutorials, and helping first-time recipients of encrypted messages understand what they need to do in order to decrypt their message.

MessageGuard should integrate with existing web applications. Users enjoy the web services and applications they are currently using and are disinclined to adopt a new system solely because it offers greater security. Instead, users prefer that content-based encryption be integrated into their existing applications [1, 27]. Equally important, content-based encryption should have a minimal effect on the application's user experience; if encryption gets in the way of users completing tasks it is more likely that they will turn off content-based encryption [15].

MessageGuard's interfaces should be usable at any size. Current web interfaces allowing users to consume or create content come in a wide variety of sizes (i.e., height and width). When MessageGuard integrates with these web services, it is important that MessageGuard's interfaces work at these same dimensions. To support the widest range of sizes, MessageGuard's interfaces should react to the space available, providing as much functionality as is possible at that display size.

A.3.3 Ubiquitous

MessageGuard should support most websites and platforms.

MessageGuard should work with most websites MessageGuard should make it easy for researchers to explore adding end-to-end encryption into whichever web applications they are interested in. While it may be impossible to fully support all web applications (e.g., Flash applications or applications drawn using an HTML canvas), most standard web applications should work out-of-the-box. For those applications which don't work out-of-the-box, MessageGuard should allow researchers to create customized prototypes that handle these edge cases.

MessageGuard should function in all major desktop and mobile browsers. Prototypes built with MessageGuard should function both on desktop and mobile browsers, allowing researchers to experiment with both of these form factors. Furthermore, MessageGuard should work on all major browsers, allowing users to work with the web browser they are most familiar with, obviating the need to restrict study recruitment to users of a specific browser.

A.3.4 Extensible

MessageGuard should be easily extensible and contribute to the rapid development of content-based encryption prototypes.

MessageGuard should be modular. MessageGuard's functionality should be split into a variety of modules, with each module taking care of a specific function. Researchers should also be free to only change the modules that relate to their research and have the system continue to function as expected. Similarly, MessageGuard's modules should be extensible, allowing researchers to create new custom modules with a minimal amount of effort.

MessageGuard should provide reference functionality. As a base for other researchers' work, MessageGuard should include a reference implementation of the various modules that adds content-based encryption to a wide range of web applications. This reference implementation should be able to be easily modified and extended to allow researchers to rapidly implement their own ideas.

A.3.5 Reliable

The usability and security of MessageGuard should be reliable, protecting researchers from unintentionally compromising MessageGuard's security or usability.

Reducing the security of MessageGuard should require deliberate intent. HCI researchers should feel comfortable customizing MessageGuard's interface without needing to worry that they are compromising security. To facilitate this, MessageGuard should separate UI and security functionality into separate components. As long as researchers limit themselves to changing only UI components, there should be no effect on security.

Modifying the cryptographic primitives should have minimal effect on MessageGuard's usability. As above, MessageGuard should separate its UI and security functionality into separate

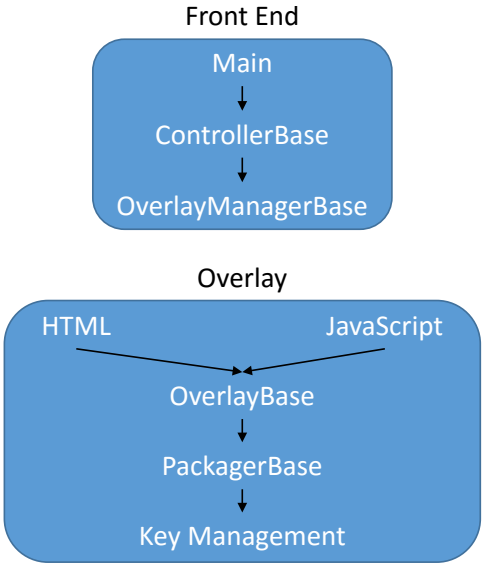


Figure 8: MessageGuard's customizable framework.

components. This will allow security researchers to modify the cryptographic primitives without worrying about how they will affect MessageGuard's usability. One caveat is if a new key management scheme requires a user interface that MessageGuard does not already make available. In this case, researchers will need to provide this key management scheme's interface, which could affect usability, but other interfaces should remain unaffected.

B. MESSAGEGUARD AS A RESEARCH PLATFORM

In this section, we describe the ways researchers can employ MessageGuard as a platform for their own research. In addition to the details described in this section, we invite researchers to download MessageGuard's source code. To help researchers quickly familiarize themselves with MessageGuard's code base, we have included instructive comments throughout the code and have provided a reference implementation that supports most websites that researchers can refer to as they build their own systems.

MessageGuard was designed to minimize the amount of code that must be changed in order for researchers to build new prototypes. The customizable classes enabling this rapid prototyping are shown in Figure 8. MessageGuard includes a default instantiation for each of the base classes (e.g. `ControllerBase`) seen in the figure. To change the global functionality of MessageGuard, researchers need to change the aforementioned default implementations. If researchers desire to implement new functionality (e.g., create a new overlay, support a new application), they can instead subclass these base classes. All classes, both base classes and default implementations, can be extended, but only allow researchers to override the methods that are unique to their functionality.

B.1 Frontend

The main class is responsible for parsing the URL and instantiating the appropriate controller (i.e., classes extending `ControllerBase`). Frontend controllers are responsible for the actual operations of the frontend, including detecting when overlays are needed and placing those overlays. Every overlay is created by and coupled to an overlay manager, which is responsible for handling communication between the overlay and MessageGuard's frontend. Currently, MessageGuard provides overlay managers for both reading and composing encrypted content.

The simplest way to modify the frontend is to change the elements that it will overlay. This can be done by changing the CSS selector that is passed to `ControllerBase`'s constructor.[18] The controller can also be configured to support additional types of overlays (i.e., creating a unified read and compose overlay for instant messaging clients). In this case, it will also be necessary to create an overlay manager to communicate with the new overlay.

Using these base classes, MessageGuard's default functionality was implemented using less than 200 lines of JavaScript.

B.2 Overlays

Overlays are composed of both HTML interfaces and JavaScript code. Researchers can either modify the existing overlays (read and compose) or create their own overlays. The steps for creating a new overlay modifying overlays on a per-application basis are as follows:

1. Create a new HTML file for each overlay. This will define the visual appearance of the overlay.
2. Create a custom read, compose, or entirely new overlay (e.g., file upload) by extending either the `OverlayBase` class or one the reference overlays (read and compose). These parent classes provide basic functionality (e.g., positioning, communication with the frontend).
3. Connect the overlay's HTML interface to its controlling code by referencing this new JavaScript class in the new HTML.
4. Create a new overlay manager to work with the new overlay. You can extend any of the existing overlay managers, or create a new one by extending `OverlayManagerBase`.
5. Add any custom communication code to both the new overlay and overlay manager.

MessageGuard's default read overlay required 70 lines of HTML and 150 lines of JavaScript to implement. The default compose overlay needed 190 lines of HTML and 670 lines of JavaScript, most of which was responsible for setting up the HTML5 rich-text interface and allowing users to select a specific key for encryption.

B.3 Packager

By overriding PackagerBase, it is possible to create custom message packages, allowing MessageGuard to support a wide range of content-based encryption protocols. This functionality can be used to allow prototypes developed with MessageGuard to inter-operate with existing cryptographic systems (e.g., using the PGP package syntax in order to be compatible with existing PGP clients). It could also be used

[18] Though unlikely to be necessary, it is also possible to modify the controller to do more complex selection that does not rely on CSS selection.

to experiment with advanced cryptographic features, such as key ratcheting [35].

B.4 Key Management

One key goal of MessageGuard is to allow existing proposals for key management to be implemented in a real system, and then compared against alternative schemes. As such, we took special care to ensure that MessageGuard would be compatible with all key management schemes we are currently aware of. In order to create a new key management scheme, the following two classes must be implemented:

KeyScheme. The KeyScheme is responsible for handling scheme-specific UI functionality for the key manager (e.g., importing public/private keys, authenticating to a key server). The KeyScheme methods are:

- **getUI** Retrieves a scheme-specific UI that will be included with the KeyUIManager's generic UI. This method is provided with the KeySystem being created/updated and a callback which notifies the KeyUIManager that the KeySystem is ready to be saved.
- **handleError** Modifies an existing KeySystem's UI to allow it to address an error. This method is provided with details about the error, the KeySystem UI to modify, and a callback which notifies the KeyUIManager that the error has been resolved. Examples of errors include not having a necessary key or expired authentication credentials.
- **create** Creates a KeySystem from the scheme-specific UI provided to this method.
- **update** Updates a KeySystem from the scheme-specific UI provided to this method.

KeySystem. A KeySystem is an instantiation of a key management scheme that allows the users to decrypt/sign data for a single identity and encrypt/verify data for any number of identities.[19] A KeySystem is responsible for performing cryptographic operations with the keys it manages. Every KeySystem has a fingerprint that uniquely identifies it. The KeySystem methods are:

- **serialize/deserialize** Prepares data that is not a part of the KeyAttributes type for storage by the KeyStorage class.
- **encrypt** Encrypts data for the provided identity. Returns the encrypted data along with the fingerprint of the KeySystem that can decrypt it.
- **decrypt** Decrypts the provided data.
- **sign** Signs the provided data.
- **verify** Verifies that the provided signature is valid for the provided data.

By default, MessageGuard will allow users to use all available key management schemes, though this can be overridden on a per-prototype basis.

[19] Key systems which don't support recipients set `canHaveRecipients` to `false` and ignore the identity parameters.

Stage	Static			Dynamic		
n	100	500	1000	100	500	1000
Chrome[1]	1.14	0.84	0.95	3.17	6.49	11.0
Firefox[1]	1.06	0.99	0.96	2.26	3.15	4.45
Safari[1]	0.45	0.63	0.53	3.73	12.8	25.5
Chrome[2]	4.27	4.39	4.60	12.9	30.2	51.1
Chrome[3]	5.68	5.97	5.94	12.4	32.0	61.2
Safari[3]	2.57	2.46	1.79	15.1	25.2	39.5

[1] MacBook Air (OSX 10.10.3, 1.7GHz Core i7, 8GB RAM). Chrome—42.0.2311.135, Firefox—37.0.2, Safari—8.0.5.
[2] OnePlus One (CyanogenMod 12S, AOSP 5.1, 64GB). Chrome—42.0.2311.47.
[3] iPad Air (iOS 8.3, 1st gen, 64GB). Chrome—42.0.2311.47, Safari—8.0.

Table 5: Average time to overlay an element (ms)

C. VALIDATION OF MESSAGEGUARD

We evaluated MessageGuard ability to support usable, content-based encryption research on a wide range of platforms. Additionally, we measured the performance overhead that MessageGuard creates. Our results indicate that MessageGuard is compatible with most web applications and has minimal performance overhead.

C.1 Ubiquity

We tested MessageGuard on major browsers and it worked in all cases: Desktop—Chrome, Firefox, Internet Explorer, Opera, and Safari. Android—Chrome, Firefox, Opera. iOS—Chrome, Mercury, Safari.

We tested MessageGuard on the Alexa top 50 web sites. One of the sites is not a web application (`t.co`) and another requires a Chinese phone number in order to use it (`weibo.com`). MessageGuard was able to encrypt data in 47 of the 48 remaining web applications. The one site that failed (youtube.com) did so because the application removed the comments field when it lost focus, which happens when focus switched to MessageGuard's compose overlay. We were able to address this problem with a customized frontend that required only five lines of code to implement.

These results indicate that researchers should be able to use MessageGuard to research content-based encryption for the web applications of their choice with little difficulty.

C.2 Performance

We profiled MessageGuard on several popular web applications and analyzed MessageGuard's impact on load times. In each case, we started the profiler, reloaded the page, and stopped profiling once the page was loaded. Our results show that MessageGuard has little impact on page load times and does not degrade the user's experience as they surf the Web: Facebook—0.93%, Gmail—2.92%, Disqus—0.54%, Twitter—1.98%.

Since MessageGuard is intended to work with all websites, we created a synthetic web app that allowed us to test MessageGuard's performance in extreme situations. This app measures MessageGuard's performance when overlaying static content present at page load (Stage 1) and when overlaying dynamic content that is added to the page after load (Stage

2). The application takes as input n, the number elements that will be overlayed in each stage. Half of these elements will require read overlays and half will require compose overlays.

Using this synthetic web application, we tested MessageGuard with six browsers and three values of n. We averaged measurements over ten runs and report our findings in Table 5. Performance for overlaying static content does not significantly vary based on the number of overlays created. In contrast, performance for overlaying dynamic content for most browsers seems to grow polynomial in the number of overlays added. Still, performance in the Firefox desktop browser demonstrates that this is not an inherent limitation of MessageGuard. Finally, we note that even in extreme cases (dynamic—$n = 1000$) overlaying occurs quickly (max 61 ms).

MessageGuard's low performance overhead indicates it is suitable for building responsive prototypes for testing by users. Moreover, if performance problems arise, researchers can be reasonably sure that the problems are in their changes to MessageGuard.

D. USER STUDY MATERIALS

This section of the appendix contains instructions and surveys from the user study that will allow others to replicate this research. The following items are included: A) instructions to the study coordinators that supervise Johnny and Jane; B) demographic questions; C) initial instructions to Johnny and Jane describing the user study scenario; D) instructions to Johnny and Jane regarding the tasks they must complete for each MessageGuard variant; E) survey questions Johnny and Jane answer after using each MessageGuard variant; F) post-study questions; G) and descriptions of the security of each key management scheme.

D.1 Study Coordinator Instructions

1. Have each participant sign two copies of the consent form. Give one copy to the participant to keep.
2. Use a coin flip to determine who is Johnny.
3. Johnny will remain in this room and Jane will go next door.
   (a) Ask the participant to sit down. Invite them to adjust the chair if they wish.
   (b) Tell them, **"You and your friend are in different rooms, and will need to work together to complete a task. During this task, we will provide you with some information that needs to be sent over email. Other than this information, you can feel free to communicate with your friend however you normally would. While you are waiting for email from your friend, feel free to relax and use your phone or the Internet"**
4. Do the following:
   (a) Start the audio recorder.
   (b) Open {Screen recording software}. Start recording.
   (c) {Open the survey}
5. Before using each system, the survey will instruct the participant to tell you they are ready to begin the next task. When they do so, complete the following steps:
   (a) (Johnny) Look at which system the participant will be using, and provide Johnny with the appropriate information sheet.
   (b) (Jane) Provide Jane with the generic information sheet.
   (c) Start the VM software and resume the snapshot.
   (d) Change the view to full screen-exclusive mode.
   (e) Notify the other coordinator which system will be used.
   (f) Record in the notes the order the systems are used.
6. During the course of the task pay attention to the following items:
   (a) (Jane) When Jane decrypts her email, give her the appropriate information sheet for her to complete the task..
   (b) Make notes of anything interesting you see.
   (c) If the participant sends sensitive information in the clear, make a note of this, then instruct them that they need to use the secure email system to send that information.
   (d) **Note how participants transmit passwords (e.g., phone call, text, email).**
   (e) During the study, participants may have questions for you. Answer any questions regarding the study task, but do not instruct participants on how to use the systems being tested. Instead, encourage them to continue trying.
   (f) In case users wrote their codes down incorrectly, we have included them at the end of this document.
7. When the task is complete, the participants will be instructed to tell you they have finished the task. When they do so, complete the following steps:
   (a) Ensure that the participants have correctly completed the task.
   (b) Exit exclusive mode.
   (c) Restore the snapshot.
   (d) Switch to the survey and have the participant continue the survey.
8. **When the survey is finished, ask the participant about their experience.**
   (a) Ask the participants about any problems they encountered during the study and how they dealt with them. Try and understand what the user was thinking. Also ask the participant if something in MessageGuard could be changed to address this issue.
   (b) Ask them about anything you felt was unusual or unique in their experience.
   (c) For each key management scheme **(follow the order they used the systems in)**:
      i. Ask participants who can read their messages. If unclear, ask them what would an attacker need to do to steal their secure email.
      ii. **Record whether the user correctly understood the scheme in the notes.**
   (d) For each key management scheme (not concurrent with previous bullet, **follow the order they used the systems in**):

i. **"I will now describe to you what an attacker would need to do in order to read your encrypted email. If you have any questions about my descriptions or how the systems work, feel free to ask."**
ii. Explain to the users the security provided by each scheme.
iii. Ask the participant if, based on this information, their opinion on any system changes.
iv. Ask the participant which system they would prefer to use in the real-world with their friends.
v. **Record this information in the notes.**

9. Close out the individual portion of the study.
   (a) Stop the video recording.
   (b) (Jane) Stop the audio recording, and bring your participant back to the main room.
10. Now that the participants are together, ask the participants about their experience.
   (a) How would your ideal email encryption system function? If you would like to, feel free to use the whiteboard to sketch ideas.
   (b) What did you think about doing a study with a friend?
11. Close out the study.
   (a) (Johnny) Stop the audio recording.
   (b) Clean the whiteboard if needed.
   (c) Thank the participants for their time.
   (d) Help them fill out the compensation form, and direct them to the CS office.

D.2 Demographic Questions

In our study, Johnny was shown these questions at the end of the survey, while Jane was shown them at the beginning of the survey. This was done to let Johnny get started working on the first task right away and to give Jane something to do while waiting for the first email.

What is your gender?

- *Male*
- *Female*
- *I prefer not to answer*

What is your age?

- *18–24 years old*
- *25–34 years old*
- *35–44 years old*
- *45–54 years old*
- *55 years or older*
- *I prefer not to answer*

What is the highest degree or level of school you have completed?

- *Some school, no high school diploma*
- *High school graduate, diploma or the equivalent (for example: GED)*
- *Some college or university credit, no degree*
- *College or university degree*
- *Post-Secondary Education*
- *I prefer not to answer*

What is your occupation or major?

How would you rate your level of computer expertise?

- *Beginner*
- *Intermediate*
- *Advanced*

D.3 Scenario Instructions

D.3.1 Johnny Scenario

In this study, you will be role playing the following scenario:

Your friend graduated in accounting and you have asked their help in preparing your taxes. They told you that they needed you to email them your last year's tax PIN and your social security number. Since this information is sensitive, you want to protect (encrypt) this information when you send it over email.

You will be asked to send this information using three different secure email systems. In each task, you'll be told which system to use and assigned a new PIN and SSN. After correctly sending the information, your friend will reply to you with a confirmation code that can be used to continue with the study.

D.3.2 Jane Scenario

In this study, you will be role playing the following scenario:

You graduated in accounting and have agreed to help a friend prepare their taxes. You have asked them to email you their last year's tax PIN and their social security number.

As part of the study, your friend will send you this information three different times. Each time, after receiving their PIN and SSN, you will be provided with a confirmation code and a PIN number to send to your friend so that both of you can continue with the study.

D.4 Task Instructions

D.4.1 Johnny's Task

Johnny repeats the following for each MessageGuard variant.

Tell the study coordinator that you are ready to begin this task.

System: **MessageGuard—{Insert encryption scheme}**

In this task, you'll be using **MessageGuard—{Insert encryption scheme}**. The system can be found at the following website: {Insert url}

Please encrypt and send the following information to your friend using MessageGuard—{Insert encryption scheme}:

SSN: {Task SSN}
PIN: {Task PIN}

Enter the confirmation code provided by your friend.
Enter the PIN provided by your friend.

Once you have received the confirmation code and PIN from your friend, send an email to your friend letting them know

you received this information. After you have sent this confirmation email, let the study coordinator know you have finished this task.

D.4.2 Jane Task

Jane repeats the following for each MessageGuard variant.

Tell the study coordinator that you are ready to begin this task.

Please wait for your friend's email with their last year's tax PIN and SSN.

Enter your friend's SSN. Include dashes.
Enter your friend's PIN.

Once you have written down your friend's SSN and PIN, let the study coordinator know that you are ready to reply to your friend with their confirmation code and PIN.

You have completed your friend's taxes and need to send them the confirmation code and this year's tax PIN from their tax submission.

Since your friend used MessageGuard—{System name} to send sensitive information to you, please also use MessageGuard—{System name} to send them the confirmation code and PIN.

- Confirmation code: {Task SSN}
- PIN: {Task PIN}

Once you have sent the confirmation code and PIN to your friend, wait for them to reply to you and confirm they got the information. Once you have gotten this confirmation, let the study coordinator know you have finished this task.

D.5 Survey

Johnny and Jane complete the following survey after each MessageGuard variant.

You will now be asked several questions concerning your experience with **MessageGuard—{Insert encryption scheme}**.

Please answer the following questions about {Insert encryption scheme}. Try to give your immediate reaction to each statement without pausing to think for a long time. Mark the middle column if you don't have a response to a particular statement.

<SUS Questions>

What did you like most about using MessageGuard—{Insert encryption scheme}?

What would you change about MessageGuard—{Insert encryption scheme}?

Please explain why.

D.6 Post-study questions

You have finished all the tasks for this study. Please answer the following questions about your experience.

Which system was your favorite? ***(Ask the coordinator if you are unclear which system is which.)***

- *First system: MessageGuard—{First system name}*
- *Second system: MessageGuard—{Second system name}*
- *Third system: MessageGuard—{Third system name}*
- *I don't like any of the systems I used*

Please explain why.

Please answer the following questions. Try to give your immediate reaction to each statement without pausing to think for a long time. Mark the middle column if you don't have a response to a particular statement.

I want to be able to encrypt my email.
<Likert scale>

I would encrypt email frequently.
<Likert scale>

In the password-based version of MessageGuard, the passwords you entered would be deleted when you exited Chrome. This meant that others using your computer would not be able to read your encrypted email.

In contrast, the PKD and IBE versions save your encryption keys, and anyone logged into Gmail on your computer can read your encrypted email. This could be changed by adding a **master password** to MessageGuard. You would select your master password when you install MessageGuard.

From then on, whenever you open your browser, MessageGuard would require you to enter your master password before functioning. This would protect your IBE- and PKD-encrypted emails from others who use your computer.

Would you prefer MessageGuard to use a master password?

- *Yes*
- *No*

Please explain why.

D.7 Key Management Descriptions

PKD: "In the {first, second, third} system you tested, your email was secured using PKD. In PKD, when you installed the system, a lock and key were created. The lock was stored on the MessageGuard website, allowing anyone to download it and use it to encrypt email for you. The key is kept on your own computer and is needed to decrypt your email. To read your encrypted email, an attacker would need to break into your computer and steal this key." "In PKD, your recipients need to install the system and generate their lock and key before you can encrypt and send email to them. If you lose or delete your key, email encrypted with your lock will be inaccessible."

IBE: "In the {first, second, third} system you tested, your email was secured using IBE. In IBE, anyone can encrypt email for you, and the key to decrypt that email is stored on the MessageGuard website. To read your email, an attacker would need to break into the MessageGuard account you created during the study, and steal your key. Because the MessageGuard website does not have access to your email, it cannot decrypt it."

Passwords: "In the {first, second, third} system you tested, your email was secured using a password you or your friend chose. To read your email, an attacker would need to steal or guess that password."

When is a Tree Really a Truck? Exploring Mental Models of Encryption

Justin Wu
Brigham Young University
justinwu@byu.edu

Daniel Zappala
Brigham Young University
zappala@cs.byu.edu

ABSTRACT

Mental models are a driving force in the way users interact with systems, and thus have important implications for design. This is especially true for encryption because the cost of mistakes can be disastrous. Nevertheless, until now, mental models of encryption have only been tangentially explored as part of more broadly focused studies. In this work, we present the first directed effort at exploring user perceptions of encryption: both mental models of what encryption is and how it works as well as views on its role in everyday life. We performed 19 semi-structured phone interviews with participants across the United States, using both standard interview techniques and a diagramming exercise where participants visually demonstrated their perception of the encryption process. We identified four mental models of encryption which, though varying in detail and complexity, ultimately reduce to a functional abstraction of restrictive access control and naturally coincide with a model of symmetric encryption. Additionally, we find the impersonal use of encryption to be an important part of participants' models of security, with a widespread belief that encryption is frequently employed by service providers to encrypt data at rest. In contrast, the personal use of encryption is viewed as reserved for illicit or immoral activity, or for the paranoid.

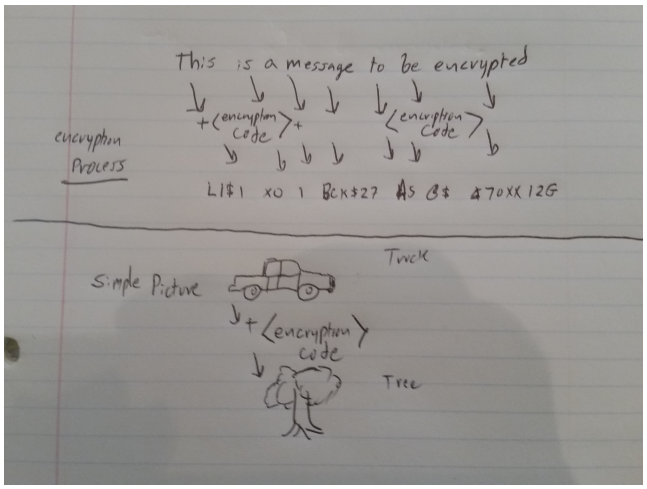


Figure 1: One participant's diagram. The top half reveals how they visualize the encryption of textual data to occur, while the bottom half shows their impression of how pictures are encrypted.

1. INTRODUCTION

Many security and privacy experts advocate for the widespread adoption of encryption, both as a security measure for protecting against third-party attackers and as a privacy preserving tool. Indeed, recent years have seen encryption incorporated by default into popular software, such as instant messaging apps like WhatsApp and in mobile devices operating systems like Google's Android and Apple's iOS. However, previous studies have shown that when users are actively involved in the process of encryption, they can struggle to accomplish this task [5, 10, 11, 15, 17, 23]. This is important because the misuse or misapplication of encryption technologies can be devastating. Those who incorrectly use encryption tools may falsely believe themselves protected by technology whose guarantees no longer hold due to their mistakes. Perhaps more dangerous still, users who do not understand the risks of the technology could well find themselves in a situation of self-imposed denial of service, permanently locked out of accounts and data which they have lost the keys for, and—unlike with passwords—no one who can help them recover them.

In the context of encryption, where the cost of users' mistakes can be grievous, circumventing the possibility of user error by transparently incorporating encryption into software, thus bypassing the user entirely, is a desirable and effective option. Indeed, in scenarios where encryption has already achieved widespread deployment—smartphone encryption, TLS / HTTPS, and secure messaging apps—it has succeeded by doing precisely this. Unfortunately, while this approach is indeed effective, its applicability is not without limits [6, 12]. Automation is not always a perfect solution; even the best software at times encounters errors that require user interaction to proceed [22]. With high levels of automation, users likely lack the context necessary to make the correct response. In two cases where encryption has been transparently applied, for example, studies of the efficacy of TLS browser warnings [2] and of the authentication ceremony in secure

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

messaging apps [14] have shown that users are confused by generated warnings and unsure of the correct action to take.

It is this context which is the focus of our study: when users *must* interact with encryption tools, how do they make sense of them? If all you knew was that a tool "used encryption," what would you understand about it, and how would your *mental model*—the representation of your thought process of how something works—guide your efforts to interact with it?

The functional nature of mental models has serious implications for design: accurate or inaccurate, someone's mental model is what they rely on when they interact with and troubleshoot systems [19]. Subsequently, a proper understanding of mental models can be beneficial in user-centered design, affecting both the intuitiveness of tools as well as the efficacy of our communication with users [21]. Extending beyond use, mental models also have implications for adoption, as users' perceptions of the utility of a security or privacy tool is a critical element in their decision of whether or not to adopt.

There is some evidence from previous research that suggests that users' mental models of encryption are flawed or incomplete [1], although there has heretofore never been a systematic effort to profile and understand what these mental models may be. To help us better understand how users perceive encryption, we have executed a qualitative study consisting of 19 semi-structured interviews focused on profiling users' models of encryption. We thus present the first directed effort to explore users' mental models of encryption.

We explored three facets of participants' perceptions of encryption: what it is, how it works, and its role in their lives. Interviews with participants were divided into two halves. The first half of the interview was designed to elicit their mental models of encryption and pertained to the mechanics of encryption. In addition to questions probing these concepts, participants were tasked with "encrypting" both text and non-textual data (a picture) in a brief, but illustrative diagramming exercise, an example of which can be seen in Figure 1. The second half of the interview focused on how encryption might be used, and presented participants with three distinct encryption use cases for discussion.

By analyzing the results from our interviews, we have categorized participant responses into a set of four mental models. These models vary in complexity and detail, but ultimately reduce to a functional abstraction of access control. Based on our observations about these models and other remarks made by our participants, we outline a number of important implications regarding the future of encryption software design and risk communication.

The contributions made by this work are as follows:

1. **Presents first directed effort at exploring mental models of encryption.** Previous research that has investigated perceptions of encryption have only done so as parts of larger efforts, such as to understand secure communication or general security behaviors. This work is the first to focus exclusively on this issue.

2. **Identifies four mental models of encryption.** Based on interviews consisting of both verbal questioning and a diagramming exercise, we identify four properties—some functional, some structural—that comprise a mental model of encryption. By correlating individual participant responses with these properties in a data matrix, we compiled a set of four mental models of encryption that highlight the differences in the way people perceive the structure and function of encryption.

3. **Outlines implications for design and risk communication that arise from participants' perceptions of encryption:**

   Encryption is restrictive access control. Despite the varying levels of detail and complexity in participants' mental models of encryption, they nevertheless produce the same functional abstraction: restrictive access control. Designers should contemplate whether encryption contexts align with this model and, if not, consider presenting users with a functional model for interaction that more closely aligns with that of access control.

   Models of encryption are of symmetric encryption. Participants' functional models for the role of keys and sharing keys coincides closely with symmetric encryption. Their structural view of what keys are, furthermore, does not align well with asymmetric encryption, and could be a major usability obstacle in the use of software employing public key cryptography unless an alternative model is presented to users.

   Confusion about encryption strength. Even participants with similar mental models described a wide array of timescales in which they believed encryption could be broken. A number of factors appear to play a role in this discrepancy, such as varying perceptions about the capabilities of attackers.

   Grassroots adoption is also a public relations battle. Participants largely viewed encryption as already being deployed by the service providers they deal with regularly, such as banks and online merchants. However, when it turned to personal use of encryption, they felt strongly that the use of encryption was the domain of those engaging in illegal or immoral activity, or the paranoid.

2. METHODOLOGY

The data presented in this work is sourced from 19 semi-structured interviews conducted with participants from the United States. Each interview lasted between 30–60 minutes, and participants were each compensated the equivalent of $15 USD for their time. Our study was approved by our institutional review board.

2.1 Recruitment

Participants were recruited using the Prolific research platform, and interviews were conducted until data saturation was reached. Prolific allows for prescreening of potential participants by filtering for a number of demographic variables. To maximize ease of communication, we limited the pool of potential candidates to only those Prolific participants who both reside in the United States and speak English as a first language.

Our study was listed as a task on Prolific, advertised as "phone interview on an Internet-related topic." Interested participants self-selected into the study and registered via an online scheduling form that was linked in the Prolific task. In accordance with Prolific's requirements, we did not collect

any personal information beyond first name and Prolific email address (Prolific provides an associated email to each account for communication purposes). Participants who scheduled were contacted via their Prolific email, and three things were communicated: (1) the study coordinator's phone number, (2) an instruction to have pen and paper ready during the scheduled time in preparation for the diagramming exercise, and (3) a request for consent to transcribe and share study data after anonymization.

2.2 Demographics

19 participants in total participated in our study. Our sample skewed heavily male (n=13, 68.4%). Participant age ranges were fairly diverse: 7 between 20-29, 5 between 30-39, 5 between 40-49, 1 between 50-59, and 1 between 60-69. We had some students (n=6, 31.6%), but most of our participants were not in school. While we did not seek explicit socioeconomic demographics, we know that 7 of our participants had full-time employment (36.8%), while the remainder worked either part-time or not at all.

2.3 Study design

Interviews were conducted remotely, by phone. They were semi-structured, with a set of questions that were asked of each participant, and others that were asked only of specific individuals as their responses warranted. The interview guide can be found in the Appendix, and interview transcripts have been made available for download at `https://mentalmodels.internet.byu.edu`.

2.3.1 Introductory information

As each interview began, the coordinator informed the participant that the topic for discussion would be technical in nature, and that it was expected that there would be questions for which they had no answer; in such an event, they were to instead offer their best guess. While such answers might be speculative in nature, and thus seem undesirable, our goal was to understand how participants would perceive software if all they knew was that it used encryption in general, which mirrors this situation.

They were then reminded of the need for pen and paper for the drawing exercise, and asked to prepare them if they had not already done so (no participants actually needed the reminder). Finally, participants were asked to give a brief description of their line of employment and/or area of study, to get a sense for technical background. They were also asked to describe what types of devices they own and use, and what types of tasks they perform on them.

2.3.2 Encryption mechanics

At this point, the main portion of the interview began. Participants were first asked to describe what came to mind when they hear the word "encryption." Follow-up questions were asked as necessary to have participants clarify their responses. They were typically asked to explain where they might have seen or heard the term used, to seek insight into the contexts in which they believe encryption appears. They were then asked to describe what types of imagery they associate with the term, with the goal being to help us understand what types of visual metaphors might work well when communicating about encryption. As discussion on these questions drew to a close, the coordinator initiated the diagramming exercise.

2.3.3 Diagramming exercise

The diagramming exercise consisted of two tasks where we asked participants to illustrate the encryption of textual and non-textual data respectively. The latter was particularly important because we supposed that imagery of textual encryption would be prevalent in popular media, and were curious how participants might react to the idea of encrypting something else.

When this segment of the interview began, participants were first informed that the point of the exercise was to help the coordinator visualize what the participant had in mind, and not a test of artistic ability. They were asked to write the following sentence on their paper, "This is a message to be encrypted." It was then explained that they were to imagine encrypting this message; their task would be to draw what they imagined would happen. No explicit instructions were given as to form to avoid influencing participants, which unfortunately led a couple participants to generate diagrams that were too lacking in detail to have interpretative value. Participants were given as much time as needed to finish the task, being told to alert the coordinator when they were done. Once they had finished this first task, they were then told to draw "a simple picture, such as a tree, cloud, or stick figure," and repeat the same task, diagramming what they imagined would happen if this picture were to be encrypted. Upon completion of this second task, they were asked to text or email a picture of their work to the coordinator.

After the picture of their drawing was received by the coordinator, discussion about its contents began. Participants were asked to walk the coordinator through their drawings and explain the various elements of their illustration, with the coordinator prompting for clarification as needed. Participants were also asked to explain how they imagined the process of decryption would work: the coordinator described a scenario in which an encrypted message was sent to a friend or family member, and asked what the receiving party would need to do to read the original message. Finally, participants were asked to characterize how difficult they expected it would be for two groups—hackers (representing individual attackers) and the NSA (representing institutional resources)—to break encryption.

2.3.4 The role of encryption in life

After the diagramming phase, participants were informed that the discussion was about to transition away from what encryption is and how it works to how it might be used. They were then asked if they thought encryption played any role in their life. If participants' responses were restricted to institutional use—such as by banks, the government, online vendors—we also asked them if they thought there might be individuals who used encryption for personal reasons.

Finally, the last segment of the interview involved introducing each participant to three encryption use cases that are available to normal users. These are mobile device encryption, HTTPS, and secure messengers. Some participants had already mentioned one or more of these prior to this part of the interview; if they did so, the topic was discussed at that earlier point. Otherwise, the three use cases were presented and discussed in this order.

With the partial exception of HTTPS, our aim with this part of the interview was not to assess participants' famil-

iarity with the use cases in question, but rather to assess their impressions of the respective utility of these encryption options. Accordingly, for each use case, a short preface was given in which we introduced the use case to the participant before discussing it with them, enabling them to share their thoughts on each scenario even if they were not previously aware of its existence.

Mobile device encryption

Nearly all iOS devices are encrypted, due to a decision by Apple to enable encryption by default on devices running iOS 8 or higher. As a result of fragmentation issues, the number of Android devices that are encrypted is much lower, although it is expected to improve with time as new Android devices now also ship with encryption enabled by default. Because of the relative ubiquity of smart devices, their importance in daily life, and the likelihood of their being encrypted, they serve as a useful first look at the perceived utility of encryption in daily life.

We explained to our participants that both Android and iOS devices had functionality that allowed for encryption. Each of our participants had, and regularly used, at least one mobile device, and thus had the necessary context needed to share their impressions. We asked what they thought it meant to encrypt their smartphone or tablet, drawing a juxtaposition to the encryption of data as had been discussed previously. Participants were then asked why they imagined someone might want to encrypt their device, that is, what would be protected by doing so? Finally, we asked them to explain who they considered enabling device encryption would protect them from.

HTTPS

User interactions with HTTPS are well-studied, and the technology is extremely widely deployed, with efforts in play to make it ubiquitous; it thus provides another useful example for examination. With this use case only, we began by asking participants if they had ever noticed an "HTTPS" or lock icon in their browser, located near the address bar. Since all had, we then asked them to describe what they believed these indicators to mean. After they had responded, we informed them that it represented an encrypted connection between their browser and the web server of the site they were visiting. We then asked them to describe what information they believed it was meant to protect and whom it would protect them from.

Secure messaging apps

Our final examined use case, secure messaging apps, are a class of instant messaging apps that use end-to-end encryption. A handful have seen fast-growing adoption, albeit not for their security properties [1], and are a growing area of interest for security research. We began by asking participants if they used any popular instant messaging apps, such as WhatsApp or Facebook Messenger, to establish a frame of reference. We then explained that secure messaging apps are a subset of these types of apps, the difference being that they encrypt all communication made via the app. Participants were asked why they imagined someone might wish to encrypt daily communications, and not just what is more commonly considered sensitive data, such as financial information. They were also asked who they thought encryption was meant to protect their communications from.

2.4 Data analysis

The audio of each interview was recorded and transcribed. These transcriptions were then jointly coded by the study researchers via open collaborative coding per the conventional content analysis approach. Coding was only performed in meetings where both authors were physically present and jointly reviewing the transcripts. Any disagreement was resolved via on-the-spot discussion, and thus we did not calculate inter-rater agreement.

We separated participant responses into two types: those to be explored as individual themes, and those that we identified as serving functional or structural roles in users' mental models. From our codes, we chose four structural and functional properties that comprise the mental model of encryption. We then went through each participants' transcripts anew, and filled in a matrix, matching each participant's responses to the corresponding mental model components. Finally, we grouped these individual models into sets based on what we identified as critical dividing boundaries derived from fundamental structural or functional differences. Each of these sets became one of our final mental models, and represents a unique abstraction of encryption.

3. RESULTS

In executing this study, our goal has been to explore the space of user perceptions of encryption and not to quantify the extent to which users possess certain views. Accordingly, we do not provide quantitative measures of the frequency with which various opinions were expressed, and instead have attempted to characterize a representative set of the issues our participants described.

3.1 Impressions about encryption

At the start of each interview, we began by asking each participant to describe what came to mind when hearing the word "encryption." This served to give us some sense of the context in which they imagine encryption exists—the environments in which it is used and the purposes it serves. Responses broadly fell into three categories: characterizations of encryption itself and then contexts and current events that they associate with encryption.

A number of participants described encryption itself, including terms such as *"algorithm," "encode,"* or *"secret code."* Participants also described imagery that they associated with encryption, such as Lloyd[1], who related encryption with *"that scene in the Matrix, where the letters are falling out of the sky and it's like the code of the Matrix."* This imagery of long, indecipherable strings of symbols was commonly shared by our participants, and particularly evident during the diagramming phase of the interviews.

Encryption was also clearly associated with security and privacy in our participants' minds. They mentioned both broad properties such as *"security," "safety," "protect,"* and *"privacy"* in addition to specific contexts where they imagined encryption was used such as *"access control," "email,"* and *"passwords."*

[1] All names used are pseudonyms

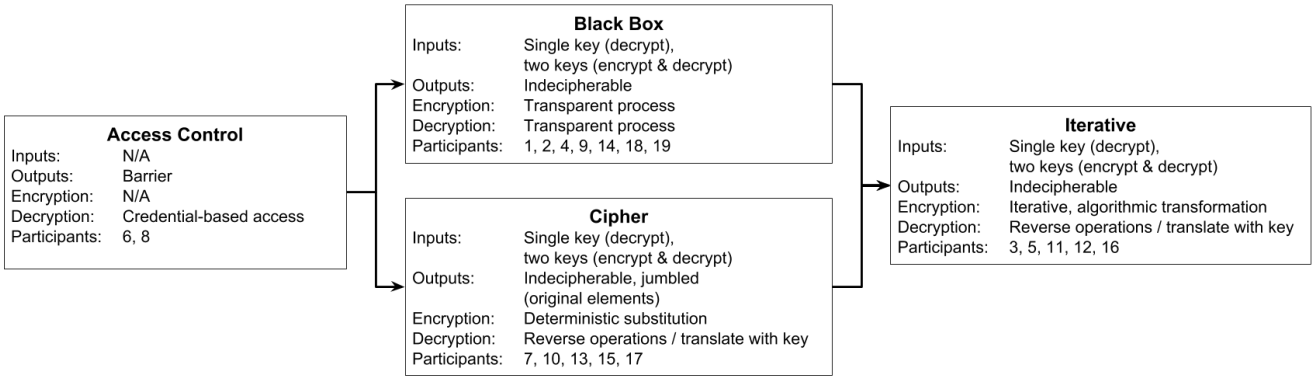


Figure 2: The four mental models we identified. Detail increases going from left to right, while models at the same level do not differ in complexity, but rather are fundamentally different in other ways.

Current events involving encryption also stood out to participants when they became notable enough to be covered by mainstream news media, as with Edward, who explained that he noticed such events *"[w]hen they're big enough to show up in the normal news. I'm not putting much effort into looking for this."*

3.2 Mental models of encryption

Mental models have been described as having alternatively structural and functional properties. Structural properties describe how participants perceive the internals of how encryption works, while functional properties characterize how participants interact with encryption. Because, as we expected, our participants largely did not have any experience with encryption tools, our focus is primarily on the structural properties, although we did evaluate some functional aspects via presented scenarios. Based on our coding of participant responses, we compiled a set of four properties that comprise a mental model of encryption:

1. **Inputs to encryption/decryption:** This property is taken from the follow-up questioning segment of the diagramming exercise, and describes what inputs, if any, participants believe are necessary for the encryption process (aside from the object to be encrypted).

2. **Encryption output format:** This property is taken from each participant's diagram, and refers to how they depicted the output of encrypting the text/picture.

3. **Encryption process:** This property is taken from each participant's diagram and follow-up questioning, and characterizes what they imagine the *process* of encryption itself entails.

4. **Decryption process:** This property is taken from follow-up questioning about each participant's diagram, and characterizes how they imagine the process of decryption occurs.

Before continuing further, there is something we wish to impress upon the reader: the "obvious" solution to the issues we explore—i.e., improving the accuracy of users' mental models—may not be as simple as it seems. If we leave tool design static, attempting instead to correct inaccurate mental models, we run into the issue of the difficulty of effective communication regarding encryption, both in terms of message (what to convey) and medium (how to convey it). Moreover, this approach places the responsibility for change upon users.

Alternatively, we can alter software design to more closely align with users' mental models. This places the onus for change upon developers, who we feel have both stronger incentives and the knowhow to do so. This is the approach we espouse in this work. We do note, however, that if the community as a whole can make headway on communication efforts, the most productive approach will be to tackle this problem from both ends.

The following mental models are listed in general order of complexity/detail, proceeding from the simplest to the most detailed, although some models may be "equally" complex, and differ instead on certain critical details. A diagram of these models and their relationship can be found in Figure 2.

3.2.1 Model #1: Access control

The first and most basic model of encryption provides only the most minimal abstraction of access control. Unlike the remainder of our participants who recognized that encryption transformed the source data somehow, the two participants who possessed this model instead viewed encryption as an extension of credential-based access. As can be seen in Figure 2, this model has multiple "N/A" entries, to signify that their model did not even allow for the existence of these properties.

When first presented with the diagramming task, Edward immediately felt at a loss to characterize what encryption might look like. He asked, *"Does it actually have a look? Is it just something that 'what comes in your mind' or does it actually have a certain look? It's all just... online. It's all zeroes and ones."* When later discussing his illustration (Figure 3) with the coordinator, he added, *"I didn't really think of a physical change, really, aside from something coming up on a barrier,"* which indicates that he had likely never previously considered that encryption might actually transform data somehow, instead associating encryption solely with a notion of access control.

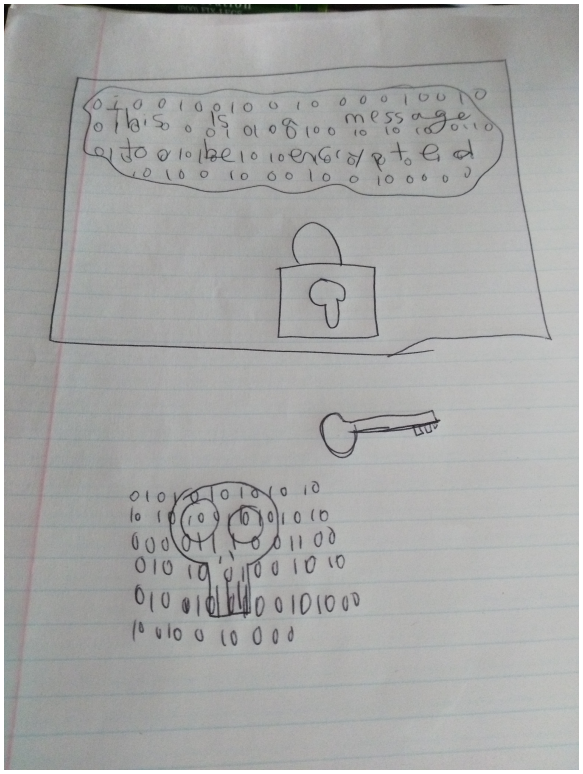


Figure 3: Edward's barrier diagram, which represents "the symbolism of locking." The skull is "just blocking, something that's just blocking the message."

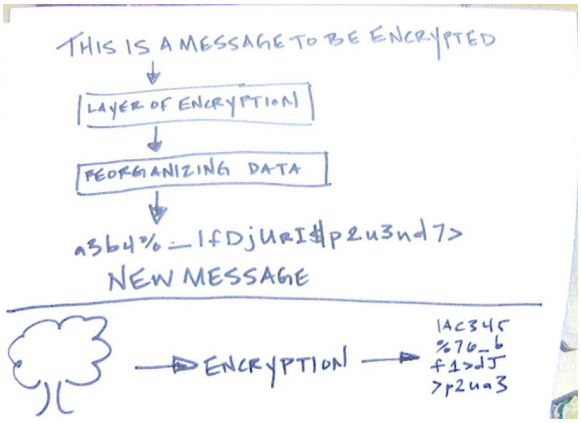


Figure 4: Wally's black box encryption. The data to be encrypted goes through a "layer of encryption" that runs "in the background."

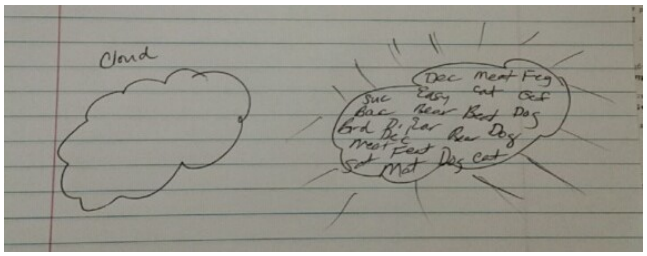


Figure 5: Eva's watermark diagram.

Selena employed a similar metaphor, analogizing encryption to a "wall": *"I mean, I've like heard of protection but I don't know exactly what it means. And I'll be honest, I'm really a person who's been—I don't want to say sheltered, but– But I'm thinking a wall."*

3.2.2 Model #2: Black box

The next model advances the previous model slightly. It is functionally similar to the first model in that participants with this model similarly viewed it as an extension of existing credential-based access. However, participants who had this model did understand that encryption would transform the source data—that is, they understood that encryption was an active process, though they did not have a strong sense of how this process functions. As Diana explained, *"You don't really know what it means, but you know that it means something; you just don't know what. I feel like that kind of works with encryption, because when they're like, 'oh, it's encrypted, your stuff is protected,' I kind of know what that means, but I don't really know what they're doing to make that happen."*

Wally explained how encryption is something that *"sort of run[s] in the background."* The software would transparently handle both encryption and decryption as it *"gets the information and gobbles it all up and translates it into something else and as it comes out the other side it's sort of put back together."* This black box perception of encryption is well-captured in his diagram, shown in Figure 4.

In another example, Figure 5 depicts Eva's diagram for encrypting the cloud that she had drawn. She, as with many of our participants, did not have a concept for the digital representation of image data, and so did not first transform her illustration into a digital representation before transforming it. Instead, she drew the *"only thing [she] could think of"*—a watermark—because she knew that *"they put watermarks to keep people from stealing,"* and she associated encryption with protection against theft.

This model is more functional than structural, with some conception of what encryption will do for you, but not how it works. Subsequently, participants had to analogize from other security mechanisms that were more familiar to them, such as a "watermark" in Eva's case or "password dots" in Diana's. Because this model correctly perceives encryption as a process, these participants did have some notion of necessary inputs, even if, due to a loose conception of how encryption worked, that input might simply be the encryption algorithm itself.

When asked how decryption might occur, Diana shared that a "key" would be needed—*"I think they'd have to send a key with it, or else I wouldn't know what to do with it."* What, then, was this key? Her response was one echoed by many of our participants: a key is a reversed list of the operations executed during the transformation (encryption) process. *"Well, if the letters were sort of mixed-up randomized, then I think, from that, they could make a key based on how it's been randomized, where the letters went or where they originally were, and they could hand me that, and from there, I could*

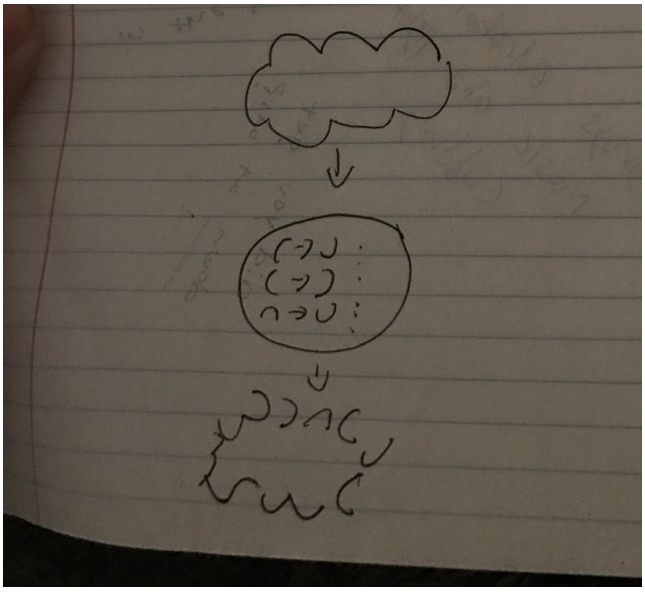


Figure 6: Allen's ciphered cloud. "The picture is split into different sections or elements and each element turns into a different element"

unrandomize the letters." (A handful of others possessed a similar, if slightly different view: a key was a like translator that guided the reversal.)

3.2.3 Model #3: Cipher

The cipher model differs from the black box model in that participants with this model had a clear sense of how the "transformation" process of encryption works: constituent portions of the source are deterministically substituted into another form, i.e., a cipher. As can be seen in Figure 6, this substitution cipher behavior extended even to the encryption of image data: Allen defined deterministic transformations of the various curves in his cloud to be enacted by the encryption process.

For example, Fred described *"a simple algorithm for it,"* where each letter would be replaced by the letter a specified distance from it alphabetically. *"[I]f the original letter is 'T,'—that's the first variable—it'll add four letters, so it'll go 'U,' 'V,' 'W,' and end on 'X.' And then every letter after that, continue to add 4 to each. The cipher is '+4' basically."*

3.2.4 Model #4: Iterative encryption

Participants with this model were the most descriptive and detailed. When it comes to the properties we have discussed in other models, they share only some superficial similarities. They all believed that encryption would transform the source data. Their notion of what type of operations are performed by the encryption process varied. Their impressions of the difficulty of breaking encryption similarly varied, although they generally imagined it would be a non-trivial task. We chose to unify these participants under a single model, however, because their models jointly exhibit a shared property not found elsewhere: they explicitly described the encryption process as an iterative one involving multiple passes over the source data in order to produce the encrypted output.

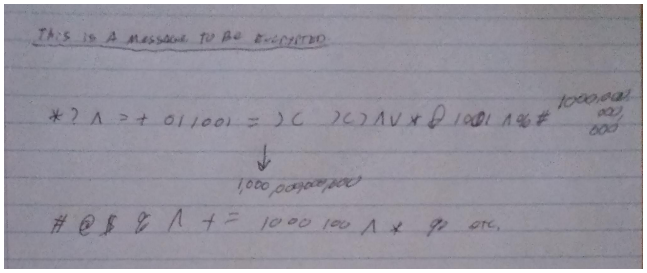


Figure 7: Franklin's iterative text encryption process.

For example, Figure 7 shows Franklin's depiction of how textual encryption would work:

"Interviewer (I): Okay, great. Let's start with the part at the top where you're encrypting the message. Can you explain what that second line is to me? You have something on the left and an equals sign and then something on the right and it says one trillion or some very big number. Can you explain that to me?

Participant (P): Yeah, it's converted to those various symbols to the power I've put it in.

I: And when you say 'to the power,' what do you mean, what are you describing?

P: The conversions are that many times or roughly that many times."

Their model for what types of operations were to be performed at each iteration varied from person to person. Lloyd, for example, described a model where *"scrambling"* happens *"a bunch of times."* Franklin believed encryption to be a mathematical process at heart, explaining that if math were responsible, then encryption would *"be uniform, it would follow certain specific laws, it would be rational."* Nicole described a complex process where encryption *"would be either swapping or rearranging and [...] adding an infinite amount of extra garbage to it or just infinitely changing all the different parts of it to be different things."*

3.2.5 Shared secrets

One vital aspect of interaction with encryption software is an understanding of the requirements for *de*cryption. To help us understand how participants envisioned this process, we asked them to imagine a scenario where they had encrypted a message for a friend or family member, just as in the diagramming exercise. We then asked them to describe what they felt the receiving party would need in order to read the original message. As mentioned earlier, participants described needing something to reverse the operations that encryption had performed, which they frequently called a key.

When asked how their friend or family member would acquire access to such a key, most participants did not have an idea or answered that they would make arrangements in-person. Some thought about sending the key over another medium, but at least one participant, Lloyd, recognized the circular nature of this problem: if the goal was to establish a secure channel, then wouldn't you need an existing secure channel to convey the key? *"I have no idea. Arcane computer wizardry?*

'cuz you think you'd have to encrypt the key and encrypt the encryption on the key–" Brent, on the other hand, imagined that perhaps keys might be tied to login credentials, such that *"maybe you just receive them when you log on or view something that's encrypted."*

3.2.6 Encryption strength

We asked participants to contextualize their responses as timescales in which two parties would attempt to break an encrypted message: hackers and the NSA. Conceptions of what it would take to break encryption were all over the board, ranging from minutes to years to practically impossible. A deeper discussion of these responses is included in Section 4.4.

3.3 The role of encryption

Perceptions of encryption extend beyond what it is and how it works to more functional views, such as the role it plays in life. As part of our effort to better understand how people view encryption, we asked participants to describe the perceived role of encryption in daily life in general and also in three use cases that we presented for discussion.

3.3.1 Encryption in daily life

When we first planned to include a question about the role of encryption in daily life, we were concerned that participants would not think that encryption was at all present in their lives, and that we would subsequently be unable to glean anything from their responses. To our pleasant surprise, however, all of our participants responded immediately with examples to this question. Indeed, their answers throughout our interviews make it apparent that encryption plays an important role in our participants' models of computer security as a whole.

Nearly all participants felt confident that any service providers they engage with that deal in sensitive information proactively encrypt their data, with banks and major online vendors first coming to mind. Most participants seemed to associate encryption with online activity, although a couple participants did mention credit/debit cards, a likely reference to EMV chip technology. Despite this online-oriented view of the entities responsible for encrypting their data, however, it was clear that—excepting those few participants who were explicitly aware of TLS and its purpose—participants simply did not have a model allowing for encryption of data in transit, only at rest. For example, Diana informed us that *"I think once you send the data or whatever, that it's not really yours anymore because now they have it, so maybe once they get it, they can do whatever they want with it, so they can encrypt it that way. Once you send the data and they get it, they can fuzz it or jumble it or do whatever they do when they encrypt it, and then you're good to go."*

Because our participants unilaterally brought up institutional use of encryption in response to this question, we also asked whether they imagined there were individuals who used encryption in personal contexts as well. Participants all believed that there were, although their responses centered on sensitive contexts, whether that be business interests such as investment information or intellectual property or for more nefarious uses, such as illicit or immoral activity.

For example, when we asked Diana about the individual use of encryption, the following dialogue played out:

"I: What about individuals as opposed to institutions? So not talking about a company or the government, do you think there are people that use encryption on a regular basis?

P: Maybe if they're doing something illegal?

I: And why do you think they would be using encryption in that case?

P: So they don't get caught?

I: So you mean to hide what they're doing from other people?

P: Yeah.

I: Any other examples that you might be able to think of?

P: Maybe if they're an entrepreneur and they're making something that they don't want to be stolen, they'd use encryption.

I: So again, basically any time what you're doing is sensitive?

P: Yeah."

Participants also recognized that there might be some who employed encryption out of generic privacy concerns, but typically classified such concerns as *"paranoid."* Nicole, who sometimes needed to use encryption tools at the request of clients, characterized the situation in this manner: *"For some people, I think it's a level of paranoia, almost. To have everything need to be encrypted. But for some people, particularly that are in the tech industry, it's almost like a biblical need. So when I'm dealing with someone who's really into encryption, I have to think of it from their standpoint of a desire for privacy and security– Of course, the paranoia that someone's gonna care."*

In general, the personal (non-business) use of encryption seemed to be viewed quite negatively: either you use encryption because you have something bad to hide or because you're paranoid.

3.3.2 Use case #1: Smartphone encryption

The first use case we presented for discussion was that of smartphone encryption. All of our participants used mobile devices—many had both smartphones and tablets—and thus all had the context necessary to understand this use case. When we asked participants to explain what they believed smartphone encryption meant, responses primarily viewed it as a form of access control, tied to the passcode lock they already had on their devices. In one example, Abe explains that, *"I imagine that it means to take the data on it and do the same thing: put it in a code that's only able to be broken by you, by something like your passcode or thumbprint."*

Some participants recognized that encryption would protect their storage medium itself, such that *"if someone stole my phone and they didn't have my passcode, they wouldn't be able to access my phone's hard drive or storage and read all my data"* (Allen). Some, however, just saw it as an additional protection over the passcode lock of unknown nature, such as Brent: *"It might just add another extra layer of it, 'cuz I thought the password lock was up there in terms of protection because it's a thing only you would know unless someone used social engineering to figure it out. I feel like it would just add another layer."*

Participants viewed their phones as important stores of personal data, with encryption potentially protecting items such

as login info (via apps), photos, contacts, and texts. When describing who encryption was meant to protect their phones against, the ever-present catchall of "hacker" was present, but participants also described the need to defend against physical threats (when devices are either misplaced or stolen) and law enforcement. For example, Mary first described smartphone encryption as protecting things one *"wouldn't want other people to see."* When asked to clarify who she meant by "people," she explained, *"I was just gonna say somebody who just steals your phone, but yeah. Probably hackers too because there was the huge photo dump with the iCloud stuff."*

3.3.3 Use case #2: HTTPS

As mentioned earlier, with our second use case—HTTPS—we began by asking if participants were familiar with seeing either "https" in the address bar or the lock icon nearby. All participants were, and so we asked them to explain what they believed these indicators represent. While a small number were fully aware of TLS and its purposes, by and large, participants responded that they were indicators of site security. Edward, for example, believed that it meant sites had been reviewed by a third party and received a seal of approval: *"maybe it's just another entity, like a government entity, that would review certain sites and give a seal of approval. But other sites, that are newer or not as established, that don't have that because they're not under review."*

Interestingly, a couple had previously clicked on these indicators and read a little about them, and knew that HTTPS involved certificates, although they nevertheless still conflated these with site security. Brent explained, *"it informs me of who it belongs to, like what company stands behind it and basically it's like a certificate of who we are, we are authorized, and we are secure."*

Participants' model for what types of information were being protected involved the sensitive data they conveyed when online, such as credit card information when shopping at online merchant sites. Interestingly, while there was some variance in their model of technique and/or target, i.e., the site or you directly, the attacker model was consistently that of hackers.

3.3.4 Use case #3: Secure messaging apps

Our final use case was that of secure messaging apps. While not all of our participants had previously used instant messengers, and thus lacked the direct comparison, all regularly texted and had at least this level of context. We explained to them that secure messaging apps were a class of instant messengers that encrypted communications via the app. We drew an explicit contrast to "sensitive information," such as financial information, and participants were asked to describe why they thought someone might want to encrypt daily communications by comparison.

While many participants did first think of sensitive, potentially damaging conversations such as those pertaining to cheating on a partner, there were a number of responses that saw potential use for privacy in more mundane settings, though they didn't personally envision such a use for themselves. This dialogue with Carol is an illustrative example of this sentiment:

"I: What I want to ask is: why do you think we might want to encrypt daily communication? For example, it's easy to see why you might want to encrypt financial or medical information, but why do you think someone might want to encrypt their daily communication?

P: Probably for illegal reasons.

I: Can you explain a little?

P: Well, if you're doing things that aren't necessarily legal, you don't want people knowing about it that shouldn't know about it or the government looking into your things.

I: Are there other use cases that you might imagine?

*P: Why a normal person wouldn't want people looking into their stuff, basically? Because it's none of their business, for one thing *laughs* You might have personal things going on too, like an affair or something like that, that you wouldn't want other people knowing. But normal people can't look into that, so it would be more like government or police.*

I: Is this something that you would ever do personally? Use something that encrypts your daily communication?

P: Personally? I doubt it.

I: Can you explain why you might not?

P: I don't have anything to hide or worry about."

4. DISCUSSION

The models we observed, and other remarks our participants made, have implications for the way encryption is presented to users, both in design and communication. We now discuss the issues we observed and include our suggestions for a way forward.

4.1 Encryption is access control with a symmetric model

From the simplest mental model we observed to the most detailed and complex, they all reduced to the same functional abstraction: restrictive access control. Structurally, participants varied in how they imagined the encryption process acted on the data it was meant to protect, but all of our participants believed that encryption served one purpose: preventing undesired access.

Simultaneously, participants' mental models of encryption coincide very nicely with symmetric encryption. They understood that a shared secret would be needed, and also struggled to imagine how key sharing could safely occur, which hints at the key distribution struggles of symmetric encryption. Moreover, while their *structural* models of what the shared secret actually was were inaccurate, this mistaken belief nevertheless carries similar *functional* properties. More specifically: (1) if you believe the shared secret to be the set of transformations itself, then compromise of the key is tantamount to compromise of the encrypted data, and (2) loss of the shared secret results in the loss of ability to decrypt encrypted data. Thus, while participants had a flawed model of what a symmetric key is, their ability to interact with one is likely unimpacted.

Relevantly, this strong correlation of existing mental models with a symmetric encryption model also implies that the asymmetric encryption model is non-intuitive. Because keys

in symmetric and asymmetric encryption fulfill such different roles, getting users to understand public and private keys—even from a functional perspective—seems an uphill battle, particularly because we have yet to find an appropriate real-world analogy for the role of public and private keys [17].

Consider, for example, the task of verifying keys (e.g., in an authentication ceremony) in secure messaging apps. In a standard access control model, authentication is a unidirectional process: the accessing party verifies to the mechanism; only one agent is active in this scenario. Verification in an asymmetric encryption model, such as the authentication ceremony, however, requires both sending and receiving parties to validate one another. Given an access-control abstraction for encryption, what does the process of verifying the *sending* party's key even mean in the context of *receiving* encrypted data? Lacking a model, when users are asked to perform verification, they are likely to fall back on ad-hoc, non-cryptographic methods done, as has been observed [18].

Recommendation

This common perception of encryption as access control can be useful in the right contexts. Because it was shared by all our participants, even those with the most simple mental models, it can serve as a lowest common denominator model off which to build, and is likely a useful and intuitive abstraction in certain use cases. For example, encryption of personal data at rest, such as that done by mobile devices, is a good fit for this functional model. Digital wallets, such as those commonly employed by cryptocurrency, are likely another use case that fits well with the access control model.

Because asymmetric encryption necessitates a functional abstraction so foreign to participants' existing models, the presentation of interaction mechanisms as "encryption" is perhaps an unwise approach in these contexts. Instead, perhaps the way forward is to present users with interaction abstractions altogether separate from encryption, with a focus on matching functional, rather than structural, models.

4.2 Grassroots adoption is also a public relations battle

Participants felt that encryption is meant to protect sensitive information. While, on the surface, this view isn't necessarily incorrect, the nuances of this belief have grave implications for the grassroots adoption of encryption because it suggests that the perceived utility of encryption is low. More specifically, from a security perspective, participants believed that companies already encrypt their sensitive customer data, such as financial information. With respect to privacy concerns, the personal use of encryption in contexts such as daily communication, by contrast, is viewed negatively because it is believed that such data would only be perceived as sensitive if the user were engaging in either illicit or immoral activity, or were "paranoid" about the value of their data.

Thus, in scenarios where encryption is seen as having value, using it is seen as the responsibility of someone else, whereas in scenarios where using encryption would fall upon them personally, its use is perceived as improper. This has serious implications for adoption: if users do not perceive encryption as having utility—or worse, see its use as stigmatized—then they are unlikely to make proactive effort to adopt encryption software even *if* usability concerns could all be resolved.

Recommendation

If our participants' responses generalize to larger populations, then it suggests that improving the usability of encryption is likely not enough: improved risk communication will also be necessary. That is to say, improving the *how* of encryption is unlikely to alone resolve adoption issues; we must also focus on the *why*. Users can perhaps be helped to understand that there are benefits deriving from their individual use of encryption; even if not personally, but then perhaps to society as a whole. One of our participants, for example, described how the personal use of encryption might make sense in a different cultural context:

Lloyd: *"So keeping all that stuff that's very personal to yourself is probably good both from a 'keeping personal stuff personal' sort of way and—although this is a little paranoid in itself, and although this isn't a big deal right now—but if in ten years, that person's engaging in civil unrest, and that information's out there, that person can be threatened indirectly. It sounds really paranoid but that sort of shit happens in China all the time, you know? [...] that sort of stuff happens when governments have the ability to straight up read all the data and you have that kind of oppressive government going on."*

4.3 Confusion about encryption strength

Participants' perceptions about what it would take to undo encryption varied greatly, even among participants with similar mental models, and even among participants' individual responses themselves.

One potential cause for this is that factors external to their mental model seem to have influenced participants' beliefs. One participant was aware of the FBI's inability to break the encryption on the iPhone of a suspect, and so decided that encryption was therefore very strong. Another participant explained that she had seen encryption in popular media and it had always been broken, leading her to conclude that *"[i]t obviously can be done pretty easily."* Other participants noted the existence of data breaches, which, in combination with their belief that companies routinely encrypt data, signified to them that encryption is regularly broken by criminals.

Participants' responses also made it evident that beliefs about the strength of encryption—and by extension, its ability to protect their data—appeared to be focused more on the capabilities of attackers than it was on the fundamentals of the technology itself. In other words, even if it takes incredible resources to break encryption, that doesn't mean anything if an attacker *has* those resources. For example, Fred assumed that encryption would take "years" to undo without a quantum computer, which would instead need just "seconds." However, because he believed that the NSA does possess quantum computers, he believed encryption to be rather fragile as far as they were concerned.

Recommendation

It seems that if we wish users to understand the protective capabilities that encryption can offer them, we must convey its strength specifically within the context of the capabilities of likely attackers. We echo the sentiments made by Wash in his mental modeling effort [20]: *"without an understanding of threats, home computer users intentionally choose to ignore advice that they don't believe will help them. Security edu-*

cation efforts should focus not only on recommending what actions to take, but also emphasize why those actions are necessary."

4.4 Learning how encryption works doesn't help

Given that the mental models we describe seem to be flawed or incomplete, one natural reaction might be to assume that the proper course of action is to simply teach users how encryption really works. Research has shown, however, that "correcting" existing mental models can be a difficult task: "one cannot merely present people who hold an incorrect understanding with the correct information" [19]. Indeed, "the 'broadcast of facts' approach has been discredited by experts in safety risk communications" [16].

In our study, two participants had been exposed to the detailed mechanics of encryption previously, and yet still evidenced confusion. Clark, having learned of the WannaCry ransomware attack when it made the news, had attempted to learn something about how encryption worked, including watching a video on "how AES-256 encrypts stuff." This glance into the inner workings of encryption had nevertheless failed to fully impress itself on him, and he explained that all he knew was that *"[i]t's scrambling the, uh, the message. Uh– By doing a lot of math. I don't know much beyond that."* Brent described how his girlfriend was well versed in security matters, and had once taught him about encryption. For that reason, he remembered the terms "public" and "private" key, although he remembered neither their function nor their purpose.

Recommendation

Rather than relying on attempts to imbue users with an accurate model of how encryption works as the path to usable encryption, we should make efforts to align designs and communication efforts with the functional models users already possess.

4.5 Consider the context

The way security indicators are interpreted is dependent on users' perception of what threats exist within the respective context. The HTTPS/TLS browser indicators (e.g., the lock icon) were very effective in the sense that our participants had all noticed their existence, and a couple of participants had even clicked them for more information. However, their interpretation of what these indicators meant is worryingly inaccurate.

Participants mostly lacked a model of the man-in-the-middle as a potential threat, and thus when presented with security indicators, believed them to be representative of site security and not connection security. Those who had clicked through and were aware of the existence of certificates similarly misinterpreted their meaning, believing they meant that a certain site had been "certified" by a third-party. As these browser indicators are not at all a direct measure of site security, but rather indicative of an encrypted TLS connection between the user's browser and the web server, one could very well have a secure connection to an unsafe website, as Lloyd suddenly realized during his interview. *"Oh, really; that's what that means. I shouldn't do that then! Because it could be an* ***encrypted*** *way to send my password to hackers!"*

Recommendation

Security indicators must be carefully designed, with an aim of not just being noticed and trusted, but also with an eye to construct validity. That is, we must take caution to ensure not only that users understand indicators, but that they do not *misunderstand* them.

5. LIMITATIONS

Our study carries with it several limitations that are a direct consequence of our sample and sampling method. First, our sample consisted entirely of residents of the United States, and results may be subject to cultural effects. Similarly, our participant recruitment requirements necessitated English speakers; it is conceivable that this would have strengthened cultural effects, if any. Finally, our sample skewed heavily male, and it is possible that this had an effect on our findings, though we did not observe any notable differences between the models of male and female respondents.

The data collected was self-reported, qualitative in nature, and subjected to a coding process. Our findings, as presented, are thus subject to interpretation, and it is entirely possible that other researchers may come to different conclusions. For this reason, we have made our data publicly available.

Additionally, while we did continue our interviews until a perceived data saturation point, it must be acknowledged that our sample size falls on the low end. It is possible that with additional interviews, we would have observed additional behavior. However, because we were exploring a topic for which people's perceptions are fairly shallow due to limited exposure, and because participants' mental models were already very similar, we do not believe it likely that we would find any substantially different results with more interviews.

Finally, as explained to our participants before each interview began, encryption is a technical topic that our participants were largely unfamiliar with, and their mental models were unlikely to have been very developed beforehand. For example, when we asked Eva how decryption might occur, she proudly declared, *"I never thought about that! You don't."*

Indeed, it was often evident that participants' models were evolving during the interview itself, as they considered issues that had not previously occurred to them; this is in contrast to investigating mental models of systems that users frequently interact with, which guide existing behavior. In one explicit example of this occurring, when Sam was asked to explain what a "key" was (a term he had used unprompted), he changed his mind mid-sentence: *"The key is kind of like the schema of the code. [...] The algorithm so to speak— No, it's not the algorithm. The algorithm is what actually does the encryption, but it has to function with a certain pattern, and the key is like the file that knows exactly what pattern that the encryption then does."*

Because participants had to think through issues such as this as part of the interview, and because such thought was typically prompted by questions from the interviewer, it is possible that mental models of encryption in practice are more shallow than as presented here.

6. RELATED WORK

Previous research in this space, as it relates to our study, can largely be divided into two categories: (1) studies concerned

with the usability of encryption and (2) mental modeling efforts in the computer security space.

6.1 Encryption usability

Beginning with the seminal work, "Why Johnny Can't Encrypt" [23], now published almost 20 years ago, many studies have documented the long history of usability struggles in the encryption domain. As our study is directed at perceptions of encryption, however, we focus on those that reveal difficulties of understanding and not use.

Whitten and Tygar's classic work [23] tested how standard usability design principles hold up in the domain of computer security by evaluating PGP 5.0, an example of encryption software that was (at the time) considered to meet principles of good design. In addition to a number of interface design flaws discovered via a cognitive walkthrough, a user test found that participants struggled to execute a simple encryption task, simultaneously evincing confusion about core concepts such as the public key model. Tong et al. [17] examined in detail one of the criticisms posed by Whitten and Tygar, evaluating the metaphors used to communicate public key cryptography concepts. They developed a new set of metaphors that focused on the actions involved in the encryption process instead of the cryptographic primitives at work. Testing revealed that these new metaphors improved the efficacy of communication and improved the user-experience. These studies reveal that the mismatch between developers' and users' models of how to use encryption tools is a major cause of usability problems, underlining the importance of first understanding how users think when designing security tools.

Two more recent studies have made some progress in understanding mental models of encryption, although only as part of an investigation of broader topics. Abu-salma et al. [1] studied mental models of secure communication as part of a larger effort to understand obstacles to the adoption of such tools. They found frequent misunderstandings, several of which reinforce our own. First, some of their participants conflated authentication and encryption, which is captured by our access control model. They also found that participants often equated encryption with some sort of data encoding or scrambling process, which we frequently observed in our interviews. Probing participants' understanding of the differences between end-to-end encryption and client-server encryption, they found nearly no one who could distinguish between them. This coincides with our findings that participants' perception of encryption was nearly unilaterally that of encrypting data at rest, with the exception of a few informed individuals had some knowledge of HTTPS.

Ruoti et al. [13] examined the risk perceptions and security behaviors of a number of suburban adults. As part of their effort, they asked participants if they had heard of encryption before. They reported that most of their participants understood the basic principles of symmetric key cryptography, recognizing that "encryption relies on a shared key." By contrast, our more focused exploration of this issue revealed that while participants indeed had some notion of a shared secret in its broadest sense, their conception of what a shared secret might be differs quite drastically from the traditional cryptographic sense. More specifically, instead of an input to an encryption algorithm, many of our participants believed that the set of operations performed during the encryption process were themselves the shared secret.

6.2 Mental modeling in computer security

While not directly relevant to our study, similar efforts have been to explore mental models in a computer security domain, their application in computer security having been previously encouraged [4, 9, 16].

Volkamer and Renaud [19] described the concept and its role in computer security. Importantly, they characterized the methods for exploring mental models, such as think-aloud and diagramming. Wash [20] presented findings of users' mental models for home security, identifying eight "folk models" that users rely on to guide them in making security decisions. (This was later replicated with a German population, by Kauer et al. [8].) His approach for synthesizing interview data into discrete mental models served as a guide for the process we followed in this work. Bravo-Lillo et al. [3] demonstrated how mental modeling can be used to understand how novice users and advanced users interpret and react differently to security warnings. They presented participants with a series of warnings and asked questions about their perceptions and thought process. Kang et al. [7] explored mental models of the Internet and how that affects user perception of privacy and security and included a diagramming portion as a central element of their work. As mentioned previously, we employed both a question-and-answer process (similar to Wash and Bravo-Lillo et al.) as well as a diagramming exercise (like that used by Kang et al.).

7. CONCLUSION

In this paper, we present our findings drawn from 19 semi-structured interviews with U.S. participants about their perceptions of encryption. Our focused effort to explore these perceptions sheds additional light on, and adds nuance to, existing research that touched upon aspects of this problem. We identify four mental models of encryption, of varying levels of detail and complexity, that convey functional abstractions of access control and the mechanics of a symmetric encryption model. We highlight concerns about the current state of how encryption is presented to users and how they are expected to interact with encryption tools.

Perhaps because our study focuses on perceptions of encryption divorced from implementation, we find an urgent need for improved risk communication efforts regarding encryption. Notably, we must make greater attempts to contextualize *why* participants should use encryption in a manner that takes into consideration their threat models. Similarly, we should strive to frame the functional aspects of encryption in a form that matches the intuitive models users possess, regardless of their technical accuracy.

8. ACKNOWLEDGMENTS

We would like to express our appreciation for Rick Wash and Emilee Rader for their guidance in the early stages of this work. We also thank the anonymous reviewers for their helpful feedback. This research is sponsored by the Department of Homeland Security (DHS) Science and Technology Directorate, Cyber Security Division (DHS S&T/CSD) via contract number HHSP233201600046C.

9. REFERENCES

[1] R. Abu-Salma, M. A. Sasse, J. Bonneau, A. Danilova, A. Naiakshina, and M. Smith. Obstacles to the adoption of secure communication tools. In *IEEE Symposium on Security and Privacy (SP 2017)*, pages 137–153. IEEE, 2017.

[2] D. Akhawe and A. P. Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *USENIX Security Symposium 2013*, volume 13. USENIX Association, 2013.

[3] C. Bravo-Lillo, L. F. Cranor, J. Downs, and S. Komanduri. Bridging the gap in computer security warnings: A mental model approach. *IEEE Symposium on Security and Privacy (SP 2011)*, 9(2), 2011.

[4] L. J. Camp. Mental models of privacy and security. *IEEE Technology and Society Magazine*, 28(3), 2009.

[5] S. Clark, T. Goodspeed, P. Metzger, Z. Wasserman, K. Xu, and M. Blaze. Why (Special Agent) Johnny (still) can't encrypt: A security analysis of the APCO project 25 two-way radio system. In *USENIX Security Symposium 2011*, pages 8–12, 2011.

[6] W. K. Edwards, E. S. Poole, and J. Stoll. Security automation considered harmful? In *New Security Paradigms Workshop (NSPW 2008)*. ACM, 2008.

[7] R. Kang, L. Dabbish, N. Fruchter, and S. 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 2015)*. USENIX Association, 2015.

[8] M. Kauer, S. Günther, D. Storck, and M. Volkamer. A comparison of american and german folk models of home computer security. In *International Conference on Human Aspects of Information Security, Privacy, and Trust (HAS 2013)*, pages 100–109. Springer, 2013.

[9] K. Renaud, M. Volkamer, and A. Renkema-Padmos. Why doesn't Jane protect her privacy? In *International Symposium on Privacy Enhancing Technologies Symposium (PETS 2014)*, pages 244–262. Springer, 2014.

[10] S. Ruoti, J. Andersen, S. Heidbrink, M. O'Neill, E. Vaziripour, J. Wu, D. Zappala, and K. Seamons. We're on the same page: A usability study of secure email using pairs of novice users. In *SIGCHI Conference on Human Factors in Computing Systems (CHI 2016)*, pages 4298–4308. ACM, 2016.

[11] S. Ruoti, J. Andersen, D. Zappala, and K. Seamons. Why Johnny still, still can't encrypt: Evaluating the usability of a modern PGP client. *arXiv preprint arXiv:1510.08555*, 2015.

[12] S. Ruoti, N. Kim, B. Burgon, T. Van Der Horst, and K. Seamons. Confused Johnny: when automatic encryption leads to confusion and mistakes. In *Symposium on Usable Privacy and Security (SOUPS 2013)*. USENIX Association, 2013.

[13] S. Ruoti, T. Monson, J. Wu, D. Zappala, and K. Seamons. Weighing context and trade-offs: How suburban adults selected their online security posture. In *Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 211–228. USENIX Association, 2017.

[14] S. Schröder, M. Huber, D. Wind, and C. Rottermanner. When SIGNAL hits the fan: On the usability and security of state-of-the-art secure mobile messaging. In *European Workshop on Usable Security (EuroUSEC 2016)*, 2016.

[15] S. Sheng, L. Broderick, C. A. Koranda, and J. J. Hyland. Why Johnny still can't encrypt: evaluating the usability of email encryption software. In *Symposium On Usable Privacy and Security (SOUPS 2006)*, pages 3–4. USENIX Association, 2006.

[16] G. Stewart and D. Lacey. Death by a thousand facts: Criticising the technocratic approach to information security awareness. *Information Management & Computer Security*, 20(1):29–38, 2012.

[17] W. Tong, S. Gold, S. Gichohi, M. Roman, and J. Frankle. Why King George III can encrypt. http://randomwalker.info/teaching/spring-2014-privacy-technologies/king-george-iii-encrypt.pdf, 2014. Accessed: 7 February, 2018.

[18] E. Vaziripour, J. Wu, M. O'Neill, R. Clinton, J. Whitehead, S. Heidbrink, K. Seamons, and D. Zappala. Is that you, Alice? a usability study of the authentication ceremony of secure messaging applications. In *Symposium on Usable Privacy and Security (SOUPS 2017)*, 2017.

[19] M. Volkamer and K. Renaud. Mental models - general introduction and review of their application to human-centred security. In *Number Theory and Cryptography*. Springer, 2013.

[20] R. Wash. Folk models of home computer security. In *Symposium on Usable Privacy and Security (SOUPS 2010)*. USENIX Association, 2010.

[21] R. Wash and E. Rader. Influencing mental models of security: a research agenda. In *New Security Paradigms Workshop (NSPW 2011)*. ACM, 2011.

[22] R. Wash, E. Rader, K. Vaniea, and M. Rizor. Out of the loop: How automated software updates cause unintended security consequences. In *Symposium on Usable Privacy and Security (SOUPS 2014)*, pages 89–104. ACM, 2014.

[23] A. Whitten and J. D. Tygar. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In *USENIX Security Symposium 1999*, volume 1999. USENIX Association, 1999.

Appendix

Interview introduction

1. **Before we start, I just wanted to say that what we're going to talk about today is likely to be a subject you're not very familiar with—so please don't worry, this isn't a test. Instead, I'm interested in hearing what you think and feel, so don't be worried if you don't think you know the answer to a question. If that happens, just take your best guess. Also, if you're ever confused by a question I'm asking, please let me know, and I'll try to explain or rephrase.**
2. **If there are no other questions, let's get started.**
3. **To begin, I'd like to get a general sense for your background, and so I'd like to ask: what do you do for a living?**
4. **Now I'd like to get a sense for your computing environment. Can I ask what types of devices you own: laptop, desktop, smartphone, etc.? What types of things do you do with them on a regular basis?**

Existing thoughts on encryption

1. **Now I'd like to turn to the topic for today. My first question is: what comes to mind when I say the word "encryption"?**
   - **What sorts of imagery do you picture in your head when I say that word?**
   - **Where have you seen or heard that term before?**
2. **Now I'd like to begin the diagramming task I mentioned in the email. Before we start, I'd like to remind you that this isn't a test of your artistic ability; this is just to help me get a better sense for how you imagine things work.**
3. **I'd asked you to prepare a pen and paper for this: do you have them ready?**
4. **Great. Now, on your piece of paper, please write the sentence, "This is a message to be encrypted." Now, what I want you to do is imagine you're going to encrypt this sentence, and draw whatever you think that looks like. Take as much time as you need and let me know when you're done.**
5. **Next, could you just draw a simple little picture for me? It can be anything, like a cloud, tree, stick figure—anything. Now, I want you to do the same thing you just did, but with the picture. Imagine you're going to encrypt this picture, and draw whatever you think that looks like. Again, take as much time as you need and let me know when you're done.**
6. **Could you take a picture of the drawing with your phone and text or email that to me? Thanks.**
7. **[Once the email with their picture arrives...] Great, I got it. Could you walk me through what you drew?**
8. **Okay, so my next question is: imagine you're going to send an encrypted message to a friend or family member and what they get is just the encrypted part. What would they have to do to read the original message?**

Examples of encryption

1. **Okay, were going to change gears a bit now from what encryption is and how it works to how it gets used.**
2. **First, what role do you think encryption plays in your life?**

3. **What about individuals? Do you think there are people that use encryption on a personal basis?**
4. **Now, I'm going to introduce a few examples of places where encryption does exist. Again, it's entirely fine if you've never heard of any of these before: I'm interested in hearing your impressions anyway, so please make your best guess.**
5. **[Smartphone encryption]: The first example is smartphone encryption. Both iOS—if you have an Apple device—and Android allow you to encrypt your smartphone. Now, my first question is: what do you think it even means to "encrypt" your smartphone?**
    - **Why do you think this function exists? What do you think encryption is supposed to protect?**
    - **Who do you think encryption would be protecting your phone from?**
6. **[Web encryption/HTTPS]: The next example we're going to discuss is encryption of data that goes out over the Internet. Have you ever noticed an "HTTPS" or little lock icon near your address bar when using a browser? What do you think it means?**
7. **So what it means is that the data going between your browser and the web server is encrypted.**
    - **What do you think encryption is protecting in this case?**
    - **Who do you think it's protecting you from?**
8. **[Secure messengers]: Now I want to talk about secure messaging apps. Have you ever used an instant messenger like WhatsApp, Facebook Messenger, Signal, or Telegram?**
    - **So it makes sense why you'd want to encrypt sensitive information like financial information. But why do you think someone might want to encrypt their daily communications?**
    - **Who do you think you'd be protecting your communications from by encrypting them?**

"It's Scary...It's Confusing...It's Dull": How Cybersecurity Advocates Overcome Negative Perceptions of Security

Julie M. Haney
University of Maryland, Baltimore County
1000 Hilltop Circle
Baltimore, MD, USA
jhaney1@umbc.edu

Wayne G. Lutters
University of Maryland, Baltimore County
1000 Hilltop Circle
Baltimore, MD, USA
lutters@umbc.edu

ABSTRACT

Cyber attacks are on the rise, but individuals and organizations fail to implement basic security practices and technologies. Cybersecurity advocates are security professionals who encourage and facilitate the adoption of these best practices. To be successful, they must motivate their audiences to engage in beneficial security behaviors, often first by overcoming negative perceptions that security is scary, confusing, and dull. However, there has been little prior research to explore how they do so. To address this gap, we conducted an interview study of 28 cybersecurity advocates from industry, higher education, government, and non-profits. Findings reveal that advocates must first establish trust with their audience and address concerns by being honest about risks while striving to be empowering. They address confusion by establishing common ground between security experts and non-experts, educating, providing practical recommendations, and promoting usable security solutions. Finally, to overcome perceptions that security is uninteresting, advocates incentivize behaviors and employ engaging communication techniques via multiple communication channels. This research provides insight into real-world security advocacy techniques in a variety of contexts, permitting an investigation into how advocates leverage general risk communication practices and where they have security-specific innovations. These practices may then inform the design of security interfaces and training. The research also suggests the value of establishing cybersecurity advocacy as a new work role within the security field.

1. INTRODUCTION

> *"From the audience's perspective, security can be characterized by three major factors: one, it's scary; two, it's confusing; three, it's dull" (P08, security consultant).*

On a regular basis, the news is filled with reports of cybersecurity attacks [27, 48, 50], with companies, government agencies, and individuals being exploited at an alarming pace [45, 47]. Despite real and evolving cyber threats, users are falling behind in defending their systems and networks. They often fail to implement and effectively use basic cybersecurity practices and technologies, due in part to negative feelings about security.

Cybersecurity advocates are security professionals who attempt to remedy implementation failures by actively encouraging and facilitating the adoption of security best practices. "Cybersecurity advocate" is an emerging term-of-art among practitioners, with few holding it as their official job title. Indeed, many perform advocacy tasks in parallel with other responsibilities. They promote security to a variety of individuals, including home users, office workers, students, faculty, technical staff, developers, and executives. Examples of cybersecurity advocates include: organizational security awareness professionals; secure development champions; security consultants; and non-profit staff who publish resources to aid others in securing their digital assets. Regardless of the scope, advocacy is instrumental to their professional success. To be effective, these advocates must motivate people to engage in beneficial security behaviors, which often necessitates overcoming negative perceptions.

Prior research studies have investigated user perceptions of security and intentions toward following security practices. This body of work reveals incomplete, inaccurate mental models and a variety of sociotechnical factors that influence people's decisions to implement security solutions (e.g., [15, 19,35,49]). However, no research has been done to explore this problem space from the perspective of those actually doing the influencing, such as cybersecurity advocates.

To address this gap, we interviewed 28 self-identified cybersecurity advocates from industry, higher education, government, and non-profits. This paper presents a subset of findings from this larger study. Here we focus on answering the following research questions: 1) What are the professional characteristics and skills that security advocates employ in their work? and 2) What techniques do security advocates use to encourage security adoption?

The findings reveal ways in which advocates attempt to overcome users' widely-held negative views of security. We found that, as a foundation, advocates must first establish trust with their audience. To overcome perceptions of security being fear-invoking, advocates are honest, yet discerning, about the risks they communicate. They also attempt to empower their audience by engendering a feeling of hope and self-efficacy. Advocates address feelings that security is con-

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

fusing, complex, and difficult by "bridging the gap" between security experts and non-experts. They do this by serving as security educators who promote recommendations that can be realistically accomplished with usable security solutions. Finally, to overcome perceptions that security is irrelevant and boring, advocates create interest by incentivizing and employing engaging rhetorical techniques.

Our research has several contributions. Foremost, it identifies the cybersecurity advocacy role and its evolving definitional boundaries. It also provides insight into real-world security advocacy techniques in a variety of contexts, including how advocates leverage general risk communication practices and where they have security specific innovations. These practices may then inform the design of security interfaces and training resources. Additionally, the research suggests that there is value in establishing cybersecurity advocacy as a new work role within the security field, and provides a foundation for the recommended attributes of those who might perform it.

2. RELATED WORK

In this section, we lay the foundation for our study and its implications by summarizing past research in risk communication and persuasion within the security context.

2.1 Risk Communication

To form a basis for the skills, knowledge, and abilities necessary for cybersecurity advocacy, it is helpful to look at the literature on risk communication. Although much of this work has been conducted outside the cybersecurity context in fields such as health, environmental hazards, disaster planning, and home security, there can be much to learn about strategies to effectively communicate risk.

Kasperson [20] found that risk communicators aim to develop trust, create awareness strategies, facilitate understanding of concepts, employ mediating skills, and motivate people to act. Rowan [38] observed that risk communication can be controversial because it involves threatening and poorly understood concepts that can invoke hostile feelings towards the communicator. Therefore, they must be able to diffuse negative feelings so as not to erode trust, reframe negative messages into positive ones when appropriate, and use negotiation skills. Since a foundational aspect of risk communication is the establishment of trust and credibility, communicators need to exhibit empathy, honesty, openness, listening skills, and commitment [9, 42]. Trustworthy risk communicators serve as the bridge between technical experts and non-experts [14]. This bridging is akin to establishing common ground, which is the mutual knowledge, beliefs, and assumptions that are believed to be essential for successful communication between people [7].

Risk communication is a learned competency that includes a variety of approaches, including: keeping communications simple, but specific and unambiguous; customizing information to target audiences; assisting people in seeing the consequences of their decisions; providing clear and precise directions for action; building self-efficacy; and presenting information in an engaging manner [9, 11, 26, 39].

The risk communication literature begins to form a picture of what is required to be effective. However, little research has been done to investigate whether these characteristics, approaches, and goals are relevant for advocates within the security realm.

2.2 Influencing Security Behavior

Understanding what motivates people to change their behavior and practice good security is essential to evaluating whether advocates' approaches address these motivations.

2.2.1 Perceptions of Security

Before determining the most effective way to persuade people to practice good security, there needs to be an underlying understanding of their perceptions of security. Numerous studies have explored these perceptions, mostly among non-technical users. Huang et al. [16] conducted a survey of over 600 individuals that revealed influential factors, including knowledge, impact, controllability, and awareness of exposure to a threat. Furnell and Thomson [12] and Stanton et al. [44] discussed "security fatigue," a weariness towards security when it becomes too burdensome. From an organizational perspective, Post and Kagan [31] found that employees view stringent security measures as counterproductive since it impedes their ability to be flexible in their day-to-day operations.

A set of researchers explored mental models of security, with some examining the general public's often incomplete and inaccurate mental models and how these perpetuate poor security practices [19, 32, 49]. Other researchers shed light on the differences in mental models of security experts and non-experts [17,30]. Bravo-Lillo et al. [5] and Raja et al. [33] examined mental models while applying risk communication principles to security warnings. Camp [6] discussed how mental models of physical security, medical infections, criminal behavior, economics failure, and warfare might be applied to communicate cybersecurity risk. Zhang-Kennedy et al. [51] extended these models, suggesting that the use of surveillance and medical metaphors within infographics and a comic resulted in better security learning. However, Brase et al. [4] investigated the impact of Camp's suggested models in cybersecurity situations and found that there was little indication that any of these resulted in significantly better outcomes.

2.2.2 Persuasive Techniques

Protection Motivation Theory (PMT) [22, 37] claims that risk behavior is based on a cost-benefit analysis in which a threat appraisal (severity, likelihood, rewards/consequences) is weighed against a coping appraisal (response cost, effectiveness of response, self-efficacy). Sommestad et al. [43] sought to determine whether the PMT held true in the information security domain and found that it did explain security behavior if the threat and coping mechanism were concrete and when the threat was personally relatable.

Several studies applied PMT to explore the effectiveness of fear appeals in changing security behaviors within organizations. Johnston and Warkentin [18] suggested that, because people naively think that bad things will not happen to them, fear appeals should emphasize the likelihood of an occurrence by using concrete examples of negative consequences related to a threat. Herath [15] found that both intrinsic (e.g., perceived effectiveness, contribution to the greater good) and extrinsic (e.g., social pressures, penalties) motivators influenced security behaviors. However, the

severity of penalty approach has a negative impact because penalties are often inconsistently applied or may generate hostilities.

Additional efforts investigated approaches for influencing security behavior change among employees. Albrechtsen and Hovden [1] found that small group workshops were more effective at changing security behaviors than mass communications. Siponen [41] suggested that security awareness programs should include reasons for why people should follow security guidelines and engender feelings of wellbeing, rationality, and logic. Other efforts examined similar techniques from a home user perspective. Rhee et al. [35] discovered that the threat of negative consequences has limited impact on decisions to implement security, whereas users with higher feelings of security self-efficacy were more likely to engage in positive behaviors. In a study on the adoption of security technologies, Shropshire et al. [40] found that negative framing (presenting outcomes in loss terms) is better suited for detection technologies (e.g., virus scanners, firewalls) than for prevention technologies (e.g., password settings, access controls). Redmiles et al. [34] investigated why people choose to accept security advice, discovering that advice sources were evaluated based on perceived trustworthiness, and that fictional narratives with relatable characters may be effective for teaching security concepts.

Although much literature has focused on persuasion in security, little research examined this topic from the viewpoint of those attempting to do the persuading. This paper seeks to understand their expert craft and how they appropriate these techniques and creatively respond to the particular context at hand. Ultimately this reveals the art of effective security advocacy. To best illuminate these practices, we had to deeply engage with expert advocates.

3. METHODOLOGY

Over a nine-month period, we conducted semi-structured interviews of cybersecurity professionals performing advocacy tasks as a major component of their jobs. We chose semi-structured interviews over other methods, such as surveys, because of the richness of data afforded, the latitude to ask follow-up questions to clarify or delve deeper into participant responses, and the ability to encourage participants to add other relevant information not explicitly targeted [8].

Our institutional review board approved the project. Prior to the interviews, participants were informed of the purpose of the study and how their data would be used and protected. Participants then signed a consent-to-participate form, also indicating whether they would allow audio recording of the interview (two declined). All interviews were transcribed from the audio recordings or field notes and stored without personal identifiers. Interviewees were not compensated for their participation.

3.1 Recruitment

Our conceptualization of an advocate originated from field observations on how this group of professionals described themselves. Therefore, we initially recruited from researcher contacts and internet searches those who self-identified as security advocates. We then were open to snowballing recommendations that allowed interviewees to identify others like themselves. Our definitional boundary of the cybersecurity advocate role continued to take shape and guided our subsequent recruitment as the interviews progressed. To ensure that a broad range of security advocacy contexts would be included in the study, we purposefully selected individuals who performed different types of security advocacy, for example, security awareness training, public campaigns, advocacy for a particular community, or security consultation. Additionally, we sampled advocates working in a variety of organizational types, including government, industry, higher education, and non-profits, to account for different viewpoints that may be inherent in each of these sectors. This yielded a collection of information-rich cases [28].

We employed theoretical sampling throughout data collection to guide recruitment [8]. Following this approach, we recruited participants four or five at a time. The next group of potential participants was then purposely chosen to include those who might be able to provide more insight on concepts or areas of interest emerging from the analysis of the preceding set. For example, when several participants raised gender diversity concerns in the security field, we subsequently made an effort to recruit additional female participants to gain their perspectives.

3.2 Data Collection

We conducted 28 semi-structured interviews lasting on average 45 minutes. If logistically feasible, interviews were face-to-face (12 interviews). Otherwise, participants were given the option of a phone (9) or video conference (7) interview.

The first three interviews were pilots to discover potential flow and timing issues. Because there were only minor revisions to the protocol following, data from these interviews are included in the final data set. In line with accepted qualitative research methods, we interviewed until we reached theoretical saturation, the point at which no new themes or ideas emerged from the data [23].

Interview questions addressed several areas: work practices, professional motivations and challenges, characteristics of successful advocacy, and how participants stay up-to-date on security happenings. The interview protocol is included in the appendix. Separate from the interview, participants also completed a short, online demographic survey that collected information about years of experience in the field, current position, sectors in which they had worked, and education. One participant did not complete the survey.

3.3 Analysis

We conducted iterative, inductive coding and analysis on the data. This commonly used qualitative research approach allows for an organic emergence of core concepts, starting with the categorization of the data into initial codes and then progressing to the recognition of relationships among those codes [13]. We began preliminary analysis at the onset of data collection to assess the quality of data and themes arising from the interviews. This allowed for small adjustments in the interview protocol over time as some questions reached saturation or when new themes started to arise as part of theoretical sampling. Throughout this process, we also engaged in axial coding to link related codes together (demonstrated by the subsections in section 5), wrote analytic memos, and identified core concepts. We regularly met to discuss emerging themes and our interpretations.

At the conclusion of data collection, both researchers began

construction of a final codebook. We reviewed five interviews (2,482 lines) individually and performed open coding to label, look for meaning, and begin to categorize the data. We then met multiple times to discuss identified concepts in those interviews. These discussions led to the development of the codebook. The first author then used the codebook to deductively code the remaining interviews.

4. PARTICIPANT DEMOGRAPHICS

We interviewed 10 female and 18 male professionals, clustered in age from 25-34 (3 participants), 35-44 (7), 45-54 (7), and 55+ (10), with one undisclosed. Overall, they were a veteran group, with all but six having more than 10 years experience in the security field, and the rest having at least five years. Table 1 summarizes participant demographics. Some details are generalized to protect confidentiality.

The participants had diverse educational and career backgrounds. Interestingly, 14 participants had at least one degree in non-technical fields as diverse as public policy, communication, history, law, business, English, and graphic design. Participants had worked in a variety of government, private industry, higher education, and non-profit organizations, with most having experience in more than one of these sectors. When asked to describe their target audience, 10 said their audience was mainly external to their organization, three mainly focus within their organization, and 15 said they advocate both externally and internally. Their diverse audiences included the general public, co-workers, professional communities, government organizations, students and faculty, policy makers, corporate boards, developers, and other security professionals. The advocates performed a number of functions, several having more than one. Some were security engineers, led organizational security awareness programs, or served as security consultants. Others were security educators, non-profit organizers, researchers, or secure development experts.

5. FINDINGS

In this paper, we focus on how advocates attempt to overcome negative perceptions that security is scary, confusing, and dull. An overview of our framework is provided in Fig. 1. We first discuss a prerequisite condition for successfully overcoming negative perceptions: the advocate's ability to be viewed as a trustworthy information source. Subsequent sections begin with a description of each underlying negative perception reported in security advocates' audiences. Subsections describe strategies that advocates employ to attempt to overcome the perception. Note that strategies gleaned from the interviews are based on participant *perceptions* of effective advocacy strategies.

Counts of participants who mentioned a concept are provided throughout this section. However, due to the semi-structured format of the interviews, we caution the reader against making quantitative inferences beyond frequency. Counts are reported to add weight to concepts that were repeatedly mentioned throughout the interviews, but the significance of an insight may not be determined solely by the number of participants voicing it.

5.1 Establishing Trust

Before advocates can overcome negative perceptions of security, they must first establish trust, which is a foundational aspect of risk communication. A security engineer who provides consultation to government customers spoke of this trust: *"To me, trust is the most important thing that I have. If they trust that what I'm telling them and what I'm doing is the right thing, then I am much more successful" (P12).* Advocates gain audience trust by relying on organizational reputation, demonstrating technical knowledge, building relationships, and leveraging insider access.

5.1.1 Relying on Organization Reputation

As noted in four interviews, organizational reputation may help to establish credibility, at least initially. One participant suggested that the most effective advocates are sometimes *"people who have the credentials and are associated with organizations that are viewed as having some authority" (P07).* This credentialing can especially be helpful when advocating to the general public, especially online where personal interactions are rare. However, when interacting directly with an audience, organizational reputation only goes so far, and must ultimately be upheld on an individual basis. A government security analyst discussed this external bump versus sustained personal reputation: *"Our agency... carries with it a great deal of credibility... And I think that helps out a lot. But [individuals have] to be able to exhibit and illustrate the qualities that go along with the respect they bring into the door" (P01).*

5.1.2 Demonstrating Technical Knowledge

One way that advocates establish individual credibility is by demonstrating technical knowledge, as suggested as an important characteristic by 19 participants. One participant exclaimed, *"First and foremost, you really do need to understand the technology... This stuff's tricky, and you don't just guess your way out of it" (P08).* Advocates that work with technical staff are particularly held to high standards with respect to technical acumen. A participant with over 30 years in the security field emphasized this: *"This is a business that is very technology oriented, and full of people... who want to one-up you. So if you can't kind of deal with that, it's going to be hard for you to be an effective advocate because people will kind of eat you up" (P04).*

5.1.3 Building Relationships

Whereas technical skill may be an important component in building credibility and trust, our findings support previous risk communication research that emphasizes the importance of exercising interpersonal skills to build relationships and foster trust. A security usability specialist emphasized the value of these non-technical skills: *"If you're a computer scientist, and all you know is the computer science, and you don't have the empathy, you don't have the skills to listen,... you don't have that psychological side, I don't think you can make it work" (P03).*

Relationship building is facilitated by demonstrating empathy (mentioned by four participants) and listening skills (by six). A participant suggested, *"The most important part is to go in and listen... to what their challenges are, what their problems are" (P05).* A technical executive at a higher education institution expressed the importance of empathy:

> *"I think people have to have a high emotional intelligence and especially empathy. Part of being successful in this is being able to have a conversation and put*

Table 1: Participant Demographics

ID	Gen	Role	Sector	Edu	Audience	Audience Description
P01	M	Security analyst	***G***	T,N	B	tech staff, managers
P02	M	Professor	***E***,G,I	T,N	B	general public, students
P03	F	Computer scientist	***G***,I	T	B	tech staff, managers, general public
P04	M	Security evangelist	***N***,G	T	B	tech staff, managers
P05	M	Security researcher	***I***,G	T	B	tech staff, managers
P06	M	Director	***N***,G,E,I	N	B	public policy makers, managers
P07	F	Senior technologist	***G***,***E***,I	T	E	general public, managers
P08	M	Security consultant	***I***	N	E	non-tech professionals, managers
P09	M	Training director	***E***,G	N	E	tech staff
P10	M	Instructor, consultant	***I***,***E***,G	T	E	tech staff, managers
P11	M	Director	***N***,I	N	E	public policy makers, tech staff, managers
P12	M	Security engineer	***I***,***E***,G	T	E	tech staff, managers
P13	M	Security engineer	***I***	U	I	tech staff, managers
P14	M	Security awareness director	***E***,G	N	B	students, faculty, tech staff, managers
P15	F	Director	***N***,***E***,I	N	B	tech staff, managers
P16	M	Computer scientist	***G***,E,I	T,N	I	managers
P17	M	Researcher	***I***	T	E	developers, tech staff
P18	M	CIO	***E***	T	B	students, faculty, tech staff, managers
P19	F	Senior Architect	***I***	T	I	developers
P20	M	Professor	***E***,***G***	T	E	students, tech staff, managers
P21	F	Company co-founder	***I***,G	T	E	end users, tech staff, managers
P22	M	Security researcher	***I***, E	T	B	developers
P23	F	Security consultant	***I***,E	N	B	tech staff, general public
P24	F	Director	***N***	N	E	general public, tech staff, managers
P25	F	Deputy CIO	***G***,I	N	B	end users, tech staff, managers
P26	F	CISO	***G***,I	T	B	end users, tech staff, industry partners
P27	M	Director	***N***,I	N	B	tech staff, managers
P28	F	Security Awareness director	***I***,E	N	B	end users, tech staff, managers

Sector (***Current***,Past): E=Education, G=Government, I=Industry, N=Non-profit; **Edu (Education)**: T=Technical degree, N=Non-technical degree, U=unknown/not reported; **Audience**: I=Internal to own organization, E=External to own organization, B=Both internal and external

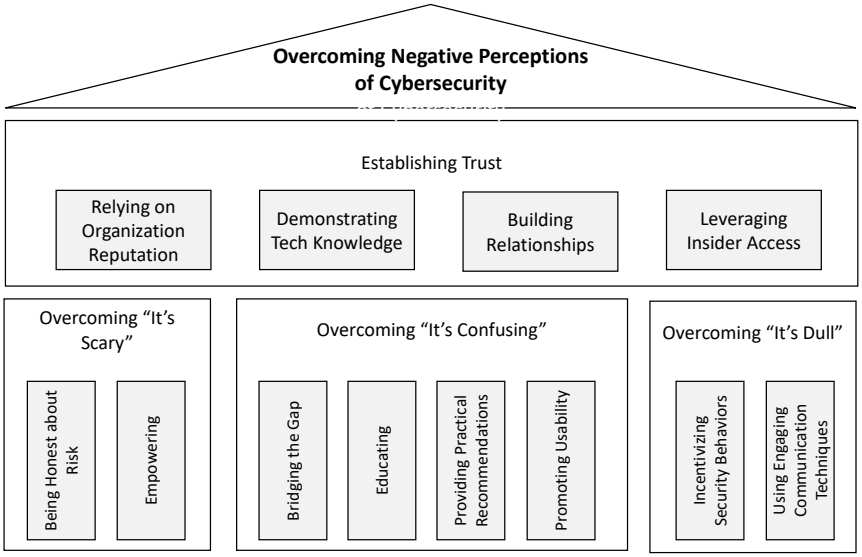


Figure 1: Framework of how cybersecurity advocates overcome negative perceptions of security

yourself in the place of the person that you're working with, and then be able to give effective advice that is not preaching, is trying to be helpful, and is letting them know that they're not stupid because they may not know how to do certain things" (P18).

Humility was mentioned by five participants as another interpersonal skill important for trust-building. Several noted that those advocates who approach a situation with an attitude of *"I'm in charge. I know best. You must listen to me" (P02)* are not generally very effective in enacting security behavior change because they put their audience on the defensive. A deputy CIO with a strong technical background remarked on the importance of not being arrogant because *"You'll never have all the answers" (P25).* A security consultant discussed his personal philosophy of humility: *"Whenever I walk in the room, I assume I'm the stupidest one there, and everything works out great" (P10).*

Trust is also created by being open to multiple viewpoints and building consensus. Consensus was especially important for the participants from non-profit organizations that relied on volunteers to inform their advocacy efforts. A founder of a non-profit group discussed their commitment to consensus building: *"We prioritize and cherish a multi-stakeholder approach. There [are] lots of voices... The goal is to surface beliefs, combine them with other beliefs, come to a set of shared beliefs" (P11).* Another participant described her collaborative role with members of her non-profit organization as *"an uber-facilitator. Our job is to get these people together and make them work for the common good" (P15).*

Interpersonal skills do not only apply to advocates who have in-person interactions with their audience; others must utilize these skills for any security guidance that reaches their audience. For example, P24's non-profit conducts extensive anonymous consumer research prior to publishing security guidance to ensure they address their audience's concerns and use language that will be easily understood. This attention to their audience's needs, in effect, demonstrates listening skills and empathy.

5.1.4 Leveraging Insider Access

Nine participants gained credibility due to their past experience in the professional communities to which they advocated. This experience helped them to be portrayed as "insiders." For example, one participant with a law background began her career in security advocacy when a legal organization recruited her to help with security compliance: *"They needed a translator to translate law to geek... And I learned that I sort of have a unique aptitude in this area where law and information security policy intersect" (P15).* Another participant remarked, *"It's very difficult to integrate yourself into someone else's daily work when you don't know what the daily work is" (P17).*

However, gaining credibility can be challenging when the advocate is perceived as an outsider. To overcome this, six mentioned the value of enlisting the support of opinion leaders and decision-makers within the target community. One participant talked about this value: *"You need to find whoever it is that you think is a change maker and make sure they have that data, that they're excited by that data, and they can use it to their benefit to make a difference" (P03).*

5.2 Overcoming "It's Scary"

The consequences of poor security can be catastrophic personally, organizationally, and societally. All participants had a solid understanding of the current state of security and potential consequences of poor security practices. One participant opined that the internet is *"getting more insecure constantly... The bad guys are getting better" (P06).* Another was concerned with global consequences, saying, *"It is so easy to imagine a really big cyber incident. And the barrier to entry is really, really low" (P16).*

Security risks are real, but several participants believed that, in some cases, these risks are sensationalized. Two participants partially blamed other security professionals, with one advocate noting, *"We're just really a fear-mongering industry" (P21).* Another who came into the security field with a humanities background observed security professionals *"tend to be really negative and really fatalistic. Everything's awful, everything's burning, everything's dead" (P23).*

Three participants also blamed media portrayals of security incidents for creating anxiety, particularly among non-technical audiences. A security consultant reflected that when people see depictions of cyber incidents on television and in the movies, *"the computer looks like some kind of magic box where somebody touches it, and zing! They attacked our network and taken our children, and look, they've wilted our lettuce!" (P08).* Another commented on how media portrayals can build fear around concepts that are unfamiliar: *"People are afraid of what they don't understand or don't want to learn... Their consciousness is kind of framed in this Hollywood... sort of approach where it's this evildoer. And that terrifies people" (P02).*

It is not surprising, then, that some people view cybersecurity as scary. To address this, advocates must strike a careful balance between being candid about security risks while being hopeful and encouraging. The latter are essential for developing a sense of empowerment in the audience.

5.2.1 Being Honest, Yet Discerning, About Risk

To convey a sense of importance and urgency to their audience, our participants said that they must be forthcoming about risks. One remarked, *"You can't appreciate the importance of security without first understanding what's at stake, what's at risk" (P14).* Another recommended, *"In terms of it being scary... take that head on. Here are all the terrible things that can happen" (P08).*

However, six participants noted the importance of being discerning: not "crying wolf" (being an unnecessary alarmist) over every little security issue, lest their audience become overwhelmed, disinterested, or skeptical. One said a mistake in security advocacy is *"being more sensational, and theoretical, and hypothetical than practical and rational... Focusing on the possibility is a very easy way to get known as crying wolf" (P02).*

In some cases, advocates may only want to engage a select group with the authority to address a security issue, especially when dealing with issues that have broader-reaching organizational or national consequences. An advocate who promotes security to industries that build safety-critical products, such as medical devices, commented, *"If I told everybody what I know, they'd freak out. I want to tell a smaller*

list of people I know so that we can quietly fix it" (P11).

5.2.2 Empowering

For many users, an overabundance of fear may result in a feeling of futility regarding their security situation. This can lead to paralysis and inaction. This was echoed by one participant when she opined, *"if you have a little bit of fear, it's actionable. But if you have too much fear, it becomes so overwhelming that you give up on it" (P21).*

Feelings of helplessness can be perpetuated by security professionals who regularly express their belief that users are unable to comprehend and practice good security behaviors (mentioned by six). An advocate who had led her company's security awareness program expressed her frustration with these professionals: *"I feel like there's just a lot of people saying, 'Oh humans are the weakest link. They're always hopeless... They haven't changed their behaviors, so what's the point?"' (P21).* Another commented on the harsh way non-experts are treated by security experts in online forums, remarking, *"smashing them and telling them they're stupid, that's not going to help. Instead, we need to be more encouraging, more open-armed in the industry" (P10).*

To overcome feelings of inadequacy, advocates must empower their audience to take action. Empowerment was a concept mentioned by 16 participants, mostly in the context of non-technical users. A prerequisite seemed to be instilling a sense of hope, as noted by eight advocates. One participant reflected:

> *"You can't last for decades in this cybersecurity business without being one of two personality types: the hopeless cynic or the hopeless optimist... You can make an entire living just pointing out other people's problems or mistakes... But I just don't find that satisfying. I'm much more interested in creating positive change" (P04).*

Advocates then use this hope to foster self-efficacy in their audience. Self-efficacy is a belief in one's own ability to exert control in specific situations or accomplish a task [3]. This is the cornerstone of independence, which was expressed by one advocate when he said, *"we have to be able to get to a point at which they can do a lot of it themselves" (P01).*

The interviews suggested that self-efficacy can be encouraged by providing people with basic, concrete actions that will allow them to be proactive in their security situation. Instead of simply raising an alarm, a security technologist believed, *"it's really important to tell people what they can do so they that don't just go, 'Oh my gosh. The world is a scary place, but there's nothing I can do about it, so I guess I just won't worry about it' " (P07).* Another commented,

> *"I love empowering people and seeing their lightbulbs go off in the moment that they understand why they are a target and what they can do about it. So, it's not a place of fear. You have to start with fear to get them to understand that there's a problem, but then you also give them the tools" (P21).*

Framing messages in a positive light and comparing security measures to more familiar, accessible protective mechanisms can also help to alleviate fear and empower. A security advocate talked about how she chooses to frame her communications during her work with senior citizens:

> *"you slip that message of 'You're going to get attacked and everything's going to get stolen' to 'Well, it's kind of like home improvement when you put a better deadbolt on your door or you decide that you're going to shore up your foundation"' (P23).*

5.3 Overcoming "It's Confusing"

Few non-professionals have the technical know-how to address security issues, so *"security is mysterious to most people" (P07).* A participant underscored the impact of this knowledge deficiency when she commented, *"people don't actually know what the names of the tools they need are. They don't know the proper, technical words that are going to lead them to a solution" (P23).* This lack of understanding leads to the perception that security solutions are confusing and difficult to implement, as noted by 20 participants.

The barrage of security messages and advice people receive at work, from the media, and from friends can create *"a lot of uncertainty of what is the right thing to do" (P04).* One participant commented on this state of being overwhelmed: *"You're getting hit from every single side... We have almost an information overload happening, and it's hard to sort through it" (P08).*

Security can also be seen as a burden, *"just one more thing to remember, one more rule" (P28)* that gets in the way of doing other tasks. A participant observed, *"there's a complete misunderstanding that to be secure takes an immense amount of time. That's a huge obstacle to get over" (P23).*

To overcome the perception that security is confusing, advocates "bridge the gap" between security experts and non-experts, educate people on how to practice good security, provide practical recommendations, and promote usable security solutions.

5.3.1 Bridging the Gap

The process of mediating between technical and non-technical audiences requires establishing common ground, which necessitates advocates to have strong communication and translation skills and an awareness of audience context. A nonprofit director underscored the importance of communicating in a manner that is meaningful to the audience:

> *"you can produce as many policies and processes as you like, if you cannot communicate them to people in a language that they understand, in a language that means they're going to be receptive to your message, then they're worthless" (P27).*

A security consultant described his role as a connector between groups: *"I'm sort of the in-between person, between the business interests of the company and the technical interests because they don't talk to each other very well. I can translate both languages" (P08).*

Bridging the gap was a concept discussed in 22 interviews. Participants described their connective capacity with various terms such as translators, boundary spanners, ambassadors, cross-pollinators, and information carriers.

Translating: Highly technical security experts often unwittingly make security seem more elusive as they rely heavily on disciplinary jargon. One participant remarked, *"There's*

also, I think, a big language issue... it is a highly technical field with a very specialized language" (P04). A lack of understanding of the skill level of their audience also results in confusion, as described by a security awareness educator: *"It's not that people are stupid, it's that we need to communicate in their language" (P09).*

To overcome the language difference, advocates act as "translators," reframing highly technical concepts using terms their audience can understand. Twenty-three participants commented that the underlying communication skills required for translation were important for security advocacy. In fact, despite being a highly technical person, when asked about the characteristics of successful security advocates, a security consultant said, *"communication skills I think are number one" (P10).* While describing the importance of effective communication in his work, a participant asserted, *"Being able to translate complicated things very simply is crucial to... advocating security" (P02).*

Being Context Aware: Context awareness is critical for effective security advocacy, as expressed by 22 participants. As much as possible, they need to be aware of the operational environment of their audience, including technology, roles, social structures, constraints, and goals. One participant commented, *"Understanding your environment, and the different, unique threats and vulnerabilities in your environment is hugely important" (P14).* A non-profit organizer used a metaphor to convey this necessity:

> *"This is more of an ambassador role where you're going to a foreign country. You need to represent your own country, but you have to assimilate to and acclimate to the language and the beliefs and the culture that you are trying to affect" (P11).*

Advocates must also use their knowledge of context to tailor the security message to the skill level and concerns of the audience. When appealing to non-technical audiences, a veteran security evangelist realized, *"You have to change your language, which means in the non-techno speak figure out how to translate what you know into concerns people have about economic and social issues" (P04).* A security engineer who advocates to a wide swath of people within his organization remarked, *"The message, even though it's going to be the same, it's going to be delivered differently depending on the level of person that you're talking to" (P13).*

5.3.2 Educating

A greater understanding of security helps to overcome confusion and leads to empowerment, as discussed earlier. To that end, advocates saw themselves as security educators. Eleven participants had served at one point in a formal educator role, but all discussed the educational component of their jobs. A security awareness director at a large university saw his role as foundational: *"The only way you can fully understand what's at risk and what's at stake is through education and awareness. So, it's the starting point for everyone. I'm ground zero in security" (P14).*

Eleven participants mainly taught non-technical audiences. Their goal was to provide simple, straightforward instruction and help people make informed decisions about their security behaviors: *"I think it's a lot like knowing when you see power lines are down, not to touch the power lines. It's just a basic level of knowledge you need to know for self-preservation purposes" (P15).* For example, P08 created *"security awareness basics"* videos targeted at the general public. For his other audiences of non-technical professionals in the legal, healthcare, and finance industries, he tailored both video and in-person presentations to their specific needs. He commented on the value of his security education courses: *"I'm not going to make you into a security expert in three hours ... But I want you to be able to have a conversation with one where you can be able to follow each other" (P08).*

In contrast, 15 participants primarily taught technical audiences of developers, IT specialists, college students, and other security professionals on issues such as secure products and network security. For example, P22 educated product teams within his organization on secure development practices. Five mentioned that it was important to educate the next generation of security professionals *"so that they don't make or sustain the same mistakes... that got us into the mess that we're in with cybersecurity" (P02).* One participant often does presentations for high school students at cybersecurity summer camps, *"just talking about information security, and just having fun and making them laugh. And talking about how meaningful this is" (P10).*

5.3.3 Providing Practical Recommendations

Our participants agreed that the amount of security information to be aware of can be overwhelming, even for them. To counter this, 16 participants discussed providing practical, prioritized recommendations. Six mentioned condensing security information into more manageable chunks containing the most important security actions to take. P11 mentioned how his advocacy group had developed a set of *"first principles,"* which are foundational security measures that should be in place within an organization before anything else.

While some security guidance is universal, other recommendations are dependent on the audience's environment. Several participants spoke out against "one-size-fits all" solutions, emphasizing the importance of context. A non-profit organization approached this issue by producing general guidance that can be customized and disseminated by others: *"Our goal is to create non-proprietary resources so that our local partners can take those and tailor them for their community... because it could mean different consequences for different people" (P24).* Others felt the responsibility to directly provide tailored security guidance that is based on the actual risk within a given situation. One participant was a proponent of this approach within organizations:

> *"I think in the security area there's a lot of mythology and a lot of things we do because we heard it's the right thing to do, and we have no idea why, but everybody else seems to be doing it, so we should do it, too. And so, trying to get people to stop and think it through, and figure out what's actually going to be effective" (P07).*

To ensure guidance was practical to their audience, an advocate in higher education described her organization's efforts to regularly poll members on their biggest security risks and challenges. These risks then became the cornerstone of their annual "top 10 list" of security recommendations:

> *"You're never going to be able to remediate or mitigate every single information security risk that you have,*

but you should be able to identify the ones that are the most likely and the ones that would be the most devastating to your environment, and take steps to mitigate those" (P15).

5.3.4 Promoting Usability

Security technologies and policies are not generally known for usability, leading to feelings of frustration and confusion [10, 52]. One participant felt that security professionals are *"putting too much pressure on the user, and the user doesn't have the knowledge" (P03).* She also observed that the volume of security-related tasks a user must perform on a daily basis (e.g., multiple logins, security warnings) can be overwhelming when viewed as a whole: *"In isolation, none of these security things are that big of a hardship or have significant usability concerns. The aggregation of them is what causes the usability concerns" (P03).*

To alleviate the complexity and burden of security, nine participants emphasized the need to advocate for systems and policies that are usable, minimize requisite knowledge, and compensate for the inevitability of user error. Three participants conducted usability research to directly influence vendor products as well as organizational and national policies. One of these participants explained her motivation metaphorically: *"Most of us drive a car, but don't know how to fix cars. We shouldn't have to know how to fix cars in order to drive them. And I think that should be true about computers, too" (P07).* Another participant who has been a champion for usable security both internally and externally to his organization, stated that the usable security challenge must address the question of *"How do we build and deploy systems that are easy to use, easy to manage, that result in cost savings?" (P16).*

5.4 Overcoming "It's Dull"

Another negative perception, voiced by 19 participants, is that security is boring, not relevant, not of concern, or not worth the investment. This drives user apathy in adopting good security behaviors.

Security can seem boring to less technical audiences, especially when a technologist fails to frame it in terms the audience can understand. This can be exacerbated by poor communication skills, for example *"presentations where the speaker's doing monotone and talking security. If you really love it, you can get through those, but for normal people, they're torture" (P08).* Additionally, the most common negative exposure users receive is from their annual security awareness training for organizational compliance, described vividly by one participant as *"a layer of Dante's Hell" (P21).* A security engineer who had once been tasked with refreshing an organization's security training noted that the original training *"was boring... there [was] absolutely nothing to get the user to buy into security thinking"* (P12). The training often mandates specific actions that are deemed unwelcome, unnecessary inconveniences. For example, one participant lamented password policies: *"You force them to change their password. We all hate that" (P28).*

Besides disinterest, people may be apathetic towards security because of not appreciating their own personal vulnerability and responsibility. A security awareness director expanded on this: *"if people don't understand why and how this affects them, they're simply not going to comply with whatever initiative it is you're trying to roll out" (P14).* Another participant discussed how security is not something most people take under consideration when acquiring a computing device: *"We don't want secure,... we don't even want to think about it. [We want] pretty, functional, cheap" (P06).* Security is also not a primary function for most: *"I think for end users, it's just nobody wants to spend their time doing security. That's not what they signed up for when they bought a computer" (P07).* Lack of concern may be partially due to a *"belief of it won't happen to me. It's like I'm a great driver, so I can text while driving because it won't ever happen to me, so I don't have to worry about it" (P21).*

From an institutional perspective, organizations may also be apathetic to security because it can be hard to show a clear return on any investment. Security measures are preventive in that they are implemented to lower the likelihood of some unwanted event occurring in the future [36]. Therefore, it is hard to measure prevented events because they typically cannot be observed. A participant discussed this challenge, remarking, *"It's hard to prove that it's working for you. Is it working because you've done such a good job and you've invested in all the right places, or is it working because you're just not the target today?" (P05).* An advocate working at a non-profit observed:

> *"One of the other trends that we see... is that of cyber fatigue in the boardroom: people constantly asking for more resources, yet they can't guarantee any form of security. There's no real return on investment, and it seems to be a black hole that we pour money constantly into" (P27).*

Cybersecurity advocates attempt to overcome boredom and apathy by incentivizing security behaviors and using engaging communication techniques.

5.4.1 Incentivizing Security Behaviors

Successful advocates must be able to persuade their audience to practice good security behaviors by appealing to both intrinsic and extrinsic motivations, as mentioned by 17 participants. A former security awareness director reflected, *"I really want to get people to want to do security instead of having to" (P21).*

Selling Security: Advocates, in effect, must market security in order to motivate people to take appropriate security actions. A participant commented on the importance of marketing skills: *"you have to be able to make a... good case... that's based on good data, that the dollar figures support, and that you can get excited and get them excited about. And if you can't... market that, you can forget it" (P03).* One advocate had an interesting and honest perspective on his use of persuasion:

> *"I am trying to drive them to make themselves more secure by using various argumentative techniques, and, in some way, that's manipulating them. But if you're manipulating somebody for their own good, that's not wrong" (P10).*

As discussed earlier, having context awareness is critical to being able to sell security in a manner that the audience understands and cares about. One advocate observed, *"you need to be able to be flexible in terms of adapting your argu-*

ment to their particular needs" (P06). Another commented: *"It's not a one-size-fits-all approach. You could take a given security concern and have to frame it four or five different ways depending on who you're talking to" (P02).* As an example, one security consultant was having a difficult time convincing an executive to spend resources to implement secure hypertext transfer protocol (HTTPS) for his company's website. However, when the consultant mentioned that Google ranks websites using HTTPS higher in its search results, the executive immediately changed his mind since *"Their biggest business risk was not being on the first page of Google" (P10).*

Ten participants said that they must also be able to communicate the reasons behind their security recommendations in order to convince their audience of potential benefits. An advocate stressed the importance of providing concrete reasoning: *"We gotta stop leading with 'what' and start leading with 'why.' Like why does this matter? If you get someone to care why, they'll seek the what and the how" (P11).* Along this vein, for those advocating within an organizational context, establishing the business drivers for security is essential. A former business executive believed, *"we should be concerned with selling security as mission assurance, revenue assurance, reputation assurance" (P02).*

Interestingly, three participants thought that lessons learned from persuasion within the public health field could inform the security advocacy field. An IT executive commented:

> *"It has struck me that we have not leveraged the hundred plus years of research in public health to really garner how to change people's behavior effectively. How do you teach people to wash their hands? How do you teach people to do the handful of basic things that we know will solve 80% of the problems is the hard part of this" (P18).*

A non-profit security evangelist echoed this thought, saying that public health is *"well-defined, it's a social expectation, and you know that it provides value even though you probably can't quote the actually medical studies... You should just do it. We're not to that stage yet [in security]" (P04).*

Creating Reward (and Consequence) Systems: Advocates encourage a culture that incentivizes security adoption. As mentioned earlier, showing return on investment in security can be a challenge. A non-profit director saw his role of influencing public policy as critical to creating an economic reward structure for organizations to practice good security: *"Most of these people are not doing what they ought to be doing with cybersecurity for economic reasons. And so we need to find ways to make cybersecurity more economically attractive to these people" (P06).*

Several participants saw economic incentives as only part of the solution in that they need to be coupled with appeals to the values of the audience. For example, one participant discussed motivating secure development practices, not by framing them in security terms, but in terms developers care about, such as *"you can avoid unplanned, unscheduled work, you'll be on time, on budget, you'll reclaim 20% boost in developer productivity across the calendar year. You'll get your bonus. You'll crush your competitors" (P11).*

We uncovered a tension regarding the use of negative reinforcement strategies based on audience type. Three participants pushed for more accountability with negative consequences for organizations that experience serious security breaches that result in the loss of sensitive, personal information. However, three others believed that negative incentives were not useful from an end user perspective. A security awareness director at a large university opined that these kinds of incentives are *"completely the wrong way to approach things in security. It's all about education. It's all about driving awareness, raising awareness, and getting people to understand the importance of security through non-punitive measures" (P14).* Another participant felt that simple, positive incentives could be effective, but observed:

> *"security teams generally have a lot of history and best practice in negative behaviors... We have very few examples where, 'Here are the compliance requirements. When you meet or exceed this, we will reward and recognize you as being a champion'... It doesn't have to be monetary, it can be a thank you" (P21).*

5.4.2 Using Engaging Communication Techniques

To overcome feelings that security is boring and irrelevant, advocates attempt to make their communications engaging and relatable while varying communication channels.

Exhibiting Enthusiasm: To overcome disinterest, participants felt that modeling enthusiasm for security to their audience captured their attention and promoted greater engagement. This was not difficult for the participants, considering 18 expressed passion for their role as advocates. The director of a non-profit effused, *"I believe in what we're doing, and I think we're making the world a better place" (P06).* When asked about effective security advocates he had encountered in his career, a security engineer mentioned those for whom *"you can really feel the energy that they believe in it" (P12).* Another participant expressed the importance of having passion for her work when she remarked, *"I can't sell something I don't believe in. I can't sell something I don't like. I mean, I'm not going to sit and lie to you. And so, I am passionate about it" (P03).*

Making Security Relatable: Our findings reveal that advocates also overcome apathy by making security relatable, described by one participant as putting *"the personal use and behavior in it so that people own what you're telling them" (P28).* To do so, they often used rhetorical devices in both written and oral communications to convey meaning and persuade people to take action. Among the rhetorical devices mentioned were anecdotes and narratives (by eight advocates), analogies and metaphors (4), imagery (3), alliteration (2), and pop culture references (2).

Narratives might involve stories about hypothetical, but plausible scenarios, or actual occurrences of security-related incidents. One advocate liked to share stories about her own experiences since she believed, *"Personalizing the message is useful, seeing that this happens to real people" (P07).* Another discussed how he shares stories of things that have happened to others, for example *"a person whose money might have been stolen out of their bank account because of the poor security they did at home, not because of what the bank did, but because of what they did" (P05).* Four advocates mentioned how leveraging narratives of current events can serve as *"an opportunity because [the audience's] aware-*

ness is already heightened" (P24).

Advocates also used analogies and metaphors to relate security to situations and phenomena that are more familiar to their audience. For example, the analogy to public health and basic hygiene (e.g., washing your hands, brushing your teeth) was mentioned several times to explain the concept of cyber hygiene (basic, fundamental security practices). This was described by one participant:

> *"Do you tell someone to exercise and get enough sleep, or do you wait until they are having some serious problems and then you're going to bring them in for surgery? Which route would you rather go? It's not exactly the same, but it's kind of analogous to what's going on there [with security]. And it's getting people to understand, OK, here's your basic network health hygiene" (P08).*

Even though analogies and metaphors can be useful, two participants cautioned that these must be meaningful and tailored to the audience. One participant thought that oversimplifying these *"can be dangerous. You can be too glib, and it's superficial" (P06).* P08 provided a critique of a security training video that depicted someone fishing to explain the concept of phishing. He felt that such a metaphor was *"corn-ball"* (trite and unsophisticated) and would not resonate with his audience of attorneys.

Visual representations were also valuable in making complex topics more relatable and memorable. For example, after the Heartbleed vulnerability [46] was announced, one participant said, *"I was trying to explain that and ended up using a cartoon to explain a very complex topic to people" (P25).* Another revealed, *"When I start talking about two-factor authentication, I like to call it two-raptor authentication. I like dinosaurs. It's more fun when you imagine that they're going to eat someone's face off. People will remember the name of the feature" (P23).*

The participants used a variety of platforms to peak interest and advocate for security. The most mentioned communication channels were: written materials (e.g., books, papers, frameworks, newsletters) (by 18 advocates); small group/individual face-to-face interactions (17); large forum and conference presentations (16); social media (12); and formal classroom training (9). Interestingly, several participants utilized particularly unique communication channels. For example, three had developed games to teach security concepts. One organization sponsored a food truck event for their employees during which people standing in line for their lunch were engaged with security trivia. Another creative idea was putting a security-themed vinyl wrap on a public bus: *"it becomes essentially... a traveling billboard" (P26).*

6. IMPLICATIONS

Although there have been research studies exploring techniques to encourage security best practices and technology adoption, there is much to learn from successful practitioners who are engaged in this activity on a regular basis. We also discuss the potential of cybersecurity advocacy becoming a formally recognized work role in the security field.

6.1 Advancing Risk Communication

6.1.1 Relationship to Non-Security Risk Domains

Our results confirm that cybersecurity advocates exercise many of the same risk communication best practices observed in other fields such as health (e.g., [9]), environmental hazards (e.g., [39]), and home security (e.g., [11]). For example, they expressed common goals, such as building trust, creating awareness strategies, and motivating people to act. To build trust and credibility, they employed a variety of non-technical "soft" skills. In communicating their message, security advocates similarly used engaging techniques, served as a bridge between security experts and non-experts, encouraged audience self-efficacy, and tailored their message to the audience based on context awareness.

Similarities suggest there may be much to learn from risk communication in non-security fields, especially those with longer histories and hard-fought successes. In particular, we see the value of greater investigation into how lessons learned in public health advocacy might be applied to cybersecurity given that this connection was raised repeatedly in our interviews.

Moving beyond these similarities, our findings suggest some unique properties of security risk which may advance the overall discipline. Specifically, we identified how advocacy in the security domain may be more urgent and challenging than in other domains, and may require additional tactics. Foremost, the security field is incredibly dynamic, having to adjust to constantly changing technologies and defend against determined adversaries who can exact significant damage with relatively little cost or sophistication. Second, security applies to everyone and every organization within an interconnected, technology-dependent society. However, most are not equipped to deal with security measures since security consists of abstract concepts not well understood by the typical person, and people are often dependent on security interfaces with poor usability. Motivation to enact security measures may likewise be problematic because consequences of poor behavior are not always immediate or easily observed. The economics and effectiveness of security are hard to measure. To better explore these similarities and differences, we see future research potential in performing in-depth comparisons of security advocacy practices to those in other domains.

6.1.2 Strategies for Communicating Security Risk

As discussed in Related Work, most prior research on persuasive methods in the security context has taken the perspective of the target end users and not those who do the influencing. Our study addressed this gap and does indeed confirm previous findings concerning the value of small group interactions [1], the necessity of framing security communications [40], the use of positive vs. negative incentives [15], and the importance of encouraging security self-efficacy [35]. However, our findings go beyond, for example, by uncovering a set of non-technical advocacy skills and competencies focused on bridging the gap between security experts and non-experts.

The study also questions the universal effectiveness of rhetorical devices like narratives, analogies, and metaphors, within security contexts. For example, even though only four participants explicitly said that they employ metaphors and analogies when communicating with their target audiences,

we observed that many advocates naturally used a variety of these to describe security concepts during our interviews. This was the case even though the interviewer was known to be a security expert, suggesting that their use goes beyond being purely instructive. Additionally, in line with the mixed findings in past studies about the efficacy of metaphors [4], two of our participants cautioned against incorrect use of these for fear that they may oversimplify security concepts and create misunderstandings leading to risky behaviors. Future work is needed to look more deeply into the use of metaphors and the level of detail and relevancy they must provide in order to affect security learning and behavior.

We also see value in applying security advocacy techniques to the design of security interfaces and training resources. To overcome negative perceptions of security, these resources should aim to empower users to take appropriate security actions and create a positive affect towards security. Security resources should create a level of concern without overwhelming or paralyzing by conveying severity and likelihood in clear, understandable terms. Resources must be usable, tailored to the audience, and encourage empowerment by providing concrete, achievable steps in simple language. Additional references to threat information that includes real stories of security incidents can lend greater credibility to risk claims. Training materials should consider the incorporation of storytelling and other creative rhetorical methods (e.g., [21]).

6.2 Emerging Cybersecurity Advocate Role

The majority of the participants in our study demonstrated an innate understanding of human behavior, even though few had formal training in the social sciences or humanities. They regularly employed techniques to combat common behavioral heuristics and biases that could negatively affect security decisions, as suggested by Pfleeger and Caputo [29]. For example, they addressed cognitive load by breaking recommendations down into manageable, prioritized chunks. Storytelling, sharing personal experiences, and referencing recent events helped with availability, which is an evaluation of the likelihood of an event based on recall of similar events happening.

Although the participants seemed to consider these behavioral aspects (even if subconsciously), they suggested that most other security practitioners do not share this basic interpersonal orientation. These professionals often contribute to negative perceptions of security by not taking the human element under consideration when describing, designing, administering, and enforcing security mechanisms. The advocates revealed a desire to move away from common security practitioner perceptions that "users are the enemy" and "users are stupid" to instead take the human element under consideration and regard users and security professionals as capable partners.

In some meaningful ways, the practice of securing a system appears to be different than advocating for securing a system. Yet, there is no professional preparation for the latter. An analogous situation was the foundation of human-factors/usability engineering as a profession distinct from other disciplines such as testing or business analysis. This was rooted in a discovery that human errors in systems were fundamentally different than system errors. As a result, this observation necessitated different approaches. Similarly, we see a new and rapidly growing need for security professionals with a special set of advocacy skills and techniques. Therefore, we propose that there should be continuing education efforts to aid in the progression to cybersecurity advocate from both security and non-security fields.

Currently, most of the emphasis within security professional development curricula [2,24,25] is on gaining technical knowledge. A quick review of cybersecurity work roles in these guidelines reveals that none contain set of skills that resemble the work of an advocate. Therefore, our future work may include developing a framework of cybersecurity advocate knowledge, skills, and abilities, along with an outline of a development program to facilitate advocate professionalization.

6.3 Limitations and Future Work

The study has a few limitations we intend to address in the next phase of the project. This study is a one-sided view through the lens of security advocates themselves. While this is critical for constructing a grounded understanding of their work, it does not provide evidence that any of the techniques they deemed successful were indeed effective with their intended audiences. Second, interviews may suffer from self-report bias in which participants may adjust their answers to appear more acceptable to the researcher, who had been revealed to be a security professional. This was a reasoned tradeoff in the study design meant to assist with recruitment and trust building as the researcher spoke the same technical language as the participants.

To address these potential issues, results can be triangulated with data from planned follow-on studies. We next intend to reverse the polarization of our lens and work with a diversity of organizations to understand their experiences with security advocates.

7. CONCLUSION

Cybersecurity advocates are emerging as a unique role in the ecology of security professionals. They employ a variety of skills and techniques to overcome negative perceptions of security. Our study confirms the applicability of past risk communication literature to the security domain while revealing additional considerations to address differences in cybersecurity. It also proposes the establishment of a new cybersecurity advocate career track to address the urgent need for security adoption.

Acknowledgements

The authors would like to thank the following people for their insightful comments that resulted in improvements to this paper: the anonymous reviewers, Yasemin Acar, Sascha Fahl, Sandra Spickard Prettyman, and Mary Theofanos.

8. REFERENCES

[1] E. Albrechtsen and J. Hovden. Improving information security awareness and behaviour through dialogue, participation and collective reflection: An intervention study. *Computers & Security*, 29(4):432–445, 2010.

[2] Association for Computing Machinery. Toward curricular guidelines for cybersecurity: Report of a workshop on cybersecurity education and training.

https://www.acm.org/education/TowardCurricularGuidelinesCybersec.pdf, 2013.

[3] A. Bandura. Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 1977.

[4] G. L. Brase, E. Y. Vasserman, and W. Hsu. Do different mental models influence cybersecurity behavior? Evaluations via statistical reasoning performance. *Frontiers in Psychology*, 8, 2017.

[5] C. Bravo-Lillo, L. F. Cranor, J. Downs, and S. Komanduri. Bridging the gap in computer security warnings: A mental model approach. *IEEE Security & Privacy*, 9(2):18–26, 2011.

[6] L. J. Camp. Mental models of privacy and security. *IEEE Technology and Society Magazine*, 28(3), 2009.

[7] H. H. Clark and S. E. Brennan. Grounding in communication. *Perspectives on Socially Shared Cognition*, 13:127–149, 1991.

[8] J. Corbin and A. L. Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage, Thousand Oaks, CA, 4th edition, 2015.

[9] V. Covello. Risk communication. In H. Waldron and C. Edling, editors, *Occupational Health Practice*, chapter 6. Arnold Publishers, London, 1997.

[10] L. F. Cranor and S. Garfinkel. *Security and usability: designing secure systems that people can use*. O'Reilly Media, Inc., Boston, MA, 2005.

[11] M. Dolata, T. Comes, B. Schenk, and G. Schwabe. Persuasive practices: Learning from home security advisory services. In *International Conference on Persuasive Technology*, pages 176–188, 2016.

[12] S. Furnell and K.-L. Thomson. Recognising and addressing 'security fatigue'. *Computer Fraud & Security*, (11):7–11, 2009.

[13] B. G. Glaser and A. L. Strauss. *The Discovery of Grounded theory: Strategies for Qualitative Research*. Transaction Publishers, 2009.

[14] J. A. Gordon. Meeting the challenge of risk communication. *Public Relations Journal*, 47(1):28, 1991.

[15] T. Herath and H. R. Rao. Encouraging information security behaviors in organizations: Role of penalties, pressures and perceived effectiveness. *Decision Support Systems*, 47(2):154–165, 2009.

[16] D.-L. Huang, P.-L. P. Rau, and G. Salvendy. A survey of factors influencing people's perception of information security. In *International Conference on Human-Computer Interaction*, pages 906–915, 2007.

[17] I. Ion, R. Reeder, and S. Consolvo. '... no one can hack my mind': comparing expert and non-expert security practices. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 327–346, 2015.

[18] A. C. Johnston and M. Warkentin. Fear appeals and information security behaviors: an empirical study. *MIS quarterly*, pages 549–566, 2010.

[19] R. Kang, L. Dabbish, N. Fruchter, and S. 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)*, 2015.

[20] R. E. Kasperson, D. Golding, and S. Tuler. Social distrust as a factor in siting hazardous facilities and communicating risks. *Journal of Social Issues*, 48(4):161–187, 1992.

[21] E. Lastdrager, I. C. Gallardo, P. Hartel, and M. Junger. How effective is anti-phishing training for children? In *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[22] J. E. Maddux and R. 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.

[23] S. Merriam and E. Tisdell. *Qualitative Research: A Guide to Design and Implementation*. John Wiley & Sons, San Francisco, CA, 4th edition, 2016.

[24] National Security Agency. National Centers of Academic Excellence in Cyber Defense. https://www.nsa.gov/resources/educators/centers-academic-excellence/cyber-defense/, 2018.

[25] W. Newhouse, S. Keith, B. Scribner, and G. Witte. NIST Special Publication 800-181: National Initiative for Cybersecurity Education (NICE) cybersecurity workforce framework. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-181.pdf, 2017.

[26] J. R. Nurse, S. Creese, M. Goldsmith, and K. Lamberts. Trustworthy and effective communication of cybersecurity risks: A review. In *Workshop on Socio-Technical Aspects in Security and Trust (STAST)*, pages 60–68, 2011.

[27] Office of the Director of National Intelligence. Assessing Russian activities and intentions in recent US elections. https://www.dni.gov/files/documents/ICA_2017_01.pdf, Jan. 2017.

[28] M. Q. Patton. *Qualitative research and evaluation methods*. Sage, Thousand Oaks, CA, 4th edition, 2015.

[29] S. L. Pfleeger and D. D. Caputo. Leveraging behavioral science to mitigate cyber security risk. *Computers & Security*, 31(4):597–611, 2012.

[30] C. Posey, T. L. Roberts, P. B. Lowry, and R. T. Hightower. Bridging the divide: a qualitative comparison of information security thought patterns between information security professionals and ordinary organizational insiders. *Information & Management*, 51(5):551–567, 2014.

[31] G. V. Post and A. Kagan. Evaluating information security tradeoffs: Restricting access can interfere with user tasks. *Computers & Security*, 26(3):229–237, 2007.

[32] S. S. Prettyman, S. Furman, M. Theofanos, and B. Stanton. Privacy and security in the brave new world: The use of multiple mental models. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 260–270, 2015.

[33] F. Raja, K. Hawkey, S. Hsu, K.-L. C. Wang, and K. Beznosov. A brick wall, a locked door, and a bandit: a physical security metaphor for firewall warnings. In *Symposium on Usable Privacy and Security (SOUPS)*, page 1, 2011.

[34] E. M. Redmiles, A. R. Malone, and M. L. Mazurek. I

think they're trying to tell me something: Advice sources and selection for digital security. In *IEEE Symposium on Security and Privacy*, pages 272–288, 2016.

[35] H.-S. Rhee, C. Kim, and Y. U. Ryu. Self-efficacy in information security: It's influence on end users' information security practice behavior. *Computers & Security*, 28(8):816–826, 2009.

[36] E. M. Rogers. *Diffusion of Innovations*. Simon and Schuster, New York, NY, 5th edition, 2003.

[37] R. W. Rogers. Cognitive and psychological processes in fear appeals and attitude change: A revised theory of protection motivation. In J. T. Cacioppo and R. E. Petty, editors, *Social psychophysiology: A sourcebook*, pages 153–176. Guilford Press, New York, NY, 1983.

[38] K. E. Rowan. Why rules for risk communication are not enough: A problem solving approach to risk communication. *Risk Analysis*, 14(3):365–374, 1994.

[39] P. M. Sandman. Getting to maybe: Some communications aspects of siting hazardous waste facilities. In T. S. Glickman and M. Gough, editors, *Readings in Risk*. RFF Press, Washington, D.C., 2013.

[40] J. D. Shropshire, M. Warkentin, and A. C. Johnston. Impact of negative message framing on security adoption. *Journal of Computer Information Systems*, 51(1):41–51, 2010.

[41] M. T. Siponen. A conceptual foundation for organizational information security awareness. *Information Management & Computer Security*, 8(1):31–41, 2000.

[42] P. Slovic. Perception of risk. *Science*, 236(4799):280–285, 1987.

[43] T. Sommestad, H. Karlzén, and J. Hallberg. A meta-analysis of studies on protection motivation theory and information security behaviour. *International Journal of Information Security and Privacy (IJISP)*, 9(1):26–46, 2015.

[44] B. Stanton, M. F. Theofanos, S. S. Prettyman, and S. Furman. Security fatigue. *IT Professional*, 18(5):26–32, 2016.

[45] Symantec Corporation. 2016 internet security threat report. `https://www.symantec.com/content/dam/symantec/docs/reports/istr-21-2016-en.pdf`, 2016.

[46] US-CERT. OpenSSL Heartbleed vulnerability (CVE-2014-0160). `https://www.us-cert.gov/ncas/alerts/TA14-098A`, 2014.

[47] Verizon. 2016 data breach investigations report. `http://www.verizonenterprise.com`, 2016.

[48] K. Waddell. Yahoo suffers history's biggest known data breach. *The Atlantic*, Dec. 14, 2016.

[49] R. Wash. Folk models of home computer security. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 11–26, 2010.

[50] J. L. Yang and A. Jayakumar. Target says up to 70 million more customers were hit by December data breach. *Washington Post*, Jan. 10, 2014.

[51] L. Zhang-Kennedy, S. Chiasson, and R. Biddle. Stop clicking on update later: Persuading users they need up-to-date antivirus protection. In *International Conference on Persuasive Technology*, pages 302–322, 2014.

[52] M. E. Zurko and R. T. Simon. User-centered security. In *Proceedings of the 1996 Workshop on New Security Paradigms*, pages 27–33, 1996.

APPENDIX

Interview Questions

1. Can you tell me about what you do in your job?
2. How did you come to do this type of work?
3. What motivates you to do this work?
4. What do you think is the importance of your role in promoting security?
5. How is your work is valued by others?
   (a) What kind of feedback do you get?
   (b) Can you talk about any times when you felt that your work wasn't appreciated?
6. What do you think are qualities or characteristics of people who are successful in promoting security?
7. Have you had experiences with or know of security advocates who you don't think were particularly effective? What was it about them or what did they did or did not do that contributed to their ineffectiveness?
8. Through what means do you promote security? For example, conferences, invited talks, blogs, social media, articles, client visits, face-to-face meetings, phone, email.
   (a) Which of those means do you think are the most effective? Why?
9. What are your thoughts about whether or not you are reaching the right population of people and organizations?
   (a) What is preventing you from reaching the right people?
   (b) What do you wish you could do to reach the right population?
10. How do you keep up with the latest in security?
11. What do you find most rewarding about your work?
12. What do you find most challenging or frustrating about your work?
13. What do you think are the biggest obstacles individuals and organizations face with respect to implementing security measures and technologies?
14. What do you see as your role in helping organizations overcome these obstacles?
15. Is there anything else you'd like to add with respect to what we've talked about today?

Ethics Emerging: the Story of Privacy and Security Perceptions in Virtual Reality

Devon Adams
University of Maryland, Baltimore County
adadev1@umbc.edu

Alseny Bah
University of Maryland, Baltimore County
abah4@umbc.edu

Catherine Barwulor
University of Maryland, Baltimore County
barwulo1@umbc.edu

Nureli Musabay
James Madison University
musabanx@dukes.jmu.edu

Kadeem Pitkin
College of Westchester
kpitkin@cruiser.cw.edu

Elissa M. Redmiles
University of Maryland
eredmiles@cs.umd.edu

ABSTRACT

Virtual reality (VR) technology aims to transport the user to a virtual world, fully immersing them in an experience entirely separate from the real world. VR devices can use sensor data to draw deeply personal inferences (e.g., medical conditions, emotions) and can enable virtual crimes (e.g., theft, assault on virtual representations of the user) from which users have been shown to experience real, significant emotional pain. As such, VR may involve especially sensitive user data and interactions. To effectively mitigate such risks and design for safer experiences, we aim to understand end-user perceptions of VR risks and how, if at all, developers are considering and addressing those risks. In this paper, we present the first work on VR security and privacy perceptions: a mixed-methods study involving semi-structured interviews with 20 VR users and developers, a survey of VR privacy policies, and an ethics co-design study with VR developers. We establish a foundational understanding of perceived risks in VR; raise concerns about the state of VR privacy policies; and contribute a concrete VR developer "code of ethics", created by developers, for developers.

1. INTRODUCTION

Virtual Reality (VR) technology aims to create "immersive, interactive, and imaginative" simulations for the user through visual, haptic, and auditory output [14]. The goal of VR is to create an entirely immersive experience that fully transports the user away from reality and into a virtual world [48]. While VR headsets have existed since the 1960s, they are fairly recent to the commercial market [48]: the first headset with fully-realized VR capabilities—the Oculus Rift—became commercially available in 2016. The VR market has been growing ever since, with VR revenue projected to grow from $12B to a $100B in the next five years [32].

VR systems may collect sensitive data such as facial muscle movements, which can be used to discern users' emotions or quality of health, and high-fidelity infrared images of users' environments [36]. Perhaps most uniquely, VR technology can lead to visceral real-world emotional pain caused by virtual crimes (e.g., physical attacks on virtual characters that the VR user feels they embody) [34], can cause seizures [9], and has been used as a medical device, including for PTSD therapy [41]. While prior work has studied user perceptions of privacy and security for augmented reality [1] (AR) [42, 27, 23, 15], as well as for other IoT devices such as drones [7, 10, 8, 49] and health trackers [39, 33, 28, 51, 26, 40], no such similar examination has focused on VR.

By studying VR early in the technology-adoption lifecycle, we have a unique opportunity to understand security and privacy perceptions and practices as they develop. Unlike fitness trackers, for example, which are already widely adopted, VR headsets only recently became widely available for consumer purchase and only 3% of the US population uses VR monthly [43]. Thus, the current users of VR are "innovators", as defined by Diffusion of Innovations theory [44], willing to explore the uncertain, adopting technologies that are not yet socially supported or pervasive. As privacy and security researchers, we rarely have the opportunity to develop mitigations before problems become wide spread. As such, VR presents a unique opportunity for future research and proactive technological and design solutions.

In this work, we use a mixed-methods approach to form a foundational understanding of human-centered privacy and security risks in VR. We conduct semi-structured interviews with VR users (n=10) and developers (n=10); survey the current state of privacy policies for VR experiences (i.e., applications); and conduct a co-design study with VR developers to create a "code of ethics" for development in VR.

In our interviews, we query users' and developers' *information sources* and *concerns*, especially around security and privacy; *knowledge and perception* of data collection; and VR fits into their *social structures*. Highlights of our find-

[1] AR adds virtual elements to a live view of the real world, may incorporate real-world bystanders into the experience [15], and does not necessarily require the user to wear a headset (e.g., PokemonGo). VR, on the other hand, creates an immersive environment without connection to reality through visual, audio, and haptic experiences transmitted through a VR headset and haptic controllers or even body suits [47].

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

ings include identifying three domains of VR concern: *well-being*, which encompasses both the physical (e.g., motion sickness, vision damage) and psychological (e.g., intensity of experiences, harassment); *privacy*, primarily data collection; and, to a lesser extent, *security*. We also identify a strong emphasis on community. Users describe the VR community as *exclusive and small*, which consequently makes them feel safe, but wary of the future, when the "general public" can afford to use VR. On the other hand, developers describe the community as *small, supportive, and open*, which facilitates knowledge sharing and learning, including discussion of privacy, security, and other ethics-related topics.

One such privacy topic brought up by the developers was privacy policies. The developers we interviewed viewed privacy policies as a method for achieving transparency around data-collection with end-users. Although prior research suggests that privacy policies may be ineffective for achieving this goal [31, 12, 38], to gain a supplemental, objective, understanding of the state of data collection transparency between users and developers we examined VR privacy policies . We randomly sampled 10% of the applications available for HTC Vive and Oculus (the producers of the VR systems our participants used most). Only 30% of the HTC Vive applications we sampled had a privacy policy posted. And, while 82% of the Oculus applications had a posted policy, only 19% of those policies explicitly mentioned VR or VR data. Thus, even if privacy policies were a good method of developer-user transparency, the current ecosystem of VR privacy policies would not achieve this goal.

Our interview results provide one possible hypothesis for the problematic state of VR privacy policies: a lack of standards for developers. The majority of developers with whom we spoke reported struggling to figure out how best to protect end-users. They cited a lack of guidelines in the community as one of the causes for this struggle. As two developers mentioned, "there are no advisors right now," and "there's a quite a big list of unknowns right now in terms of what's best etiquette for a user and what's gonna keep them the most [safe], comfortable, and satisfied." As a first step toward filling this gap, and better aligning the concerns and desired protections expressed by the VR users and developers, we conducted a co-design study open to 11 online communities of developers. In our study, developers from these communities came together, with our moderation, to create a "code of ethics" for VR development. The resulting code includes 10 principles for development in VR, including principles focused on accessibility, maintenance of safe social spaces, and avoiding causing physical or psychological harm to users.

The relative success of our co-design study, and the sustained views and comments on the document after the study period ended, suggest that collaborative work with developer communities may be an effective method for ensuring end-user protection and more secure applications, even beyond VR. Such collaborative processes may be especially successful in small, open communities such as that described by the VR developers we interviewed. While the close-knit nature of the community described by users and developers has benefits—it elicits the users we interviewed feel safer and appear to support developer learning—it can also lead to the exclusion of certain demographic or social groups, risking the development of a technology tailored toward the needs and interests of only those with enough resources to become early adopters. Finally, our results suggest a number of emerging concerns in VR including harassment, transparency about data collection, and security vulnerabilities, which future work should push to address early, before VR becomes more widely adopted.

2. RELATED WORK

We briefly review prior work on VR risks and potential mitigations and on IoT privacy and security, more broadly.

2.1 VR Risks

VR risks to users fall broadly into three categories: data collection and inferences [36, 42]; physical harms [11, 54]; and manipulation and violation of immersive experiences [30, 25]. VR systems collect haptic, audio, and camera inputs that can be used to infer or even treat medical conditions, enhance simulations, and drive profits [36, 41]. Such information may be collected even when the user believes the system is off, as many headsets are "always on", enabling developers to gain data without the users' knowledge [42]. This data may then be sold to third parties [36] or be leaked through known vulnerabilities [30], which may have consequences such as modifying the quality and pricing of goods or services advertised to users.

Finally, O'brolchin et al. theorize that virtual reality social networks will create a 'global village' with stronger discourse and interaction than is available in current social networks [36]. While enhanced community is a great potential benefit of VR, it also increases the risk of users sharing personal and sensitive information with unknown and untrusted third parties or being harassed. VR also enables virtual crimes (e.g., physical attacks on virtual characters, stealing of digital goods), which prior work has found generate strong emotional reactions similar to real-world crimes [34, 50, 25]. To protect against these threats, early work has explored defenses for VR, including specialized authentication systems for 3D environments [55, 18, 5, 6].

While there has been no systematic exploration of risks in VR, Roesner et al. and Lebeck et al. survey the space of AR threats [42, 27]. They point out similar concerns in AR as listed above for VR, in addition to raising concerns about output security: the integrity of the users' virtual experience. Additional work by Denning et al. investigate raises an additional AR concern: bystander effects—the incorporation of a bystander into the virtual experience. While real-world bystander effects are unlikely to occur in virtual reality, virtual avatar representations of users may become bystanders to other users experiences in VR [15]. Finally, Jana et al. work explores methods for fine-grained permissioning in AR, including the development and evaluation of "privacy goggles" that can help users visualize the kinetic data that AR systems can collect about them [23]. The authors of this prior AR work emphasize the importance of addressing AR threats early, before issues occur [42]; we argue that the same can be said of threats in VR—especially given that the more immersive nature of the VR experience presents uniquely different psychological threats as described above.

A key component to identifying and prioritizing the mitigation of VR risks, and developing legislation and policy protections for VR users, is understanding users' and developers' concerns. Only one piece of prior work, to our

knowledge, has explored user privacy and security perceptions around VR: Motti et al. collected online comments about digital glasses and other head-mounted devices (which included a small number of VR headsets) from forums, social media, and various websites [33]. We expand on Motti et al.'s findings, focusing exclusively on VR and collecting more in-depth data than is available through online comments.

2.2 Privacy and Security in IoT

Users' perceptions of privacy and security risks have also been explored in related domains, such as drones and fitness trackers. Prior work has found that people are acutely aware of the privacy and security risks around drones [7, 10, 8] and worry about the potential sale of data collected from their fitness tracking devices [33, 28, 51, 26, 40].

However, despite these concerns, Rader et al. found that fitness tracker users often struggle to consider the broader consequences of inferences that can be made with this data, making it challenging for them to self-manage their privacy around sensor data collection [39]. This finding, together with prior findings that transparent information helps users make more informed decisions [10, 49, 8], underscores the importance of assessing user and developer awareness of risks in VR so that we can increase awareness and provide strategies to mitigate emerging risks.

3. METHODS

In this section we describe our interview methodology and analysis approach, our privacy policy analysis, and our co-design study with VR developers and subsequent trace ethnography analysis. [2] We conclude with a discussion of the limitations of our approach.

3.1 Interview Study

3.1.1 Recruitment

We were interested in studying home consumers of VR and the developers of content for commercially available VR systems. Given the low adoption rate of VR (3%) [43], we do not focus on users or developers for one particular VR platform or application (e.g, Oculus Rift users or VR gamers). We recruited both users and developers by posting advertisements in 17 VR-related Reddit communities, Facebook groups, and online forums (list of communities and advertisement text is included in Appendix A). Participants completed a consent form for the entire study prior to completing short screening questionnaire containing demographic questions and asking them to indicate whether they were a VR user or a developer. To verify that users and developers were authentic in their answers, users were required to upload an image of themselves using their VR headset (they were informed of this requirement in the consent process, images were stored anonymously, and were deleted after the study closed), while developers were required to briefly describe a VR experience they are developing and what language or tools they use to develop.

3.1.2 Protocol

Eligible participants were invited to participate in a 20 minute semi-structured interview via phone, Skype, or Google hangouts, and were compensated with a $15 Amazon gift card for their participation.

We used different protocols for the developers than for the users (see Appendix B for full protocols), however, both protocols covered the same high level topics:

- VR Background. We attempted to capture background information regarding the participants' VR use to better contextualize our findings. This included capturing the VR platform used or developed on (e.g., Oculus Rift, HTC Vive), the VR domain (i.e., what users do with their headsets or the type of experiences developers are creating), participants' goals for using or developing in VR, and evangelizing experiences (i.e., whether the user or developer recommends VR to others).
- Information Sources. For users, how they learned about VR and what heuristics they used to select their VR platform. For developers, how they learned about the possibility of developing for VR and what resources they used to learn necessary development skills.
- Concerns. Our questions about concerns began generally, "Did you have any concerns about starting to [use/develop for] VR? Do you still have concerns?" With follow up questions probing specifically about security concerns or envisioned threats and privacy concerns or threats.
- Data Collection. what data they thought was being or could be collected in VR (for developers what data their experiences collected or could collect), recommendations for others/evangelizing of VR.

Six different researchers conducted the interviews in pairs, researchers of different ethnicities and genders were used to randomize and minimize interviewer biases [37].

3.1.3 Analysis

Each interview was transcribed word-for-word. Then, six researchers reviewed four of the twenty interview transcripts to develop a qualitative codebook. Separate, but similar, codebooks were created for the developer and user interviews. Each interview was double coded: interviews were coded by Researcher 1 and by one of the five other researchers, such that there was a single researcher who had coded every interview for comparison consistency. The researchers achieved an average Krippendorff's alpha of 0.72 across all the transcripts, which is above the minimum suggested threshold for exploratory studies such as this one [29].

3.2 Privacy Policy Analysis

To better understand data collection and transparency between developers and users in VR, we analyzed VR experience privacy policies. To do so, we randomly sampled 10% of the experiences in the "experience stores" for the two commercially available headsets that were used or developed for most by our participants: Oculus Rift/Gear (90 applications) and HTC Vive (50 applications). We labeled the sampled applications for whether they had a posted privacy policy posted and, if so, whether that policy mentioned VR (e.g., VR data, sensors, or experiences). If the policy

[2]The user-study portions of our work were approved by our institutional review board.

mentioned VR, we recorded what was mentioned. Three researchers labeled the sample of 140 applications; as the labeling was objective (had a privacy policy or not; mentioned VR or not) we did not double-label.

3.3 Code of Ethics Co-Design Study

A majority of the developers we interviewed mentioned, unprompted, that they wished for an agreed upon standard of practice or "code of ethics" for development in VR. To fill this gap, we conducted a co-design study in which we invited VR developers and content creators to collaboratively develop a code of ethics for VR development.

3.3.1 Advertisement

To reach developers, we posted in most of the same Facebook, Reddit, and forum communities (11 developer-specific or developer-heavy communities, identified in Appendix A) as we did when recruiting for interview participants. In our post, we briefly described our motivation for our project and then directed readers to the document described below and asked them to help create a VR developer code of ethics (Appendix A). We offered a $2 Amazon gift card to any developer who made a meaningful contribution to the document and emailed us about their contribution.

3.3.2 Document Collaboration

To help the developers get started on the code of ethics, we created a starter document in Google Docs [3], a collaborative document editing platform. The document contained seven potential principles such as "Do No Harm" that emerged from our interview results. We provided the following instructions: "We have placed some ideas for a set of standards for ethical development in VR based on our research findings and the thoughts raised by participant in our research. Please feel free to completely rewrite the standards, discuss your changes in the chat, and etc."

3.3.3 Analysis

To analyze the process by which developers collaborated on creating the code of ethics, we use trace ethnography [17]. Trace ethnography is an extension of document ethnography specifically designed for digital environments and has been used to analyze similar online collaborative processes such as Wikipedia editing [17]. We explore which sections of the document were most edited or commented upon, the different roles of those engaged in developing the document, and the process by which consensus was reached. As is typical of ethnographic research, one of the researchers who was trained in ethnography conducted this portion of the analysis.

3.4 Limitations

In qualitative studies, sufficient sample size and participant diversity are necessary to decrease bias and increase the generalizability of findings. In the interview portion of our study, we conducted interviews until new themes stopped emerging and reaching a sample size within qualitative recommendations [16] for exploratory, foundational studies such as ours. We attempted to ensure that our participants were demographically diverse through demographic screening; however due to bias in the demographics of potential participants who signed up for the study and subsequently attended interviews, and the fact that VR users and developers make up less than 3% of the US population, our sample skews male, more Asian, more educated, and young. Finally, privacy-sensitive participants may have dropped out of the study due to the requirement to upload an image. However, we find that our screening survey drop-out rate (17%) is in line with typical online survey drop-out rates [], suggesting that there was not significant privacy bias.

For the privacy policy portion of our study, we sampled only 10% of the apps in each provider's online store. This led to a maximum margin of error of 7.86%, however it is still possible that the apps we sampled were not representative.

Finally, for the co-design portion of our study, we did not control which developers chose to edit our document nor did we collect any information about those who viewed, shared, or edited the document. Our co-design study, and resulting code of ethics document, is thus biased by those who chose to participate. However, when considering our research within the broader context of the VR community, we felt that it was most important to ensure organic, open participation.

4. RESULTS

In this section we present the results of our three-part study, beginning with a description of the interview participants and findings, followed by the results of our VR privacy policy analysis and, finally, the results of our code of ethics co-design.

4.1 VR Privacy and Security Perceptions

Overall, we find that developers and users express concerns around three classes of risks: well-being, security, and privacy. These concerns vary by their role (developer or user) and, for about half of them, their experiences in the VR community. Figure 1 summarizes the types of risks that users and developers discussed, as well as the background information about their VR use, goals, and community perceptions that we analyze. Figure 2 summarizes the similarities and differences between user and developer concerns: overall we find that developers focus more on well-being—especially physical and psychological—while users focus more on security; both groups mentioned privacy concerns with near equal frequency and emphasis. Neither group was as concerned about security as about well-being and privacy.

4.1.1 Participant Overview

Participant Pool. 98 potential participants completed our demographic screening form. According to the data from the form, sixty-eight were males (69%) and thirty were female (31%). Sixty-three (64%) identified as White, fifteen percent as Asian (15%), eleven as Black or Hispanic (11%), and the remainder as "Other". Ninety-six hold a high school degree or higher (98%) and fifty-nine hold a bachelors degree or higher (60%). Fifty-two are under the age of 29 (53%), forty-two are between 30-49 (43%), and four are over the age of 50 (4%).

Participant Sample. From this sample, we selected 72 participants for interviews, attempting to stratify on age, race, gender, and education to achieve diversity. 20 of those selected attended their interview appointment (see Table 1 for an overview). Sixteen of the participants are male (80%), eleven are White (55%), seven are Asian (35%), and one participant identified as Hispanic (5%) and as Other (5%), re-

[3] https://www.google.com/docs/about/

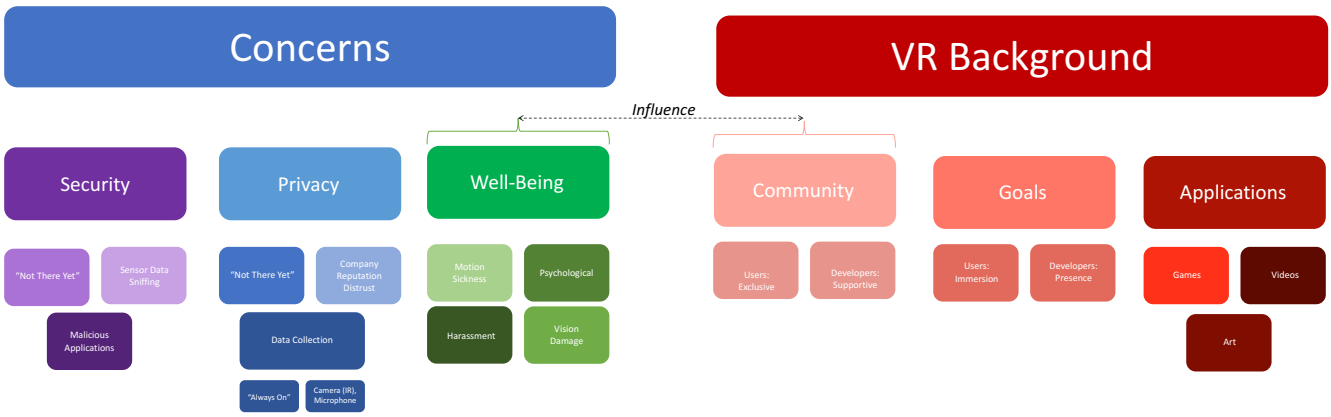


Figure 1: Diagram of the classes of concerns described by users as well as the components of their VR background that we analyze, and in some cases, find influences their concern perceptions.

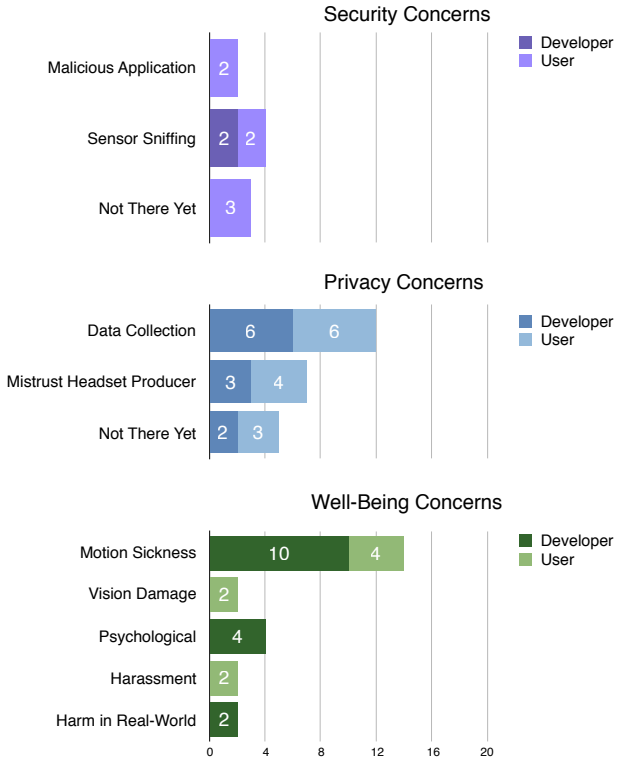


Figure 2: Counts of the number of developers and users who mentioned each type of concern during our interviews.

spectively. All participants hold a high school degree (100%) or higher and ten hold a bachelor degree or higher (50%). Fourteen are under the age of 29 (70%), five are between 30-49 (25%), and one is over the age of 50 (5%).

Representativeness of Participant Pool and Sample. Both our sample of potential participants and our 20 participants are more male than the U.S. population (51% Male) [2], as White as the general population (62% in the U.S. population), more Asian (8% in the U.S.), more educated (87% hold a high school degree or higher and 30.3% hold a bachelors degree or higher), and younger (40% are under the age of 29, 26% are between 30-49, and 34% are over the age of 50 in the U.S. [1]).

ID	Sex	Age	Race	Educ.	Plat.	App.
U1	M	50-59	W	SC	C/O	V
U2	M	30-39	W	SC	O	V/G/S
U3	M	40-49	W	B.S.	O	V/G/A
U4	F	18-29	A	SC	H	O
U5	M	18-29	A	B.S.	O	V/G
U6	M	30-39	A	>B.S.	O	G
U7	F	30-39	A	B.S.	C/H/O	V/G/A
U8	M	18-29	W	SC	H/O	V
U9	F	18-29	W	B.S.	O	G
U10	M	18-29	A	B.S.	O	V
D1	M	18-29	W	SC	O	G
D2	M	18-29	A	H.S.	O	A
D3	M	18-29	W	SC	O	G
D4	F	18-29	H	SC	O	S
D5	M	18-29	O	B.S	H/O	G
D6	M	18-29	W	H.S.	H	G
D7	M	40-49	W	>B.S.	H/O	A/E
D8	M	18-29	A	SC	H	Other
D9	M	18-29	W	>B.S.	H/O	G
D10	M	18-29	W	B.S.	O	A

Table 1: Participant Demographics. Educ. is education level, Plat. is VR platform(s) used or developed for (O: Oculus Rift or Gear, H: HTC Vive, C: Google Cardboard), and App. is VR application types used or developed (V: video, G: games, A: art, S: social, E: education).

VR Platform and Usage. Finally, nine of our ten users reported using an Oculus product, either the Rift or Gear, while the other used an HTC Vive. Of the nine Oculus users, three also used other platforms, two a Google Cardboard and two the HTC Vive. Four users used their headsets for multiple applications, while the other six used their headset for multiple uses. In total, seven users used their headsets to watch videos, six used the headset to play games, and two used the headset for art experiences. Eight of the developers developed for Oculus Rift and Gear, three of whom also developed for HTC vive, the other two developers developed only for HTC Vive. Five reported developing games, three reported creating interactive art or videos, one reported creating a social application, one reported creating an educational tool, and the final developer reported creating a work-specific simulation.

4.1.2 Developer Interview Results

Goals of VR development center around presence. The VR developers that were interviewed were creating a variety of experiences. Across these varied domains, the majority of developers (six) mentioned that their primary goal when developing was to facilitate and ensure a sense of "presence." For example, D9 says, "you have to focus all your design actions on making sure that you don't break the state of presence and you add to the immersive experience." D4 notes that well-being and security are closely intertwined with this goal: "motion sickness breaks presence, while presence enhances risks of psychological threat or someone screwing with people by getting into the virtual environment."

VR developer community is strong and enhances learning. When asked how they learned to develop for VR, five developers reported using online tutorials and three reported signing up for a more structured online bootcamp or course. The other two developers, as well as six of the eight who also reported learning from only tutorials or bootcamps, also mentioned asking questions to other VR developers in various online communities. More broadly, four of the developers, without prompting, mentioned the strength of the VR community. For example, D10 says, "you could fit them all in a small room really, and it's really close, it's really tight knit. I actually became close friends with most of them even still to this day I consider them almost family in a way. And yeah, I have been developing for VR ever since [meeting other VR developers online]." Similarly, D7 describes an open, supportive community, "we are still in the phase in the industry where people are very open and willing to help each other and that's a huge blessing because we are still you know its still evolving and...instead of like hoarding knowledge, we need to sort of cross develop."

Concerns for user well-being encompass the physical and psychological. Developers' concerns for their users often focused on well-being. All of the developers raised concerns about motion sickness. For example, D6 says, "motion sickness isn't really a concern at all when developing for a game you are going to play on a computer screen. But when you're looking at VR, [motion sickness] is...a driving factor, you could say, in development."

Additionally two developers raised concerns about participants being unaware of danger in their real-world physical environment (e.g., not hearing fire alarm, bumping into objects as a result of game play). On this point, D1 says, "the biggest issue is probably letting people know that there needs to be a specific and safe environment to use VR. Because obviously you're not interacting with the rest of the world [while you're in VR]...[this is an] issue that I don't actually see very many people looking into. [For example,] say that you're in VR and a alarm fire goes off in the building, who is to say that you actually are going to hear that alarm...this disconnect from the actual world and the VR experience is a definite issue."

Four developers mentioned concerns with the psychological well-being of their users. D9 and D4 mention that harms (e.g., bullying or intentional scary moments) in VR may feel more realistic and thus may be more traumatizing. D9 explains, "VR is a very personal, intimate situation and when you wear a VR headset...you really believe it, it's really immersive. So if someone harms you in VR—either verbally or something else—you will also feel that after taking off the headset." Similarly, D4 says, "VR content is a lot more impactful...the feeling of like being scared in VR is much more real. Because everything does feel so real and so tactile, so you have to be extra careful with the content you are introducing to people." D8 and D5 express similar concerns, and also raise a connection between psychological swell-being and security—potential psychological harms that may come from a malicious entity being able to alter the VR experience. D8 says, "I think that it's on the developer to try and limit the user to being able to only experience what the developer was intending for them experience in the first place."

Developers mention privacy concerns about variety of issues. Six developers mention privacy concerns about VR, but none of them feel these concerns are relevant for their own products. Two mention concerns with the fact that the headsets are "always on" and users could be unaware of the data collection that is happening when they are not using their headset. Three others expressed concern about the ability of the headset to use camera sensors to detect users locations or to access the microphone in their space: "what somebody is doing while in VR is recordable and trackable on a different level" than on other digital platforms, which is something you "have to acknowledge," D8 notes.

Three developers mentioned privacy concerns specifically related to Facebook's ownership of Oculus and the developer's perception of Facebook's reputation around privacy issues. For example, D10 remarks on his perception of Facebook's attitude toward privacy, "they are not afraid to manipulate to see if you're happy or sad, they are not afraid to get caught, in the end, it's all about the money to them. It's not about these altruistic goals and that is definitely one of my biggest concerns hands down. That's why you know, Facebook acquired Oculus, so they could get a monopoly over the next form of advertising and control media and connecting people." D7 expresses a similar sentiment: "I think Facebook is pouring money into VR because it is going to generate this kind of super personal data that will help create a biological map or biological key, of who their users are."

On the other hand, two developers felt that VR "was not there yet" to worry about privacy. D4 likens VR to the early days of the Internet, "remember the beginning of the Internet and chat rooms, [in VR] potential issues haven't been addressed yet because it hasn't happened yet." The final two developers did not explicitly comment on privacy, despite being prompted.

Developers suggest permission requests to mitigate privacy concerns, yet no such capability exists for most headsets. Four developers suggest that using permission requests could help mitigate privacy issues for end users. For example, D8 recommends that VR should do something "identical to current privacy methodologies in terms of your requesting permission from the end user ahead of time." However, no desktop VR applications include any such permission requests (the Samsung Gear VR which runs off a Samsung phone does include permission requests, although it is unclear from the documentation whether there are per-

mission requests made for e.g., camera sensors on the VR headset rather than phone sensors). D9 also recommends adding permission requests within the VR environment, but notes that this may be difficult to design because there is no single view point (e.g., screen) in VR: "if you want to [request] some information from the player you cannot simply display it on the screen because it is not there."

Finally, five developers recommend using privacy policies rather than permissions to help users make informed privacy and data collection choices. However, as summarized in Section 4.2 we find that, currently, few VR applications offer privacy policies that discuss VR data collection.

Little mention of security. Overall, when discussing potential concerns about VR, the developers we interviewed spoke the least about security. Two developers mentioned security concerns for health applications. For example, D6 says they would be concerned about security for applications used for "medical or for education [purposes]." For those applications, he says, "the issues with hacking are more serious. I think [that is where] protecting data [would be] important. But you know, there has got to be some serious push for that or else there would be no incentives...to do that right."

Two other developers mention security, but they reference passing off security responsibilities to others. For example, D5 explains that they use Unity and Google cloud storage to store data collected from the user. When asked about security they explain that they do so in part because, "it means that we don't have to deal with securing information ourselves. It makes it their problem and not ours."

Developers appear to take concerns about users' privacy and well-being on themselves. While these two developers passed off security responsibility to the storage services, it is interesting to note that no developers mentioned "passing-the-buck" for well-being or privacy. For example, while the OS would typically be key for managing permission requests and ensuring that information was shared only when it should be, no developers mentioned the responsibility of Windows- or Android-system developers to mitigate vulnerabilities and enforce permissions. (It is possible that the four developers who mentioned permissions meant to imply this.) More generally developers appeared to take responsibility for end-user privacy and well-being onto themselves, never mentioning that Oculus, HTC Vive, Windows, or Android could make improvements to address such issues. "it's gonna be something that developers have to...keep an eye on and implement" (D8).

Marked disconnect between developers' general security and privacy concerns and concerns about their own products. However, despite feelings of responsibility, there seems to be marked disconnects between developer's privacy- and security-related concerns for VR users in general and the privacy and security risks they see in their own products. While most of the developers who mention well-being concerns also mention working to mitigate these concerns for their own projects, the majority of those who mention privacy and security risks (5/6 for privacy and 2/2 for security) do not see privacy and security risks with their own products. For example, even D9, who raised security concerns and is working on an application that infers users health conditions via sensor data and then provides VR-based treatments says, "yeah I actually [can]not think of a privacy or security problem...[with my product] maybe it's an issue in the future."

Developers express desire for standards and ethical guidance. Finally, and perhaps relatedly, D5 notes that no one in the VR development community "is experienced" or is "an advisor" about ethics and privacy and security issues: "just the fact of the matter is there are no VR power users. You know I can count on the number of fingers the number of experienced 'devs' I've actually met." This lack of guidance, he says, makes it hard to "know where the right line is." Five other developers expressed similar sentiments, with D8 explaining "there's a quite a big list of unknowns right now in terms of what's best etiquette for a user and what's gonna keep them the most comfortable and satisfied in the experience. That has already been hashed out for web development over the last couple of decades [but not in VR]...[the VR] industry needs to start using standards."

Thus, D10 suggests that a "mantra" or code of ethics is needed, and suggests that the big companies are not going to step in, so an emphasis on ethics will need to come from developers themselves. He explains, "I would encourage developers to be transparent and to just not talk down to customers, don't treat them as numbers. Don't do onto others what you wouldn't want onto you. I'd like to have a mantra like that. And just because Facebook does something different doesn't mean I or anybody else [in the community] has to do that." Thus, we hypothesize that the disconnect between developer's general concerns for VR users and concerns about their own products may, in part, arise from inconsistent or ill-defined guidance about what should be a concern and what needs to be addressed in development. As explored more closely in Section 4.3 we take a first step toward helping the VR developer community define their desired code of ethics via a co-design study, resulting in the code of ethics shown in Figure 5.

4.1.3 User Interview Results

Immersiveness and, to a lesser extent addictiveness, are key user desires. We find that developer and user goals seem to be in alignment: the developers we spoke with focused their development around achieving "presence" and, similarly, we find that most users (7/10) sought out VR for the "immersiveness of the experience." As U8 explains, "I think it's really just about being immersed in a different type of world...As opposed to a TV where I can constantly be distracted. When I'm in VR...100 percent of my attention is dedicated to that because I can't text or I can't just multitask." Two users also mentioned that they *wanted* VR to be addictive: "VR needs that addicting label...those features that keep people going back again and again and again" (U4).

Majority of users learn about VR online. When asked how they learned about VR and selected the VR platform they use, users mentioned three primary information sources: friends and family (two users), the Internet (seven users), and school (one user). Of those who mentioned learning about VR from the Internet, three specifically mentioned learning about VR from Reddit: "Reddit is awesome for anything" U1 says. As U6 explains, one of the ways he learned about VR is by being a "group reddit with some friends and

stuff". Further, after learning about VR, three users, including two of whom did *not* learn about VR from Reddit, reported relying on Reddit reviews to select their headset: "you read a 100 posts on something you're gonna get a good gauge on whether a product is good bad or otherwise" (U1).

Users' concerns about well-being are focused on the physical. Four users expressed concern about motion sickness in VR. However, their concerns, are more muted than developers were. U2 explains, "it happens to some people and doesn't happen to some others...it's basically an individual kind of thing so I just had an awareness...[it was] not so much a problem to be weighed." Two users also expressed a different type of physical concern, not considered by developers: vision deterioration from VR use.

Only one user—U2 who also brought up "awareness" of motion sickness—brings up concerns around psychological harms. He worries about other users who "aren't mentally as strong," elaborating, "some people don't have the mind to handle things...I'm sure if you put a soldier into VR and play the wrong experience like Call of Duty or Battlefield or something like that. That could trigger some sort of... flashback or bipolar moment...really, what VR is trying to do here is duplicate reality where it tricks your mind into feeling like you are somewhere else. Some people might not be ready for something like that, some people might not be mentally developed enough to take something like that and not be messed up over it, you know?"

Finally, two other users bringing up a different type of well-being concern: cyberbullying / harassment. U8 expresses concern about harassment in the future, "For me, VR has just been pretty much stand alone. I haven't interacted with others in VR...[interacting with others] is, you know, a big concern. The type of people who are online: spouting racism, sexism. I mean if they have ability to [use] VR they're probably going to...know it will [become] like any message board."

More users than developers raised security concerns. Seven users raised concerns about security: four raised current concerns while the other three raised concerns about the future. U1 and U5 expressed concerns about the security of applications the experience stores. U1 says, "as soon as I moved over to the Gear [4], I didn't have as much concern as with the you know the old Cardboard glasses [5] where third parties could produce content that I could see. I'm aware of the concerns with vulnerabilities in those applications. I'm much more comfortable you know going through the Oculus store for the Gear [because] they do all the vetting and stuff up front." U6 and U10 raise concerns about malicious attackers modifying their virtual experience or gaining access to headset sensors. U10 believes that "someone could hack into your systems network...take control of your centers and using his camera to spy on you", but is "not really concerned about that". Similar to U10, U6 acknowledges that "there are different hacks you can do to change the game to have someone password or whatever".

U2, U4, and U7, on the other hand, are not concerned about security now, but would be concerned in the future as the VR industry expands. U4 says, "if VR gets the chance that it needs...that's when you're going to get to...worrying about hackers altering your experience. What's going to be crazy is at that point...[is] just like your buddy can pick up your phone and post on your Facebook, and everybody thinks it's you...someone can put on a VR head unit and go into a virtual world assuming your identity. I think that identity theft, if [VR] becomes mainstream, will become rampant." More generally, U7 explains, "I'm sure someone will figure out a way to exploit what we do. For now, everything is still new...we still haven't even figured out typing in VR. Like I feel like someone needs to invent technology [to monetize VR]...when people actually start making money in VR," that is when she thinks issues will arise.

Figure 3: Image of Oculus Rift user's real-world environment captured by the infrared sensors on the headset [3].

Users worry most about microphone and infrared camera data collection. Six users expressed concern about privacy related to data collection. All six focused on microphones or infrared sensors in the headsets collecting data because these sensors are "always on, which I find is weird" (U6). U5 says, "the Rift actually has a microphone in it...[so I realized] oh crap people can hear me...I've [also] seen somebody who posted a picture of what the sensors actually picked up and it was a pretty clear view of the room and what not" (see Figure 3 for an example of an infrared image captured by the sensors on an Oculus Rift).

Two users mentioned knowing that their headset collected this type of data about them, but said there was no reason to be worried unless "you were up to no good" (U7). For example, U2 explains, "if you're worried about something, you're up to something you shouldn't be doing. As far as what these things are going to collect, yeah you know...they could be collecting something...[but unless] you're doing something bad...what could they be collecting?"

Similar to security, three users felt that they would have more concerns about privacy in the future, but VR was not there yet. U8 explains, "I don't think there's probably anything. Because I'm just playing you know these little games...I think [privacy's] going to be a big concern of mine going forward especially when you know VR is more mainstream and more affordable."

Users raise privacy issues around headset producer reputation. Just as three developers raised concerns about privacy in the Oculus products due to the reputation of Face-

[4]The Samsung Gear is a headset produced by Oculus, which is owned by Facebook.
[5]U1 is referring to the Google Cardboard headset.

book's approach toward privacy for their other services, four users raise similar concerns. U3 explains that these concerns about reputation are, "one of the reasons that I didn't install the Facebook app within virtual reality...it can read and write your contacts, it can call phones, it can send data to whoever whenever without you knowing." Similarly, U5 worries based on what he's heard in the news about Facebook, "considering that Oculus Rift is owned by Facebook, I [am] concerned...you know Facebook has been in the news recently about just how much information they pick up based on your habit, posting activities and other things like that."

Users vary in their comparative perception of privacy risk in VR. Overall, four users felt they were at more privacy risk on VR than on other platforms, four felt that VR exposed them to the same level of risk as any other platform, and two felt that less data was collected about them on VR than elsewhere. U6 explains that he feels VR is the same as anything else because, "I've reached a point where I guess it's pessimism. Where I realize you know there's all these data breaches and hacks, you know, all of our information is out there so that after I got over that concern you know I just learned not to stress too much about it...So, I kind of took a pessimistic view towards privacy that way and I realize hey, they already have this information."

Users perceive the VR community as exclusive and, consequently, safe. Interestingly, four participants, unprompted, describe the community of other users on VR. They describe their community of peers as an exclusive one, which requires money and technical savvy to get in. For example, U4 says, there's a "high entry barrier to even get started in VR. Usually it's pretty high. You know people with disposable income and who are you know tech oriented. It's not just you know [anyone] typing on a keyboard."

Similarly, U2 describes the typical user he meets in VR as "somebody who has a lot of money and has a premium setup you know...I mean you are talking people with 4 plus sensors." This sense of exclusivity makes these four users feel safe, especially from well-being concerns around harassment. For example, U4 continues her above comment to explain that she will be more concerned about virtual crimes and bullying once VR becomes more accessible to the "general public." Similarly, U2 continues, "people in virtual reality are a lot more open, a lot nicer, they're a lot more accepting. You know, online, some people can be really rude, some people can be helpful, some people can be just annoying. I found that in VR you kind of bring back that element where you feel like you are looking at somebody in the face...The way that the market is right now, there is a specific group of people that are using these devices. So, it makes for interesting conversation. Usually the people you would meet online are not on Reddit. But, if I play with you on big screen [e.g., VR] most likely you would be on Reddit because there's a certain type of crowd that's really into this, you know?"

Some users evangelize VR, even buying headsets for others. Three of the users with whom we spoke specifically mentioned evangelizing VR to others. U6 says, "Oh I've already recommended to every person that came over to my house I've already brought my rig up to my parents to let them just play with it...I would recommend it to anybody." U2 even bought multiple headsets—and gave a headset to a friend—so that he could use VR with others and so that they could "experience the magic." Part of his motivation for doing so is social, he says, "one thing I noticed about virtual reality that kind of sucks is it can be a very solo, anti-social experience if you think about it...what you end up having is one person with glasses saying oh wow wow and everybody else is sitting there scratching their head like okay, hopefully I can try that in a few minutes. [I] found that the most you can get from these things will be when you actually link up a couple units. The social experience makes the entire thing a completely different game changer. When you are doing it with a couple other people, the social aspect completely turns VR into a totally different animal."

Two more users mention more tempered evangelizing, saying, "Like I think everyone should try it. I don't think everyone should necessarily buy it" (U7) and raising concerns around making recommendations too broadly because VR is so expensive. Finally, two users mention explicitly not recommending VR to others. U1 explains, " I'm certainly not the evangelical type to say oh you have to like it...I let them know it's out there and what's available and what the future holds" but, he says, some people get sick or don't have the right depth perception to make it right for them.

4.2 VR Privacy Policies

Overall, we find that 82% (74) of the Oculus experiences and 30% (15) of HTC Vive applications have a privacy policy posted on the page from which users can download the experience. [6] Of these privacy policies, 19% (14 of 74) of the Oculus policies mentioned VR or VR-specific data collection; 33% (5 of 15) of the HTC Vive policies did the same.

Some policies that did mention VR or VR-specific data provided vague descriptions (4 of 14 Oculus, none of the HTC Vive applications), refering the reader to the Unity or Oculus privacy policies or state that they will "collect the data you share with us" with no further detail. Seven of the 14 Oculus policies and 3 of the 5 Vive policies stated that they would only collect personal information such as the user's email address, billing code, phone number, and etc. Four of the Oculus policies explicitly mention inferring the user's movements, for example the *Virtual Desktop* privacy policy states, "Information about your physical movements and dimensions when you use a virtual reality headset." *Sprint Vector*'s policy spoke more broadly about biometrics and biofeedback, saying, "We may collect biometric and biofeedback information relating to your use of the Services, including information about your physical movements and dimensions when you use a virtual reality headset."

Finally, one Oculus policy and two of the five Vive policies warn that the experience captures audio or IR camera data. For example, *MetaTable Poker*'s policy explains that once you join the application, your microphone will be activated and everything you do and say will be transmitted to every player in room (and will be stored by the application).

4.3 Code of Ethics Co-Design

Our code of ethics document received 1053 views from our posts to 11 online communities. Of these viewers, we anticipate that 245 were able to potentially make an edit. The

[6] See `https://www.oculus.com/experiences/rift/733640976736718/` for an example page from which an experience can be downloaded.

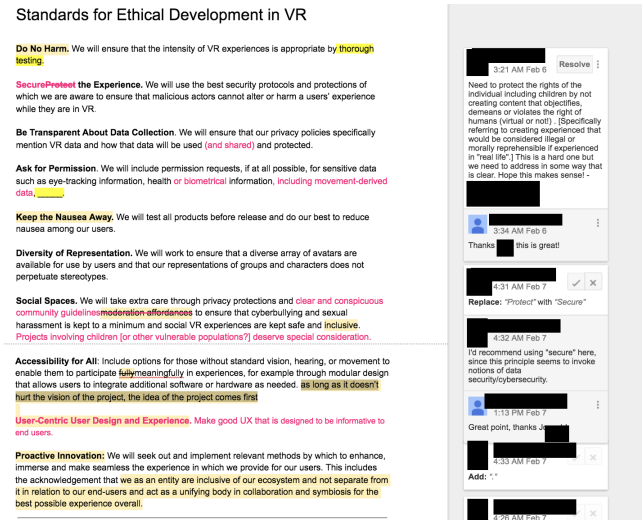


Figure 4: A screenshot of the code of ethics document about three-quarters of the way through the co-design process. Contributor names and researcher names have been blinded from the figure.

remaining viewers were on a mobile device, on which it is only possible to edit a Google doc if the app is installed and even then it takes multiple additional clicks to edit. Of these 245 people that we estimate could make an edit, 19 people made contributions—edits, comments, or additions of new sections—to the document. Figure 4 shows a screen-shot of the document about three-quarters of the way through contributions being added. Interestingly, contributions were made only asynchronously: no participants used the chat feature in GoogleDocs [7]. This may be because they did not want to communicate synchronously or because the chat was difficult to locate.

An additional seven people were "sharers"—people in the communities who indicated that they did not edit themselves, but were passing along the document to specific people to ask them to edit (or promote). Thus, we observe a phenomena similar to that observed in other editing scenarios such as Wikipedia editing: a large proportion (approx. 90% in our study) of lurkers in comparison with a small proportion of active editors [35].

Interestingly, while we offered a $2 incentive to those who made a contribution to the document, only one of the 19 contributors requested their incentive. We hypothesize that this may be due to the type of people choosing to contribute (those concerned about ethics may be more altruistic) or out of concern about anonymity (this hypothesis is less supported, as 10 of the 19 contributors revealed their names).

The initial code of ethics that we proposed had seven principles (the first seven shown in Figure 4). The developers contributing as part of our co-design modified the title and body for all but one of our proposed principles (Diversity of Representation was untouched). The contributors also added three additional principles: Accessibility for All, User-Centric User Design and Experience, and Proactive Innovation; all of which were subsequently edited or commented

[7] `https://support.google.com/docs/answer/2494891?co=GENIE.Platform%3DDesktop&hl=en`

Standards for Ethical Development in VR

Do No Harm. We will ensure that the intensity of VR experiences, and effects caused (e.g., seizure risk from flashing lights) is appropriate by thorough testing. Avoid creating content that objectifies, demeans or violates the rights of humans or animals (e.g., creating experiences considered illegal or morally reprehensible if experienced in "real life").

Secure the Experience. We will use the best security protocols and protections of which we are aware to ensure that malicious actors cannot alter or harm a users' experience while they are in VR.

Be Transparent About Data Collection. We will ensure that our privacy policies specifically mention VR data and how that data will be used (and shared) and protected.

Ask for Permission. We will include permission requests, if at all possible, for sensitive data such as eye-tracking information, health or biometrical information, including movement-derived data.

Keep the Nausea Away. We will test all products before release and do our best to reduce nausea among our users.

Diversity of Representation. We will work to ensure that a diverse array of avatars are available for use by users and that our representations of groups and characters does not perpetuate stereotypes.

Social Spaces. We will take extra care through privacy protections and clear and conspicuous community guidelines to ensure that cyberbullying and sexual harassment is kept to a minimum and social VR experiences are kept safe and inclusive. Projects involving children or other vulnerable populations deserve special consideration.

Accessibility for All: Include options for those without standard vision, hearing, or movement to enable them to participate meaningfully in experiences, for example through modular design that allows users to integrate additional software or hardware as needed.

User-Centric User Design and Experience. Make good UX that is designed to be informative to end-users.

Proactive Innovation: We will seek out and implement new methods to enhance the immersive and seamless experience we provide to our users. We will not consider end-users as entirely separate: we will act in collaboration and symbiosis with them to achieve the best possible experience overall.

Figure 5: The final VR developer code of ethics.

on by contributors other than the ones who proposed them.

The majority of contributions were edits (29 in total). Sometimes, edits were briefly explained—for example, as shown in Figure 4 one contributor changed "Protect the Experience" to "Secure the Experience" because they felt that "Secure" more clearly indicated a cybersecurity focus. The Social Spaces, Accessibility for All, and Ask Permission principles were the most edited, with six contributors editing Social Spaces and four contributors editing the other two, respectively. Each of the other sections had at least two edits. There were also 11 comments left on the document, with most sections receiving one or two comments.

After the 29 edits and 11 comments made by 19 contributors [8], the code of ethics shown in Figure 5 was produced.

Below, we present a case study of the process through which one of these new principles—Accessibility for All—was developed. We conclude with a discussion of future use of our "Ethical Co-Design" method.

Case Study: Developing the "Accessibility for All" principle. One developer added the Accessibility for All section, after feeling that there was not enough emphasis on inclusivity in the existing code. She commented on the word inclusivity (originally in the Social Spaces principle), saying, "need to add [inclusivity] as a different heading and not un-

[8] Five days after the last activity on the document we accepted outstanding edits and made small editorial corrections for reading ease.

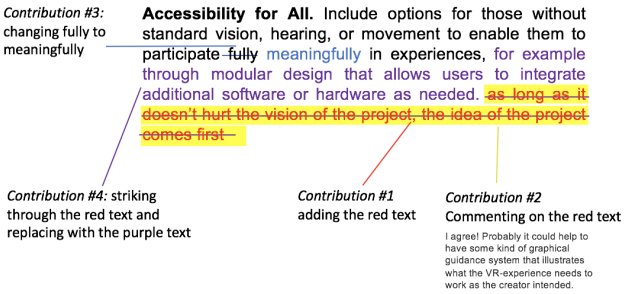


Figure 6: A diagram of the editing process for the Accessibility for All principle in the code of ethics document.

der Social Spaces: "Accessibility for all", having options for those without standard vision, hearing, or movement. (If you don't add this in from the beginning, then we will be having to kludge that in afterwards, and it just don't work well.)" She then added an Accessibility for All section. Subsequently, four other contributors commented on or edited the section. We diagram the changes in Figure 6.

The first contributor added a sentence weakening the proposed section, suggesting that inclusivity should not come before the vision of the product. The second contribution was a comment following up on this edit, agreeing, and, as we interpret it, suggesting that people with disabilities could be told what extra hardware they would need to get the same experience—seemingly a compromise between the original contribution and the edit. The third contribution was an edit, changing meaningfully to fully. This contribution also appears to have been executed in response to the conversation about the added line from contribution #1. This third contribution was explained by the contributor as follows: "I'd amend this to "meaningfully" instead of fully. Substitution of faculty is only a substitution and not a full restoration. "meaningful" interaction that facilitates their involvement and acknowledges their particular needs is more appropriate - even if [special] hardware to interface as required without specifically tailoring the experience around their needs. This would be a better approach (I believe) and therefore address the need for maintaining the vision of said project without negating [inclusion] at any point." Finally, a fourth contribution was made, which synthesized the comments from the second and third contributions—the fourth contributor removed the red line added by contributor one and added a more moderate statement recommending modular design so that users with accessibility needs could add hardware (as suggested by contributor #2's comment) to obtain a similarly meaningful experience.

Reflections and Recommendations for Future Work Using Ethical Co-Design. In sum, our work offers the first proof-of-concept for the method of *ethical co-design*, in which people (in this case developers) collaborate and reach agreement on an ethical topic (in this case, a code of ethics). We note findings of interest from our first implementation. Although a chat system was offered to participants, all collaborations took place primarily asynchronously. While discussions became heated at times – an interesting sign of engagement / care – contributors appeared to exhibit respect for each other and consistently worked to incorporate otherâĂŹs comments or intentions into their future revisions, never entirely destroying or ignoring a previous edit. Although we offered compensation, only one participant requested their reimbursement, perhaps suggesting that more altruistic people are more likely to participate in such studies.

Finally, while ethics were, at times, controversial – as exemplified in the Accessibility for All case study – in all cases consensus was reached. Combined with with recent work on algorithmic fairness showing that people may have a "common fairness mapping" [19], this suggests that people may have commonly shared ethical views. Ethical co-design may help "non-expert" or unheard stakeholders express these views, which may otherwise remain unconsidered in favor of the normative decisions of more powerful stakeholders.

However, ethical co-design is not without pitfalls. Participants may not always have a *good* code of ethics, even if they have a consistent one. For example, in this work we found that multiple participants reported a desire to exclude people different from them in order to maintain safety. If they used such views to inform a standard of ethics, the result may be harmful. Thus, researchers must take care before blindly applying such methods, and should consider vetting co-designed codes of ethics with panels of experts in a relevant domain.

5. DISCUSSION & FUTURE DIRECTIONS

Below, we highlight takeaways and areas of future work.

Collaboration with Developer Communities May Improve Application Privacy and Security. In our code of ethics co-design study we found engagement levels typical of Wikipedia editing communities and observed that VR developers were able to effectively work together to reach consensus on a code of ethics. Our interview results suggest that VR developers rely on each other, through the small and supportive community they describe, to figure out how best to build applications for end-users, including how to secure applications, respect user privacy, and ensure well-being. VR developers do not appear to seek out this guidance from the companies creating the platforms, as some developers express distrust of the headset producers, with one developer saying, for example, "Don't do onto others what you wouldn't want onto you...just because Facebook does something...doesn't mean [I] have to do that."

The success of our co-design study and developers' sustained engagement with the document (100 views every three days, plus additional shares, since the study period ended) suggests that collaborative work with developers, such as future co-design work for additional standards and/or training of security or privacy peer advocates (such as those in workplaces [21])—who could provide guidance to their peers on how to design affordances for privacy and well-being or technical advice for avoiding code insecurities—may be an effective methods for improving applications for end-users.

While other prior work has similarly investigated how developers make ethical decisions in different domains [46, 45, 20], neither the work presented here nor this prior work has moved from inquiry to action: using collaboration and social influence with developers to drive privacy and security improvements. There is, however, support that such an approach may be useful, as social influence has been shown to be effective for promoting security behavior among

end-users [13]. Thus, future work may wish to investigate whether such strong communities exist for other types of development (e.g., IoT apps, certain types of mobile phone application development) and, if so, how to leverage collaborative interventions with these communities to solve developer-driven security and privacy issues raised by prior work [4, 52].

Users' Threat Model Includes Exclusivity of Community. Our results underscore a strong role of community not only for developers but also for users. Four of the users we interviewed described the VR community as small, exclusive, and consequently: safe. They mentioned that they would start to have concerns about security, privacy, or harassment later on but they were not currently concerned because the community was "nice" and the "people aren't like that."

While the close-knit nature of the user community makes users feel safer, such exclusivity has a downside: lack of diversity (as observed at a small scale in the demographic bias of the pool of potential participants recruited for our interview study). Lack of diversity among technology "innovators" is a known problem and exacerbates the digital divide [53, 22]: if no "innovators" from particular social groups are present, concerns these groups have may not be identified or addressed, applications of the technology that are of interest to these groups may not be developed, and these groups do not have an influencer to expand adoption of the technology in their community.

Further, the attitude expressed by some of the VR users in our study—that they did not want the "general population" to begin using VR—suggests that expanding the groups using VR may be difficult not just due to problems of access, but also due to exclusionary attitudes and narrow perceptions of who the users of VR are or should be (e.g., one user explained "the way that the market is right now, there is a specific group of people that are using these devices...Usually the people you would meet online are not on Reddit. But, if I play with you on big screen [e.g., VR] most likely you would be on Reddit because there's a certain type of crowd that's really into this, you know?").This desire for exclusivity among users contrasts with the emphasis that developers in our co-design study placed on Accessibility for All (accommodating those with disabilities) and Diversity of Representation (offering diverse avatars).

To increase the diversity of early adopters of VR, producers of VR headsets or technology activism organizations may wish to place VR booths or arcades in communities to enable access to those who cannot purchase a headset for their home or may consider providing headsets to Beta testers (e.g., "influencers") who sign up within communities with no adopters [24]. Future work may wish to explore the efficacy of such approaches and investigate the risks and experiences of populations who were not well represented in this study.

Looking Forward: Harassment, Security, and Policy. Overall, we find developers to be more focused on and concerned about well-being, including both motion sickness and psychological well-being (e.g., insuring that experiences are not too intense) than users, perhaps because developers are doing a good job at mitigating these issues. However, as new developers join it will be important to ensure that addressing these well-being related facets of VR risk remains a high priority, as emphasized by a recent news piece on one user's traumatic seizure in VR [9].

We find that one well-being risk not mentioned by developers is harassment. Only users mention any harassment concerns, suggesting that such concerns may be an emerging issue especially with increasing adoption of VR and increasing release of social applications.

Additionally, very few developers, and relatively few users, expressed security concerns – many explaining that VR did not have a big enough user base and was not monetized enough to be concerned. Given this attitude, it is likely that many early VR applications will have a number of security vulnerabilities and that vulnerabilities may increase with accelerating adoption. Raising developer awareness about potential problems early, which may require additional VR-focused research similar Roesner et al.'s work on AR threats [42] and can perhaps be achieved through approaches like those discussed above, may help stop problems before VR becomes a more enticing target for attackers.

Both users and developers did raise privacy concerns. Developers primarily suggested mitigating concerns around data collection (the majority of privacy concerns expressed by both groups) through "notice and choice": that is, the use of privacy policies. However, our findings show that VR privacy policies are currently lacking – either not posted or not mentioning VR data (e.g., what data they are collecting) – and prior work shows that privacy policies are hard for users to read and largely ineffective [31, 12, 38]. Further, as one developer in our study noted, desktop VR does not currently use permissions, in part because of the difficulty of presenting a permission screen in the virtual environment. Future work may wish to expand beyond exploring VR authentication [55, 18, 5, 6] to also consider permissions and data transparency solutions.

Finally, the application developers with whom we spoke felt that they needed to lead and take responsibility for addressing risks to end-users. This emphasis seemed largely to be due to concern with the reputation of the developers of the headsets, which was expressed by both users and developers. While it is important for application developers to be part of the ecosystem designed to keep users safe, future work may wish to explore policy and system-level guidelines, especially for privacy policies (GDPR legislation which explicitly requires companies to discuss the type of data they are collecting, its' uses and a justification for that use, and a way to opt out of many of the data collection types may be a step in the right direction) and medical and educational applications that touch on HIPPA- or FERPA-regulated data, such that the burden of protecting users does not fall only on application developers.

In sum, our initial results are encouraging: developers express significant concern about end-user risks and exhibited an interest in engaging in developing solutions in our co-design study. But, our results also underscore that issues around harassment and security may be coming, especially as many developers exhibit a disconnect between identifying general concerns for users and concerns with their own products and many users feel that they do not yet need to worry.

6. REFERENCES

[1] U.S. Census Bureau Age facts. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_16_5YR_S0101&prodType=table.

[2] U.S. Census Bureau QuickFacts. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_16_5YR_DP05&src=pt.

[3] What does the cv1 sensor/camera see?

[4] Y. Acar, S. Fahl, and M. L. Mazurek. You are not your developer, either: A research agenda for usable security and privacy research beyond end users. In *SecDev*. IEEE, 2016.

[5] M. Agarwal, M. Mehra, R. Pawar, and D. Shah. Secure authentication using dynamic virtual keyboard layout. In *ICWET*. ACM, 2011.

[6] F. A. Alsulaiman and A. El Saddik. A novel 3d graphical password schema. In *VECIMS*. IEEE, 2006.

[7] D. Bajde, M. Bruun, J. Sommer, and K. Waltorp. General public's privacy concerns regarding drone use in residential and public areas. 2017.

[8] V. Chang, P. Chundury, and M. Chetty. Spiders in the sky: User perceptions of drones, privacy, and security. In *CHI*.

[9] P. A. Clark. Someone had a seizure in vr and nobody knew what to do, 2018.

[10] R. A. Clothier, D. Greer, D. Greer, and A. Mehta. Risk perception and the public acceptance of drones. *Risk analysis*.

[11] S. Cobb and et al. Virtual reality-induced symptoms and effects (vrise). *Presence: teleoperators and virtual environments*, 1999.

[12] L. F. Cranor. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.*, 2012.

[13] S. Das, A. D. Kramer, L. A. Dabbish, and J. I. Hong. The role of social influence in security feature adoption. In *CSCW*. ACM, 2015.

[14] L. De Paolis and A. Mongelli. *Augmented and Virtual Reality*. AVR, 2015.

[15] T. Denning, Z. Dehlawi, and T. Kohno. In situ with bystanders of augmented reality glasses: Perspectives on recording and privacy-mediating technologies. In *CHI*. ACM, 2014.

[16] J. Francis and et al. What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology and Health*, 2010.

[17] R. S. Geiger and D. Ribes. Trace ethnography: Following coordination through documentary practices. In *HICSS*. IEEE, 2011.

[18] C. Goerge, M. Khamis, E. von Zezschwitz, M. Burger, H. Schmidt, F. Alt, and H. Hussmann. Seamless and secure vr: Adapting and evaluating established authentication systems for virtual reality. In *USEC*, 2017.

[19] N. Grgić-Hlača, E. M. Redmiles, K. P. Gummadi, and A. Weller. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. *WWW*, 2018.

[20] I. Hadar, T. Hasson, O. Ayalon, E. Toch, M. Birnhack, S. Sherman, and A. Balissa. Privacy by designers: software developers' privacy mindset. *Empirical Software Engineering*, 2017.

[21] J. M. Haney and W. G. Lutters. The work of cybersecurity advocates. In *CHI*. ACM, 2017.

[22] E. Hargittai. The digital divide and what to do about it. *New economy handbook*, 2003.

[23] S. Jana, D. Molnar, A. Moshchuk, A. M. Dunn, B. Livshits, H. J. Wang, and E. Ofek. Enabling fine-grained permissions for augmented reality applications with recognizers. In *USENIX Security Symposium*, 2013.

[24] V. Kameswaran, L. Cameron, and T. R. Dillahunt. Support for social and cultural capital development in real-time ridesharing services. *CHI*, 2018.

[25] O. S. Kerr. Criminal law in virtual worlds. 2008.

[26] P. Klasnja, S. Consolvo, T. Choudhury, R. Beckwith, and J. Hightower. Exploring privacy concerns about personal sensing. *Pervasive Computing*, 2009.

[27] K. Lebeck, T. Kohno, and F. Roesner. How to safely augment reality: Challenges and directions. In *International Workshop on Mobile Computing Systems and Applications*. ACM, 2016.

[28] L. N. Lee, S. Egelman, J. H. Lee, and D. A. Wagner. Risk perceptions for wearable devices. *CoRR*, 2015.

[29] M. Lombard, J. Snyder-Duch, and C. C. Bracken. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human communication research*, 2002.

[30] G. Maganis and et al. Sensor tricorder. In *ACM MCSA*.

[31] A. M. McDonald and L. F. Cranor. The cost of reading privacy policies. *ISJLP*, 2008.

[32] T. Merel. The reality of vr/ar growth. https://techcrunch.com/2017/01/11/the-reality-of-vrar-growth/, 2017.

[33] V. Motti and K. Caine. *Users' Privacy Concerns About Wearables*. 2015.

[34] J. W. Nelson. A virtual property solution: How privacy law can protect the citizens of virtual worlds. *Okla. City UL Rev.*, 2011.

[35] B. Nonnecke and J. Preece. Lurker demographics: Counting the silent.

[36] F. O'Brolcháin and et al. The convergence of virtual reality and social networks: threats to privacy and autonomy. *Science and engineering ethics*, 2016.

[37] A. N. Oppenheim. *Questionnaire design, interviewing and attitude measurement*. 2000.

[38] I. Pollach. What's wrong with online privacy policies? *Communications of the ACM*, 2007.

[39] E. Rader and J. Slaker. The importance of visibility for folk theories of sensor data. In *USENIX SOUPS*, 2017.

[40] A. Raij, A. Ghosh, S. Kumar, and M. Srivastava. Privacy risks emerging from the adoption of innocuous wearable sensors in the mobile environment. In *CHI*. ACM, 2011.

[41] A. Rizzo, J. Pair, P. J. McNerney, E. Eastlund, B. Manson, J. Gratch, R. Hill, B. Swartout, et al. Development of a vr therapy application for iraq war military personnel with ptsd. *Studies in health technology and informatics*, 2005.

[42] F. Roesner, T. Kohno, and D. Molnar. Security and privacy for augmented reality systems. *Communications of the ACM*, 2014.

[43] J. Roettgers. Study predicts fewer than 10 million monthly u.s. vr headset users this year, 17 million by 2019. `http://variety.com/2017/digital/news/vr-headset-data-mau-2017-2019-1202440211/`, 2017.

[44] E. M. Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[45] K. Shilton. Four billion little brothers?: Privacy, mobile phones, and ubiquitous data collection. *Communications of the ACM*, 2009.

[46] K. Shilton. Values levers: Building ethics into design. *Science, Technology, & Human Values*, 2013.

[47] The Franklin Institute. What's the difference between ar, vr, and mr?

[48] Virtual Reality Society. The history of virtual reality.

[49] Y. Wang, H. Xia, Y. Yao, and Y. Huang. Flying eyes and hidden controllers: A qualitative study of people's privacy perceptions of civilian drones in the us. *PoPETS*, 2016.

[50] I. Warren and D. Palmer. *Crime risks of three-dimensional virtual environments*. PhD thesis, Australian Institute of Criminology, 2010.

[51] D. Wen, X. Zhang, and J. Lei. Consumers' perceived attitudes to wearable devices in health monitoring in china: A survey study. *Computer Methods and Programs in Biomedicine*, 2017.

[52] P. Wijesekera, A. Razaghpanah, J. Reardon, I. Reyes, N. Vallina-Rodriguez, S. Egelman, and C. Kreibich. âĂIJis our children's apps learning?âĂİ automatically detecting coppa violations.

[53] S. Willis and B. Tranter. Beyond the 'digital divide' internet diffusion and inequality in australia. *Journal of sociology*, 2006.

[54] R. Yao and et al. Oculus vr best practices guide. *Oculus VR*, 2014.

[55] Z. Yu, H. Liang, C. Fleming, and K. Man. An exploration of usable authentication mechanisms for virtual reality systems. In *IEEE APCCAS*.

APPENDIX

A. ADVERTISING

A.1 Groups in Which We Advertised

We advertised in the following groups to recruit our interview participants.

- Reddit
  1. R/GearVR: Forum for users of Oculus Gear headset. `www.reddit.com/r/gearvr`
  2. R/googlecardboard: Forum for users of Google-Cardboard VR headset. `www.reddit.com/r/googlecardboard`
  3. R/oculus: Forum for users of Oculus to discuss VR. `https:www.reddit.com/r/oculus`
  4. R/RiftForSale: Forum for people to buy or sell VR tech. `www.reddit.com/r/RiftForSale`
  5. R/steamVR: Forum for VR users to discuss STEAM VR games. `www.reddit.com/r/steamvr`
- Facebook
  6. Virtual reality group: Facebook group for users and developers of VR to discuss VR. `www.facebook.com/groups/virtualrealitys/?fref=ts`
  7. Oculus Rift group: For Oculus and VR users and developers to discuss VR platforms that focus on the Oculus Rift. `www.facebook.com/groups/OculusRift/`
  8. Women in VR/AR group: For women developing in VR and AR to discuss opportunities and etc. `www.facebook.com/groups/womeninvr/`
  9. Oculus Rift Users Au group: Users of VR to discuss topics about VR that pertain to the Oculus Rift. `www.facebook.com/groups/277312492704516/`
  10. Google Cardboard: Users of Google Cardboard. `www.facebook.com/groups/1568155100117690/`
  11. Google Cardboard Developers: Developers of Google Cardboard. `www.facebook.com/groups/cardboarddev`
  12. Virtual Reality Gear: Oculus Rift, HTC Vive, Gear VR, Microsoft MR, PS VR, Oculus Go, Virtual Reality. Oculus Santa Cruz, Vive Focus, Occipital Bridge, Daydream, ODG R8, ODG R9, Pimax 8K, and OSVR users and developers. `www.facebook.com/groups/gearvr/about/`
  13. Daydream and ARCore: For professional enthusiasts, UX Designers, Programmers, Unity & Unreal Engine Developers, Artists, and other VR professionals who use Google Products like ARcore and Daydream. `www.facebook.com/groups/daydreamvirtualreality/`
  14. AR & VR Developers: Everything developers need to know about: augmented and Virtual reality (VR), Mixed reality, VR/AR apps & games development, and Hardware `www.facebook.com/groups/ARVRMR/about/`
  15. Two institution-related groups omitted for blinding.
- Forums
  17. Oculus General Forum: The Offical Oculus website forum. `forums.oculusvr.com/community/categories/general`

We advertised our co-design study to groups that were explicitly developer focused or from which we recruited the most developers, groups 1-5, 6, 7, 8, 13, 14, 15.

A.2 Advertising Text

A.2.1 Interview Study

The following advertising text was posted in the 17 online communities to recruit participants to complete our screening questionnaire for the interview study.

> Join an Exciting Study on Virtual Reality

Are you 18 or over the age of 18? Do you use VR systems or applications?

If you answered YES to these questions, you may be eligible to participate in a virtual reality research study.

We want to talk to you about your experience using a VR system. We want your input for a 20 minute interview! Interviews will be conducted over the phone or through Skype. Participants will be compensated with a $15 Amazon gift card.

A.2.2 Code of Ethics Co-Design Study

The following advertising text was posted in nine VR developer online communities to recruit developers to contribute to the design of a VR developer code of ethics.

tl;dr edit this document: [url] to help create a collaborative VR developer code of ethics. Email [address] to get a $2 amazon gift card for helping out!

Long explanation: You might remember us from a post a little while back. We are a team of researchers studying development, security, and privacy in VR. As part of this project we interviewed developers from the VR community (thank you for participating!) about their experiences developing, what they see as the safety, security, and privacy concerns in VR, and etc. We also interviewed VR users about their use of VR and their concerns.

One of the key points raised by developers was that there is no standardized "code of ethics" or "instruction sheet" for what to do with user data, how to notify users of data use, and how to practice ethical VR development.

We would like to invite you to come together as a community (the strength and openness of the VR development community was also a common theme mentioned in the research), with our support, to develop a set of standards for ethical development in VR.

Every contributor to the code of ethics will receive a $2 amazon gift card as a "thank you" from our team. We will host the code of ethics on a public website once it is finished and credit you (if desired) for your hard work, as well as publish the code in the research paper we are preparing.

B. INTERVIEW PROTOCOL

The following protocols were used during the 20 minute semi-structured interview conducted via phone, Skype, or Google hangouts.

B.1 Developers

Introduction

Hello. My name is [INSERT NAME] and this is [INTRODUCE OTHER PERSON]. Today we will be conducting a study on virtual reality.

Today we are going to chat about your experiences with virtual reality. I expect that our conversation will take approximately 30 minutes.

Motivations

I'd like to start our conversation with a discussion of what made you want to develop applications or systems for virtual reality.

1. How did you get into developing for VR?
2. Why did you choose VR?

Skill Acquisition

Next I would like to talk about how you learned the skills for your VR development.

1. How did you learn to develop on VR?
2. What resources or tools did you use to learn VR development?
3. Which ones?
4. Did you talk to anyone to learn to work with VR?
5. What do you feel is different about developing for VR?
6. Do you have any different concerns when you are developing?

Concerns

1. What are you currently developing or what have you developed for VR?
2. Why did you decide to develop this product?
3. What does your product do?
4. Do you foresee any barriers [if product already released: are there any barriers you feel are currently] preventing your product from reaching the market saturation you want to achieve?
5. How do you plan to address these barriers?
6. Do you forsee any privacy and security concerns with your product or VR in general?
7. For each concern: why?

Data Collection

1. What user data does your product collect?
2. For each data type: What are you planning to do with this data?
3. If no collection reported: What type of data could you collect? What might it be used for?
4. Do you think that users will be [/would be] concerned about this data being collected?
5. Why/why not?

Recommendations

1. (if sensible based on prior answers) How would you educate other developers on privacy and security risks for VR?
2. What do you wish you had known?
3. What materials would you like to have had access to?
4. In general, what advice would you give to someone wanting to develop with VR?
5. What information or resources would you point to for someone wanting to learn about developing VR?

B.2 Users

Introduction

Hello. My name is [INSERT NAME] and this is [INTRODUCE OTHER PERSON]. Today we will be conducting a study on virtual reality.

Today we are going to chat about your experiences with virtual reality. I expect that our conversation will take approximately 20 minutes.

Motivations

I would like to begin with a few questions about your current use of VR.

1. How long have you been using VR?
2. How did you learn about VR?
3. What made you decide to buy/use a VR headset?
4. Ask why for each thing they mention?
5. Why did you choose your particular VR software over others?
6. Where did you go to find out information about the systems?
7. What do you usually do with VR?
8. What do you see as the benefits of virtual reality?
9. What were your goals when you started using these systems?

Concerns

1. When you were making your purchase, did you have any concerns?
2. Why/why not?
3. What about with your specific headset?
4. Did you worry about privacy at all when you were deciding whether to purchase the system?
5. Could you tell me a bit more about your concerns with <each item>
6. Do you still have these concerns?
7. Do you do anything to try to prevent this?
8. How about security, any concerns there?
9. Could you tell me a bit more about your concerns with <each item>
10. Do you still have these concerns?
11. Do you do anything to try to prevent this?

Data Collection

1. Do you think your virtual reality system collect information about you?
2. What do you think it collects?
3. How do you think this information is used?
4. Are you concerned by this?
5. Would you say you feel differently about this than about data that gets collected by your other devices? Why?

Recommendations

1. How likely is it that you could recommend VR to a friend or colleague?
2. What concerns would you share or discuss?

Introducing the Cybersurvival Task: Assessing and Addressing Staff Beliefs about Effective Cyber Protection

James Nicholson
PaCT Lab
Northumbria University
Newcastle, UK
james.nicholson@northumbria.ac.uk

Lynne Coventry
PaCT Lab
Northumbria University
Newcastle, UK
lynne.coventry@northumbria.ac.uk

Pam Briggs
PaCT Lab
Northumbria University
Newcastle, UK
p.briggs@northumbria.ac.uk

ABSTRACT

Despite increased awareness of cybersecurity incidents and consequences, organisations still struggle to convince employees to comply with information security policies and engage in effective cyber prevention. Here we introduce and evaluate *The Cybersurvival Task*, a ranking task that highlights cybersecurity misconceptions amongst employees and that serves as a reflective exercise for security experts. We describe an initial deployment and refinement of the task in one organisation and a second deployment and evaluation in another. We show how the Cybersurvival Task could be used to detect 'shadow security' cultures within an organisation and illustrate how a group discussion about the importance of different cyber behaviours led to the weakening of staff's cybersecurity positions (i.e. more disagreement with experts). We also discuss its use as a tool to inform organisational policy-making and the design of campaigns and training events, ensuring that they are better tailored to specific staff groups and designed to target problematic behaviours.

1. INTRODUCTION

The number and scale of cyber-attacks targeted at organisations over the past few years is unprecedented. These include hackers compromising 55 million voter records in the Philippines, hospitals worldwide hit by ransomware attacks, 33 million Twitter user names and passwords being compromised, and 11.5 million documents relating to offshore accounts of international politicians, business leaders and celebrities being leaked from a law firm [51]. Many major breaches still go unreported, with only a quarter of businesses in the UK reporting their major breaches last year [33]. Business email compromise, ransomware, and phishing are cited across industries as the top vector of compromise. In many of these cases, the attack vector involves the employee. Organisations and their employees understand that they have a responsibility to change employee behaviour as an important tool in their defence strategy, yet there is very little consensus about exactly what protective behaviours are to be advocated and prioritised. Security practitioners, policy-makers, managers, and employees tend to advocate different approaches and the end result is that users receive conflicting advice, become sceptical about the information they are given, and are consequently less proactive in cyber defence than they might otherwise be [9, 35].

Organisations typically have one or more policies addressing appropriate cybersecurity behaviour, referred to as *security policies* from here on. There is now significant literature that describes those factors that influence employees' intentions to comply with security policies [21, 31, 37, 46] and further literature documenting poor outcomes from cybersecurity awareness campaigns and organisational training initiatives [5, 41, 49, 55]. Sometimes the reason for these failures is straightforward. For example, security policies are often inaccessible or buried deep within an organisation's website, tend to be over-complex, incomprehensible and/or poorly tailored to staff needs and workload [39]. They are generally poor calls to action, not least because of the aforementioned confusion about the protective actions they promote. This is a particular problem when we consider the psychology of threat, where we know that highlighting the threat to a user, without also offering them a simple, consistent response to that threat, produces 'defensive' reactions that can include simply ignoring the problem and continuing to engage in old behaviours [29].

One example is the conflicting advice surrounding the password, where standard advice was once to create strong, unique passwords for every user account involving combinations of letters, numbers and 'special' characters. Recently, this advice has been supplanted (e.g. by NIST and GCHQ) with a 'three random words' instruction for password creation [25]. This would seem to constitute an advance, but can lead to greater confusion on the part of the end user as many current accounts still enforce 'strong' passwords requiring multiple character types, effectively rendering GCHQ and NIST advice useless in that particular context.

In this paper, we focus on the consensus problem in cyber protection and describe a tool (*The Cybersurvival Task*) that highlights the many different behaviours encompassed by a cybersecurity policy and the mental models held by members of an organisation. The task requires users to rank protective behaviours in terms of their effectiveness as a cybersecurity defence. Unlike other self-report measurement tools (e.g. [19]), these rankings provide a means for staff to disclose their assumptions in a structured way, so that organisations can understand where employee confusion and associated defensive responding might be taking place. Most importantly, the process allows for organisational security experts to reflect upon their policy and training priorities, based on direct feedback from their own

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018. August 12 -- 14, 2018, Baltimore, MD, USA.

employees. The ultimate aim is that the Cybersurvival Task could inform the development of organisational policy-making and the design of campaigns and training events, ensuring that they are better tailored to specific staff groups and/or misconceptions. Here, we describe the development of the task and describe a process whereby we piloted the task in one institution, made some refinements, and then conducted an evaluation of the final task in a second institution. We show how the task highlighted misconceptions and revealed behavioural discrepancies between experts and employees, and between different employee groups, and discuss how organisations can benefit from the Cybersurvival Task.

2. BACKGROUND

2.1 The Human Factor in Cyber Protection

Organisations face a growing range of security threats, including denial of service (DoS) and ransomware attacks that aim to take down a business or service, as well as social engineering attacks that are designed to obtain and exploit private information. While it may be possible to stay safe from such attacks by improving and maintaining the organisation's technical defences – e.g. firewalls and anti-virus software – employees have often been labelled as the 'weak link' in the security ecosystem (e.g. [53]). In recent years, this weak link argument has been replaced by an understanding that humans, far from being 'the enemy' [2] are an integral part of the *whole system* and that a proper understanding of human behaviour and of employee motivation should inform the cybersecurity design process [45].

Much of the work in this space has focused upon the fact that cybersecurity does not comprise the primary task for most employees. Unsurprisingly, people attend to their primary work tasks and tend to overlook security actions. Beautement, Sasse, & Wonham [8] have suggested that employees have a relatively small 'compliance budget' that they can allocate to security procedures and that this can shrink when job demands are particularly high or the protective behaviours demanded of users are too onerous. There are unrealistic expectations that users will create a strong password for every unique account they have [56], that they will be vigilant in checking for phishing emails they receive [54] or that they will simply not click on any links or open any email attachment in the workplace [27]. The reality is that the vast majority of people reuse simple passwords [1, 26] and that almost half of all users are likely to fall for phishing emails, with some 17% on average entering credentials on phishing websites [12].

2.2 The Non-Compliance Problem

There is often a disconnect between how organisations would *like* their employees to behave and how the employees *actually* behave and this is an important consideration for computer security (e.g. [8, 30]). Much of the existing organisational research tends to focus upon this as a 'policy compliance problem' rather than see it more holistically as an issue around the ways that employees come to understand both the cybersecurity threat and the kinds of protective security behaviours they can use to ameliorate that threat. This is important, because employees do not typically gain their understanding directly from security policies, but rather from their work peers and from the media, building up a set of *shadow security* beliefs and behaviours [34] that deviate from company policy. In other words, employees reach a compromise between security and productivity that allows them to achieve their work goals by utilising non-compliant but sufficient security behaviours.

Regardless, many organisational policies and procedures are simply not fit for purpose. There are issues with policies that are too dense and contain tracts of information that are irrelevant for many users. There are also issues with policies that are too vague and provide very little in the way of useful information [3]. Unsurprisingly there are also many organisations that have no security policies in place and many users who are simply unaware of their own organisation's stance on cybersecurity behaviour. In short, it is not easy for organisations to develop usable cybersecurity policies to keep employees safe.

2.3 Choosing the 'Right' Behaviour

We noted that users often struggle to protect themselves and their organisation online, and part of the problem is that they are given inconsistent advice about what actions to take. As cyber security experts differ in their opinion of the skills and behaviours that are important [13], so too do the security policies they create. This means security policies vary between organisations and include many different behaviours associated with accessing, categorising, storing, and transferring data – but may also cover general computer user policies including internet and email behaviours and use of external devices (USBs, personal devices). With this in mind, researchers at Google distributed a survey to both security experts (those having at least 5 years of experience working or studying in computer security) and security non-experts (Mechanical Turk workers) and found a discrepancy between online security behaviours reported as essential between the expert and non-expert group [32]. Most importantly, the researchers compiled a list of advice considered 'good' by experts consisting of 20 items. While this list constitutes a step in the right direction for identifying security behaviours that are important for staying safe online – for both policy creation and advice-generation – this advice is based on both academic and industry experts which may have contrasting views on a number of topics [32]. Additionally, this list is based on 'good' advice, defined as advice that is both effective and realistic, which potentially means that security behaviours that are very important for the organisation may have been pushed down the list. Finally, the list was compiled for the average internet user, meaning that some behaviours may not apply to everyone and this is already a problem faced by users who are overloaded with occasionally irrelevant advice [30]. In a corporate environment where job roles are clearly defined and responsibilities differ across individuals, such a generic list will likely offer excessive or irrelevant advice to individuals.

2.4 Measuring Security Behaviours and Beliefs

We have highlighted the problems that organisations face when writing security policies, so it is no surprise that enforcing the policy becomes even more challenging. But how can organisations understand what their employees are doing in the security spectrum?

Direct measurement of actual security behaviour in a live environment has proved elusive for cybersecurity researchers and many have adopted self-report scales as workable alternatives. These, of course, measure intentions to behave in a certain way and assume there are no barriers to converting these intentions into actual behaviour. A range of psychometric scales have been developed and these typically include different behavioural items where participants are asked to rate the likelihood of complying or agreement with the behavioural statements. For example, Egelman

& Peer [19] start off with 30 items which they reduce to 16 security behaviours covering 3 security topics, whereas Parsons et al. [48] list 63 different behaviours covering 7 topics. However, Wash et al. [61] found that people are poor at self-reporting security behaviours, as they may not understand what the behaviours are and may underreport less salient behaviours. This has important implications for the validity of such scales.

With this in mind, a different approach to measuring and observing behaviour may be necessary and in this paper we consider the advantages of ranking behaviours instead of rating them. The inspiration from this work comes from two seminal examples of ranking tasks used both to facilitate group discussions and to study group dynamics in occupational settings: The Desert Survival Situation [38] and the Moon Landing Task [16]. While these tasks do not measure organisationally-relevant behaviours and beliefs, they are worth considering here as they have been used for over four decades to understand the kinds of decision-processes individuals and groups make within the work context and to determine which factors are most likely to shape attitudes within the workplace.

2.5 Ranking Tasks as a Measure of Work-Related Behaviour

The Desert Survival Task [38] places participants in a simulated scenario where they are stranded in the desert after a plane crash and must rank 15 items in order of importance for survival. Participants' answers are then compared to the 'correct' answers – i.e. the rankings offered by experts – in order to indicate the accuracy of the individual and group rankings. The task has been a popular tool for understanding the behaviour of leaders in groups (e.g. [23, 42, 52]), evaluating group facilitation techniques (e.g. [58]) and exploring both individual and group decision making and problem-solving processes (e.g. [15, 24, 44]). The Desert Survival Task has also been used in disciplines other than management as a tool for understanding gender differences in schools [6], understanding what features of embodied conversation agents are most important for communicating feedback [40] and for understanding reactions to different computer personalities [20] amongst many others.

Similarly, the Moon Landing Task [16] requires participants to rank 15 items in order of importance for surviving a trip to a rescue vessel off the moon's surface. Individual and group rankings are then compared with an expert list compiled by the National Aeronautics and Space Administration (NASA). The Moon Landing Task has been used largely as a problem-solving task in studies, e.g. for understanding role of stereotypical context on the judgement of groups [4], cognitive busyness [17, 28], and teasing [10]. The task has also been used to understand group interactions amongst children [16] and as a tool for facilitating intelligence expectancy judgements on peers [43].

3. THE CYBERSURVIVAL TASK

The Cybersurvival Task asks *participants* (employees in an organisation) to *rank* the security behaviours that would best help protect their own organisation. This process is different from other security questionnaires that operate on a self-report basis, where users are asked to disclose whether or not they perform certain behaviours [19, 48]. By asking users to rank behaviours, we ensure that participants prioritise certain behaviours over others. By asking users to justify these rankings, we ensure that they articulate their beliefs about the benefits and drawbacks of these behaviours.

Table 1: Overview of the Cybersurvival Task stages.

Stage	Approx. Duration
Generate appropriate list of behaviours with the organisation's security experts tailored to workplace	30 minutes
Workshops with employees	60 minutes (each)
Reflection with experts	45 minutes

The task involves a process similar to the Moon Landing and Desert Survival tasks – in which participants engage in both individual and group ranking decisions and compare them against previously-obtained expert rankings. The major difference in this implementation is that the task items are highly salient to the cybersecurity context. In other words, the Moon Landing and Desert Survival tasks allowed exploration of a problem that was not directly relevant to the organisation in order to understand group dynamics in a 'neutral' problem space. In contrast, the Cybersurvival Task is highly relevant and allows not only the exploration of group dynamics, but the elicitation of specific mental models (at group and individual level) that are cybersecurity relevant. Critically, the Cybersurvival Task also incorporates a final reflection stage (see Table 1) not present in similar ranking tasks, where *experts* (those responsible for setting the security agenda in an organisation) can be presented with data capturing employee rankings, assumptions and beliefs.

Table 2: Overview of workshop activities.

Activity	Approx. Duration
Introduction by Facilitator	2 minutes
Individual ranking of Cybersurvival Sheet	10 minutes
Reveal of top 3 and bottom 3 behaviours (from individual rankings), plus suggestions for new behaviours	10 minutes
Group ranking of Cybersurvival Sheet – assisted by facilitator	10 minutes
Group ranking of Cybersurvival Sheet - independent	15 minutes
Reveal of expert rankings & scoring	10 minutes
Debrief	3 minutes

The task itself is simple: each participant is initially presented with a sheet (the *Cybersurvival Sheet*) consisting of *n* relevant security behaviours (agreed in advance with the organisation's security experts), listed in a random order, and is required to rank those behaviours in order of importance for staying safe online. The task is conducted individually, then conducted as a group, where participants are encouraged to discuss and agree on the importance

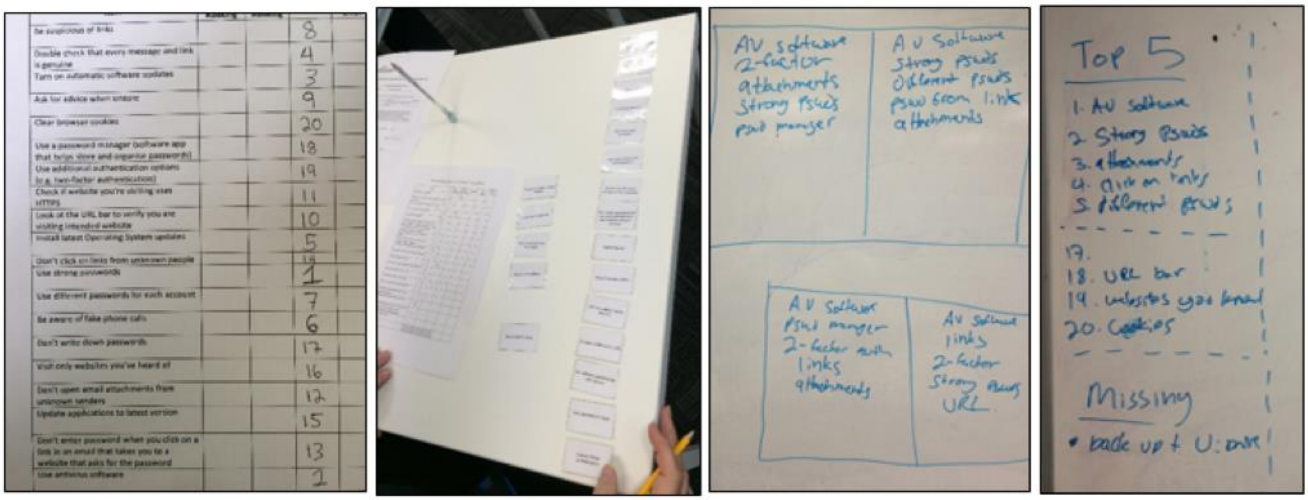


Figure 1: (a) Example of the ranking sheet used in Phase II; (b) laminated note cards used to support the individual ranking process; (c) the 'individual reveal' from each participant and (d) the group's top five agreed behaviours.

of the different behaviours. The discussion through which participants come to a group consensus is as important as the rankings themselves, which are then compared with those derived from security experts in their organisation. Each participant is given a set of laminated note cards with the printed behaviours that they can use to facilitate both the individual and group ranking process (e.g. by arranging the notes before committing pen to Sheet (see Figure 1).

The security experts' rankings are obtained via a similar process where each expert is asked to rank an initial list of behaviours (see 3.1 below) individually, followed by a group discussion where all experts have to agree on an order that suits their organisation. Experts are allowed to add, rename, and remove any behaviours at any time during the process. The initial expert ranking exercise lasts approximately 30 minutes, while the final reflection stage lasts approximately 45 minutes.

This final reflection stage highlights a striking difference between the Cybersurvival Task and the Dessert Survival or Moon Landing tasks. While the latter two tasks operate under an absolute and 'best set' of rankings, the Cybersurvival Task challenges the quality of the expert rankings in the final stage where they are encouraged to reflect on (and re-assess) their priorities and training programmes. The reflection stage consists of the researchers presenting the findings to the experts and allowing them to seek clarification on any of the findings (or specifics on behaviour choices). See Section 3.3 for more information on this stage.

We acknowledge that experts can be wrong (as we will show later), and by no means do we believe that the expert rankings from each institution necessarily represent 'best practice', but we do see the value in comparing employee rankings to their institutional experts as they have been tasked with setting and enforcing the security culture within their organisation.

Below we describe the multi-phase process undertaken to refine and evaluate the Cybersurvival Task, comprising a first deployment, task refinement and second deployment and evaluation in an institution of similar character and size.

3.1 Phase I Deployment

The first Cybersurvival Task deployment was in a large academic institution (approximately 3,000 members of staff). The goal was to understand the 'face validity' of the task from the point of view of experts and employees and to see whether any improvement should be made to its structure, activities, and delivery. We were also interested in whether the organisational experts and employees believed there was any value in engaging with the Task.

We first needed to develop a list of protective behaviours that were deemed relevant to the organisation, and so we conducted an initial workshop with two security experts from the organisation (the Head of IT Security and the Head of IT Services). We began with an initial list comprising the 20 behaviours from Ion et al.'s [32] study described above (see Appendix A). The two experts were asked to work individually and to rank the list of behaviours in order of their importance *for protecting their organisation*, and they were also given the chance to add and remove behaviours. Both experts were then asked to work together to rank the complete set of behaviours, including any new ones they had added. Their final ranked list, the 'expert agreed list', presented in randomised order, formed the Cybersurvival Sheet for employees (see Appendix B). We then used this sheet to run the Cybersurvival Task in four workshops (see Table 2 for activities) with staff in the same organisation, followed by one final workshop with the same experts who generated the initial list. Both this and the subsequent deployment received ethical approval from our university.

Twenty employees were recruited using strategically-located flyers and email distribution lists. There were 13 support staff with roles ranging from procurement to personal assistants and 7 academic staff responsible for either research or student learning. The 20 participants were split into four workshops of five participants each. One workshop consisted of solely support staff and one of solely academic staff, with the remaining two mixed. The activities and procedures in all four sessions were identical (see Table 2). Each workshop involved the participants ranking the behaviours on their own, discussing any additional behaviours with the group, and then ranking the behaviours again as a group, with a final ranking order agreed by all members of the group (see Figure 1).

Participants were then shown the agreed 'expert rankings' and were given the opportunity to discuss any differences between their rankings and those of the security experts. The sessions lasted approximately one hour. Thus, we collected the ranked list of behaviours for every participant (n=20) and the ranked list of each group (n=5) as well as the qualitative discussions during the group ranking activity (n=5).

Finally, the organisation' security experts were briefed on the findings and allowed to reflect on these (see Section 3.3).

3.1.1 Lessons Learned

The Phase I deployment of the Cybersurvival Task provided us with very valuable feedback and led us to improve upon the procedures and materials for Phase II. Below we cover the most important lessons that we learned from Phase I.

Experts expressed major interest in the reflection stage and viewed this as the most valuable aspect of the Task. However, its importance was not evident at the beginning of the task, thus leading to lower engagement with the initial ranking task. Therefore, in Phase II we were clearer with experts upfront about the entire process and highlighted the benefits of tailoring the initial set of behaviours to their own organisation.

In Phase I we focussed on the ranking of the top 5 behaviours, which meant that subsequent discussion between the group members centred around those 5 behaviours with less discussion around the lowest-ranked items. In Phase II we decided to facilitate the ranking of the top and bottom 3 behaviours, thus resulting in a more balanced discussion of 'good' and 'bad' behaviours.

We also improved the presentation of the Cybersurvival Sheet based on feedback from participants. The most important change was numbering the behaviours on the sheet to facilitate discussion amongst participant (e.g. "cookies, number 7, should go below two factor authentication, number 17"). Feedback from participants also highlighted their appreciation for the laminated cards, so we continued using these facilitators during Phase II.

Finally, we observed from the initial set of 4 workshops that the most insightful data was generated by the group ranking activity, where participants were forced to directly compare the pros and cons of behaviours which led to uncovering flawed mental models and/or shadow security measures, as well as exposing issues with the security policy. Thus, we altered the timings during Phase II to allow staff more time in the group ranking activity and less time in the expert reveal, where participants predominantly dismissed the expert rankings.

3.2 Phase II Deployment

The second phase involved a deployment of the revised Cybersurvival Task in a larger but structurally similar university (approximately 5,300 members of staff). This meant that the lessons learned from Phase I were appropriate to the new context and that the kinds of attacks and protective behaviours described were appropriate, recognising that universities are prime targets for attackers due to publicly-available information [50].

The list of behaviours for Phase II was again developed through an initial workshop with security experts from that organisation – the Chief Information Security Officer and a Faculty deputy (see Table 3 for ranked list). The format of the workshop was similar to that used in Phase I, with a greater emphasis on the potential benefits of the task during the initial briefing. The initial list of behaviours comprised the list from Phase I, plus additional behaviours that were recommended in the new organisation's security policy. This resulted in the experts spending more time adding, removing, and rewording behaviours on the list in order to tailor it to their specific organisation.

Again, the participants were 20 non-expert employees who were split into 4 groups of 5 participants each. In Phase II, we kept support and academic staff separate – with two groups of each. Note that while we chose to separate academic and support staff due to differences observed during Phase I, it is possible for organisations to separate staff as they see appropriate (e.g. by job role or subjective experience). In fact, the Cybersurvival Task can serve as an exercise for identifying potential subgroups of employees who may share similar misconceptions.

Table 3: Final ranking of behaviours by the security experts for Phase II. Keys correspond to Figure 3.

Ranking	Behaviour	Key
1	Ask for advice	ASK
2	Save files to the network	SAV
3	Use different passwords for accounts outside the organisation	DIF
4	Keep passwords safe if written down	WRI
5	Report any data loss incidents	REP
6	Turn on automatic software updates	AUT
7	Do not disclose your personal password, even to the IT department	DIS
8	Use anti-malware software and keep it up to date	ANT
9	Use strong passwords	STR
10	Educate yourself on how to avoid fraud	EDU
11	Use additional authentication options (e.g. two-factor authentication)	ADD
12	Restrict physical access to computers and removable media	PHY
13	Check if website you're visiting uses HTTPS	HTT
14	Look at the URL bar to verify you are visiting intended website	URL
15	Don't open attachments from unknown senders	UNK
16	Don't open unnecessary attachments	UNN
17	Don't click on links from unknown senders	LIN
18	Don't enter password when you click on a link in an email that takes you to a website that asks for the password	PAS
19	Clear browser cookies	COO

Participants were recruited via snowball emails across all Faculties of the university with the exception of Computing Science (who were excluded on the basis that they may have had particular cybersecurity expertise). Academic participants included PhD students, researchers and lecturers, while support participants included receptionists and staff in finance and human resources departments. All these 'non-expert' participants were compensated with a £10 voucher.

Behaviour	Individual Range	Individual Mean Rank	Group Mean Rank	Expert Rank	Key Reason
Ask for advice	7-19	18	17	1	-Not practical to do due to time required for response. Use own experience instead to make the correct judgement. -More likely to ask a colleague for advice than ask IT
Save files to network	1-16	9	4	2	Default "best practice" behaviour; pragmatic: do not want to regenerate data if lost.
Use different passwords for accounts outside organisation	2-18	10	12	3	Acknowledged as an important behaviour, but practically not feasible as cannot be expected to remember 20 unique passwords.
Keep passwords safe if written down	3-19	8	15	4	Tensions between "never" and "limited access" points of view, but generally seen as a poor security behaviour.
Report any data-loss incidents	5-15	15	13	5	Important as if data has been lost then that is an indication that there is something wrong (and it may happen again).
Turn on automatic software updates	4-19	17	16	6	-Seen as necessary for functional improvements but not security patches -Those aware of security implications also rated behaviour low due to unlikelihood of being hacked due to out-of-date software
Do not disclose personal password, even to IT	1-17	6	6	7	Very important, as disclosing password means any effort put into the creation of the password is rendered useless.
Use anti-malware software and keep it up to date	3-19	13	5	8	Seen as IT's responsibility (installation and maintenance). Although they do not need to actively manage it, AMS is seen as a very important safeguard.
Use strong passwords	1-10	1	7	9	Worthless if given away or stolen.
Educate yourself on how to avoid fraud	8-18	16	11	10	Seen as losing battle against fraudsters as they are always ahead of the curve, so time-intensive to stay up-to-date.

Figure 2: Screenshot of the report produced for experts.

The sessions consisted of a quick introduction by the facilitator, an individual ranking task followed by a 'reveal' of each participant's top and bottom three behaviours (see Table 2 for activities and timings). Staff were given a chance to suggest new behaviours to add to the list. These were written on the board by the facilitator (see Figure 1). Participants were then asked to rank all the behaviours as a group with everyone having to agree on the final list at the end of the process. The group discussion was facilitated by a researcher for the top 3 and bottom 3 behaviours, and once those were agreed participants were allowed to continue with the group ranking activity unassisted. Once all participants were in agreement, the expert agreed list was shown to the group and they were allowed to discuss discrepancies both with the group and with the facilitator. Finally, participants were debriefed and allowed to go.

We collected the ranked list of behaviours for every participant (n=20) and the ranked list of each group (n=5) as well as the qualitative discussions during the group ranking activity (n=5). Once all data was analysed, experts were briefed on the findings during the Reflection Stage (see below).

3.3 Reflection

The purpose of the reflection stage was to brief the organisational security experts on the findings from the workshops and collect their thoughts on the process and understand their reaction to the findings. In total, the session lasted 45 minutes.

Half of the session consisted of an oral presentation describing the methodology of the Task, a reminder of their rankings, and an overview of the main findings including the graphs in Figure 3.

A brief physical report was generated for the experts that summarised the purpose of the Task, the methodology used to collect the data, and the most salient findings (see Figure 2 for example). The main section contained a table that included each behaviour (ordered according to the expert ranking), the individual range for the employee scores, the individual mean rank for the scores, the group mean rank, and the expert rank (for easy comparison). The key reasons for the overall scores were also included. The behaviours were highlighted where the individual and group scores were markedly different – in green if the change resulted in a higher score, or red if it resulted in a lower score. In this specific case, different tables were created for academic and support staff to highlight the differences between the groups. Finally, a section with the main takeaways (summarising the most controversial opinions or differences) closed the report.

Following the presentation, experts were engaged in a brief semi-structured interview where they were asked to comment on the Cybersurvival Task and reflect on the findings. Experts were also encouraged to seek clarifications on conflicting behaviours and were asked about future actions based on the presented data.

4. RESULTS

Below we present both quantitative and qualitative results from Phase II, including insights from both employees and experts. The quantitative data was analysed by averaging the scores across groups for each behaviour (e.g. Ask for Advice). All tests carried out were two-tailed. The qualitative data was obtained from the employee discussions during the group ranking activities and was analysed using thematic analysis.

4.1 Comparison of Rankings Between Experts and Staff

The rankings of experts were plotted against those given by staff (academic and support). These are presented in Figure 3. The identity line (dotted) shows perfect calibration between experts and staff. However, the further away the behaviours are from the identity line, the bigger the discrepancy between staff and experts' security priorities. Behaviours above the identity line represent those that are most important to staff, while those below the line represent behaviours that are most important to experts. Figure 3 also shows the difference between those rankings made as individuals and those made following group discussion with arrows indicating the shift between mean individual and group scores. One important thing to note here is that group discussion seldom moves staff towards better agreement with the experts. This is important given the way that social norms can intervene in determining staff security priorities (e.g. [35]). This will be explored in more detail below.

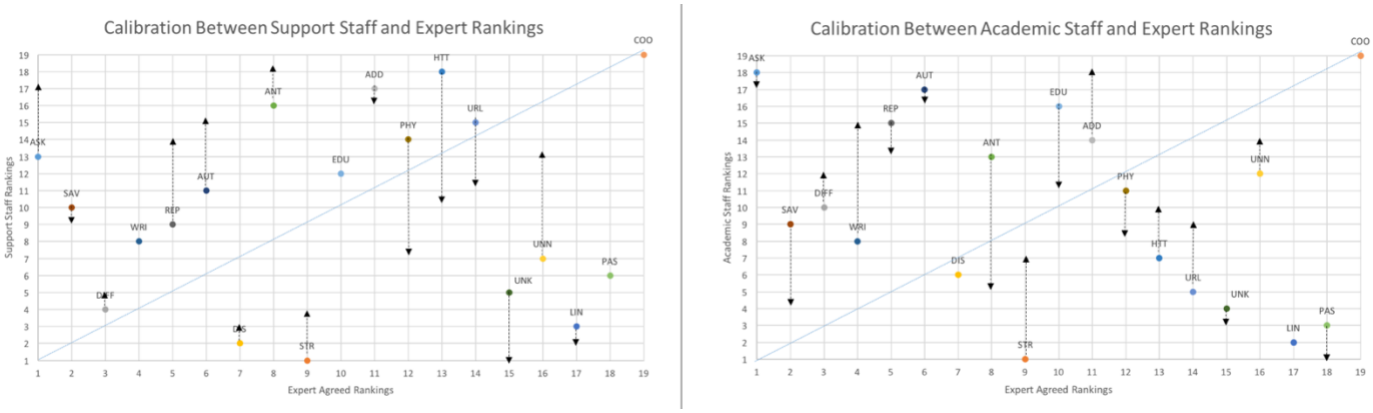


Figure 3: Scatter Plots comparing the rankings of experts (X-axis) against academic staff (right) and support staff (left). Arrows show the shift from mean individual rankings to final group rankings (dots). See Appendix C for high quality graphs.

4.2 Discrepancies in Rankings

At first glance, our graph shows poor calibration between experts and staff, with experts wanting staff to prioritise *asking for advice* and with staff (both academic and support) opting for the creation of *strong* passwords as their number one priority (again, an interesting issue in the light of recently shifting password advice).

Strong passwords have been the subject of many campaigns and are one of the few 'engineered' behaviours that staff are likely to encounter (as password systems often force the inclusion of upper and lower case, numerical and special characters as a means of creating stronger passwords). There is a certain irony here, given the new advice of *three random words* issued by GCHQ. Interestingly, support staff were more aware than academic staff of the need to use different (unique) passwords for each account, something that could possibly be tied to their use of systems that could hold sensitive data (e.g. finance, student performance, student identity, etc.).

A set of behaviours around not opening attachments or links from unknown senders were also seen as very important by academic and support staff but not so much by experts (who tended to place more trust in automated detection of malware). Similarly, checking URLs and checking HTTPS (to a lesser extent) were seen as important behaviours by the academic staff while they were rated low by the experts, although these behaviours were ranked as being less important after the group discussion and more in line with the expert scores.

Experts prioritised *asking for advice* as the single most important behaviour, but this was very poorly ranked by staff who generally believed that asking for advice was unnecessary and they could not envisage a scenario when that would happen (see below). *Reporting data loss* and *turning on automatic updates* were also seen as less important by academic staff when compared with the expert agreed rankings.

4.3 Individual vs. Group Rankings

In order to understand the differences in individual and group rankings, we calculated the absolute difference between the staff scores and the expert agreed scores (i.e. expert ranking minus staff ranking) for each of the subgroups: academic staff's individual rankings, academic staff's group rankings, support staff's individual rankings, and support staff's group rankings. The lower the added score, the closer the rankings were to the expert ones (0 = perfect, 361 = complete opposite). We then ran a Wilcoxon Signed Ranks Test between the individual scores and the group scores to measure any significant changes to ranking scores that emerged as a function of group discussion (in relation to the expert ones).

Table 4: Mean scores in the Cybersurvival Task (0 = perfect score; 361 = worst score).

	Individual Score	Group Score
Academic Staff	131.13	131
Support Staff	117.11	162

We found no significant difference in individual and group rankings for academics, $Z=-.140$, $p=.889$. We did, however, find a significant difference in individual and group rankings for support staff where individual scores were higher (i.e. more secure) than group scores, $Z=-2.668$, $p=.008$.

These results are worrying as they show that a group discussion about the importance of different cyber behaviours led to a weakening of the support staff's cybersecurity position (i.e. more disagreement with experts). This finding is also reflected in a statistical comparison of academic and support staff, where a Mann-Whitney U test did not find a statistically significant difference in the performance of *individuals*, $U=21$, $p=.167$ but where there was a significant difference in *group performance*, with academic staff generating rankings that were much more closely aligned with experts, $U=10$, $p=.011$.

4.4 Expert Assumptions

Here we report some of the qualitative data from the discussions within the different groups. We start by detailing some of the assumptions made by security experts about staff behaviour. Firstly, experts were adamant that there was an onus on employees to learn about security threats and to educate themselves. This notion was thoroughly rejected by our employees who felt that such behaviour would be too time consuming:

> *Academic Group 1 (Male): "Yeah, I think it's one of the things on my list that, I would really like to do, but you never get the time to actually get*

around to it. Presumably the fraudsters are getting cleverer and cleverer, so you have to keep up to date with new ways of helping and keeping yourself stay safe"

This is possibly one of our most predictable findings, given the extensive research literature on 'productive security' that notes the unrealistic and unacceptable 'cost' of cybersecurity policy compliance [7].

Secondly, experts assumed that users would save all their work regularly to the network drive in order to allow immediate restoration in the case of infections or attacks. In reality, this was common practice, but many staff chose convenience over security and downloaded a local working copy, which would then be uploaded to the network once access was no longer required.

Support Group 2 (Female): Force of habit, it's just a habit. I don't not for any particular reason, just lazy I guess 'cause it saves me the click for going into that, then going into that – and instead I'm like 'it's there on the desktop'.

Support Group 2 (Male): A lot of the time for me it's something that I'll only need access to for a limited time so once I'm done with it I'll just delete it.

Finally, experts believed that users would report any data breaches immediately. Employees, however, questioned how they would know if a data breach had occurred:

Support Group 2 (Female):"But how do you know that you've lost something? I'm not sure I would recognise a data loss unless it said to me, 'you've lost some data'."

This assumption highlights an important problem for experts: employees do not possess a concrete understanding of the consequences associated with cybersecurity – e.g. what actually happens when you have suffered a data breach? A possible remedy would appear to be for experts to contextualise advice and policy in order to encourage compliance.

4.5 Employee Misconceptions and Disagreements

Next, we explore employees' misconceptions about security behaviours and their failures to come to any agreement about appropriate actions. Firstly, staff believed that software updates – whether applications or an operating system – were primarily a means to access new features, arguing that updates could be delayed without any adverse impact, a finding previously reported by Vaniea et al. [59].

Additionally, they erroneously believed that if the update was important, it would get pushed through by the IT staff regardless. This misconception is in line with work showing how updating software was rarely seen as a key security behaviour [60].

Academic Group 2 (Female 2): "I've got the turn on automatic software updates, because I thought software was quite general and there's the other one that covers the anti-malware software – so any software updates could be anything. Uhm, that's why I thought it was not specific to internet security"

There was extended discussion regarding the threats from email attachments and links. While staff were generally aware that clicking or downloading items from emails could harm their computer, the exact nature of the harm was disputed. Some employees believed that links were more dangerous than attachments as clicking them automatically compromised the computer, while others argued that attachments were harmless if you did not allow them to install. While most points argued were true to an extent, it was worrying how varied their perspectives of the threats were.

Academic Group 2 (Female 1): "But if you opened it, I wouldn't anyway, but open an attachment from someone I didn't know – I would just delete it – but if I did open it I would assume that unless I clicked on a link within that attachment then the attachment couldn't, unless, you know like a Word attachment, if they sent me some kind of attachment that could be actually downloading a virus."

Academic Group 2 (Male 1): "I think an attachment is more important because that's a file that you download to your computer and could potentially run directly on your computer"

Academic Group 1 (Female 1): "to actually open an attachment itself may be important, because I know that you don't need to put your password in and malware starts to come, and there are many of those everyday. So if we put that as a priority behaviour then we can prevent a lot of malware from coming in. And it's very simple as well – that's my opinion."

Academic Group 2 (Female 2): "The more that I talk the more I realise I don't know"

This last observation is important. Employees lacked a good mental model of the nature of the threat and the way that they could realistically guard against it. This led to disagreements about the most effective forms of protection. For example, there were heated discussions about writing passwords down, with the majority of participants agreeing that it was a 'must not do' behaviour and should be avoided at all costs. In the meantime, password reuse was seen as a negative, but necessary, behaviour – especially given that mapping personal accounts to work accounts would be difficult for attackers. This demonstrates a mental model where most staff prioritise the need to protect themselves against colleagues rather than against external threats.

"Academic Group 2 (Female 3): See it's funny because I've put keep passwords safe if written down before I've put use strong passwords because obviously if you have it written down it doesn't matter how strong it is – people can get it.

Academic Group 2 (Male 1): But if you've got it written down there is maybe only a handful of corrupt people who could get their hands on it…"

While previous literature (e.g. [56]) has reported this as a flawed mental model, other work [14] argues that this might not actually be a serious security threat, while Zhang-Kennedy et al. [62] suggest that this rule should be changed, promoting the keeping of written down passwords secure. Again, these academic disagreements demonstrate the difficulties with generating security advice. Ultimately, both GCHQ and NIST have taken the stance of promoting secure storage of written down passwords in their new guidelines.

4.6 The Sources of Guidance

We now look at some of the issues around where employees would turn to for education and guidance. Firstly, as we noted, participants were reluctant to ask experts for advice as they felt it was time consuming and unnecessary. They seldom knew who they could turn to for advice either within the Department, the Faculty or the University. Participants generally agreed that learning from each other or from their own personal experience was more realistic than asking for advice from an expert:

Academic Group 2 (Male 1): "I think people are more likely to ask their immediate colleagues for advice about things."

Support Group 2 (Female 1): "Would you not just tend to ask for advice once you've done something wrong or something bad has happened?"

This is a problem when local knowledge is based upon poor mental models of both threat and effective deterrence. The tendency to rely upon peers and to trust social norms is a known problem in cybersecurity research, leading to the development of shadow security cultures within an organisation [36]. We know that teams do have an important role to play in the development of security behaviours, but we also know that these teams can appropriate security behaviours and practices, moulding them to better fit their own work context, but occasionally introducing vulnerabilities and misconceptions as a result [47].

Our employees felt that they could not be expected to stay on top of the latest advice and information. They were aware of certain 'rules' (such as not opening attachments and clicking on links) but they felt that they should not be held responsible for cyber defence as they could not be expected stay current with that knowledge and were unwilling to put extra time into learning.

Support Group 2 (Female 3): "Yeah, just come and ask us – spend an hour educating you. I mean, nobody has that time, so…"

Finally, employees recognised that certain issues were out of their control. They believed that the IT department was responsible for cyber defence and that this defence was primarily undertaken with automated detection and control systems. This is an interesting issue as it reflects the kinds of culture that evolves around staff who have restricted access in relation to installing or updating software. Knowing that IT services have control over such matters brings with it the assumption that staff have no real responsibilities in this area.

Academic Group 1 (Male 2): "So the anti-malware thing, because it's the university computer I just take it that's it's all sorted out anyway. It's not like you're meant to keep it up to date yourself personally."

Again, this speaks to the way that employees are empowered in the cybersecurity space. We know from the psychology literature on social loafing that in the presence of others, an individual user may not react to a request, assuming that others will make the required response [11, 22].

4.7 Feedback to Experts

The final step in the Cybersurvival Task was to present the findings to the university experts. Below we cover the lessons learnt from that session as well as feedback regarding the findings and the methodology.

Firstly, the experts were surprised at some of the misconceptions shown by employees. They had made assumptions that certain behaviours or terms were common knowledge, and the results of the exercise made them realise that extra effort was required to better understand their audience.

CISO: "It forced us to re-evaluate our desired behaviours. Because, I have, based on years of experience, developed a prejudice towards certain desired behaviours that I now think, based on this, perhaps I've allowed that prejudice to drive my own personal baseline. And I think this tool helps break that and forces me to re-evaluate my concept of desired behaviours."

Secondly, experts took the output from the Cybersurvival Task as evidence that their one-size-fits-all training approach was failing the university.

CISO: "One size fits all is a fallacy. It's not going to work. You need to cater your risk management programmes specifically to the people within their respective work areas. I think that's what I'm taking from this."

While this school of thought is not necessarily new for the academic security community, it is important to note that it is still being employed in organisations (this was a common finding across both Phases I & II). By utilising this tool, the CISO was able to make this realisation for himself and thus can seek more effective ways of promoting secure behaviours.

Thirdly, they argued that the task would be an excellent tool for establishing a baseline prior to undertaking training development and then using this baseline data to deliver more targeted training:

Faculty Deputy: "It's critical because this provides a mechanism for determining, not to find out whether our programmes are successful, but whether our programmes are correctly designed and catered for the intended audience. Because that's the initial hurdle. Because if the programme isn't adapted for the culture then it will fail"

Finally, experts expressed their support for the Cybersurvival Task, focusing on the fact that the issues raised by the tool were specific to their organisation:

CISO: "I don't think I've come across a tool that's quite so powerful. I've come across metrics. I've challenged metrics, but this tool is different because it's using my metrics that I've provided and compared them against other people's metrics to see how they match and there's no way I can argue against that data because it's data that I've provided, as an individual, and data that other people have provided. I can't see any weakness in there. I'm struggling to find a weakness. I think it's a very powerful tool"

We note here, that one useful aspect of the Cybersurvival Task is that the output for different staff groups can be easily quantified in terms of the kinds of visualisations shown in Figure 3. This was important as it is not easy to use purely qualitative data to illustrate discrepancies between the beliefs of different groups, but we found these illustrations, used in combination with the discussion data, were very effective as a means of organisation-specific highlighting issues.

4.8 Summary of Findings

The Phase II deployment of the Cybersurvival Task in a large institution involving two organisational security experts and 20 employees demonstrated the benefits of this tool by highlighting differences between the cybersecurity beliefs and attitudes of security experts and employees. We also found that a group discussion around desired security behaviours actually led to *less* agreement between employees and experts, which raises interesting questions regarding the social construction of cybersecurity within workgroups and related issues of how best to disseminate security information in organisations.

A follow up session with the organisational security experts found that they valued the information uncovered by the tool, and they had a clear understanding of how that information could be used to improve their organisation in the future – for example in understanding what content should be covered in mandatory training sessions.

5. DISCUSSION

In this paper, we described two deployments of the Cybersurvival Task in two large universities and showed how the task revealed security misconceptions of staff and some behavioural discrepancies between security experts and employees. We specifically highlight how the organisation's security experts were able to reflect upon flawed assumptions regarding certain employee behaviours, as well as realising how their approach to training was not fit for purpose. While the reported employee misconceptions are not all novel, it is important for the organisational experts to know which security issues exist within their realm so that they are able to address problematic behaviours or beliefs. Importantly, the fact that the task has highlighted some well-known behavioural issues serves as a sanity check that participants were being truthful and that the task is externally valid.

Here, we discuss the benefits of using the Cybersurvival Task over other existing security behaviour measurement tools and explore how organisations can use the tool to improve training programmes, tailor their security policies, and understand the development of non-compliant attitudes and shadow security behaviours. We should note that the Cybersurvival Task has two quite discrete functions. Firstly, in keeping with the Desert Survival task and the Moon Landing task, the Cybersurvival Task can highlight individual and group opinion differences between staff groups and see how they are resolved. Secondly, the task can produce useful cybersecurity data about staff behaviours, understanding and possible compliance with security policies. We will explore these functions in more detail below.

5.1 Measuring Individual and Group Decision-Making

In terms of the first function – to observe the processes of individual and group decision making – it was very interesting to note the differences between groups within the organisation, but perhaps more intriguing to note that group discussion *never* resulted in more secure rankings overall, when compared to individual rankings. Indeed, in the case of support staff, group discussion resulted in a set of beliefs that were less secure (i.e. less aligned with expert opinion). Earlier we talked about this in relation to the development of a shadow security culture within the organisation in which social norms can come to dominate [36]. However, we should also note that this resonates with other studies using ranking tasks to measure group behaviour, when the dynamics of the group can result in sub-optimal decisions. For example, in a 'Desert Survival' study involving mixed gender groups, expertise tended to be ignored in group settings if the experts were women, resulting in poor group performance, but not if they were men [57].

While it is certainly interesting to observe the differences between individual and group scores, some may argue that cybersecurity is predominantly an individual task. We disagree given the social nature of organisations and the data suggesting that users are more likely to turn to colleagues rather than experts for advice. However, it is possible to build the visual representations (e.g. Figure 3) using the individual scores, although we would recommend running the group ranking sessions regardless due to the insights they generate (see below).

5.2 Measuring Cybersecurity Attitudes and Behaviours

In terms of the second function of the task – to measure cybersecurity attitudes and behaviours – we should ask how the Cybersecurity Task compares with other available measures. The most obvious point of comparison – albeit serving a different purpose – is the Security Behaviour Intentions Scale (SeBIS) which was initially developed in 2015 with the aim of becoming the standard tool for assessing the security behaviours of end-users [19]. This has since been validated to show how some security behaviours can be reliably predicted using the scale [18]. One of the interesting differences between SeBIS (and self-reporting

questionnaires in general) and the Cybersurvival Task is the request in the latter to *rank* behaviours, rather than indicate compliance level. There is no obvious 'correct' ranking and so we can attenuate the problem of 'social desirability' (giving the 'right' answers to questions irrespective of behaviour). Additionally, by having to justify priorities, participants in the Cybersurvival Task reveal underlying assumptions and/or flawed mental models that can then be used by experts to deliver appropriate remediation.

However, there are two further issues that come to light when comparing tasks. The SeBIS is not resource heavy – it can be completed quickly and can therefore give organisations rapid, actionable data about the beliefs and reported behaviours of their staff. In contrast, the Cybersurvival Task when done properly (involving both individual and group stages) can be quite resource intensive but also allows for training opportunities while also providing a baseline measure of security knowledge within the organisation. This is not a negative thing if it results in greater understanding and ownership of the problem. In addition, the SeBIS has a static set of items to be used in any organisation, despite the fact that there are always disagreements over what items should be included and prioritised depending on the context (e.g. [32]). In contrast, the Cybersurvival Task, as we have described it, sees cybersecurity as an evolving process and adapts this list to those set up by a specific organisation. At the beginning of our study, we asked the organisation's CISO to generate 19 important behaviours and compared these with the 16 items on SeBIS [19] and the 20 "good" behaviours identified by Ion et al. [32]. There was a substantial overlap, but our CISO added certain behaviours (*ask for advice* and *educate yourself on how to avoid fraud)* which he ranked very highly. Staff did not prioritise these items and so it would be easy to argue that they were unimportant, but this would be missing the point. The Cybersurvival Task is designed to show differences between the beliefs and opinions held by the CISO and those held by employee groups throughout the organisation. Where there is disagreement, then there is an opportunity to consider whether staff communication has been adequate or whether expectations are unrealistic.

5.3 How Can Organisations Benefit from the Cybersurvival Task?

It is clear from our expert feedback session that security experts in organisations make assumptions about their institution's security culture and that these assumptions are not always correct. This means that organisations may not be providing staff with the necessary and/or relevant training programmes. While the Cybersurvival Task does not measure employee compliance – it is possible that employees engage in all behaviours on the list – it can be used to obtain a snapshot of security subcultures within an organisation, and to identify any misinformation that might be circulating in those subcultures. This would allow experts the opportunity to tailor solutions that would help prevent the proliferation of non-compliant security practices.

The Cybersurvival Task can also serve as a sanity check for an organisation's security policies. During the first step when the organisation's security experts modify and rank the list of behaviours, they can identify any policy items that may no longer apply, or others that they may not have considered before. Additionally, this process should make experts aware of what the most important message to staff should be. The act of having to rank a particular behaviour as first or second on a list can give pause for thought – how are these important behaviours being communicated to staff across the organisation? Note, too, that rankings may change in keeping with the dynamic cybersecurity threat landscape.

Lastly, it may be possible to use the Cybersurvival Task as a training tool, exploiting the way it can readily highlight misconceptions and promote discussions about why the experts prioritise certain behaviours and why staff might find these behaviours challenging to execute in their own work contexts. While such an approach would require a greater degree of co-ordination (e.g. scheduling for both employees and experts), the direct outcome with regards to mutual understanding by both parties would seem to be beneficial. The task certainly generated high levels of engagement across all groups – something which is not always said of cybersecurity training material.

5.4 Limitations and Future Work

The most obvious limitation regarding this implementation of the Cybersurvival Task related to the time taken to conduct the workshops and collate and present the findings. Despite our participants finding it an enjoyable task, we do recognise that length could be an issue for both organisations and individuals. In future settings, the individual rankings could be completed online and analysed before the group meeting to discuss differences and agree a consensus ranking (thus speeding up the process). We are hesitant to suggest running the complete task online as it is currently presented, as this would miss out on valuable qualitative data that shows the reasoning behind the rankings and reveals any underlying misconceptions or erroneous mental models that management can then address. However, it may be possible to redesign some of the activities (e.g. the group ranking task) to accommodate digital technologies for carrying out the workshops in a distributed manner and reducing the time taken to complete them.

We also recognise that our deployments have been restricted to academic organisations and so, in future work, we aim to take the tool into other sectors, streamlining some aspects of the data collection process, and exploring the automatic generation of reports.

5.5 Conclusions

In this paper, we have shown that security experts and staff do not always agree on the most important security behaviours and this will be a big concern for organisations. Ideally, all members of an organisation should be working towards the same security goals and should understand their role in achieving those goals, yet we have found that group discussions on cybersecurity behaviours in fact led to more disagreement between staff and expert priorities. We have shown that a simple ranking task, conducted individually and then in groups, can highlight such disagreements and illustrate the different normative beliefs held by specific staff groups as well as illustrating the differing priorities shown by security experts and employees at different levels of the organisation. We believe the Cybersurvival Task would be useful for any CISO seeking to understand the kinds of sub-optimal security subcultures that develop within their organisation.

6. REFERENCES

[1] A Hacker's Dream: American Password Reuse Runs Rampant: 2017. *https://www.infosecurity-magazine.com/news/american-password-reuse-runs/.*

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

[3] Alotaibi, M., Furnell, S. and Clarke, N. 2016. Information Security Policies : A review of Challenges and Influencing Factors. *The 11th International Conference for Internet Technology and Secured Transactions* (Dec. 2016), 352–358.

[4] Ashburn-Nardo, L. and Johnson, N.J. 2008. Implicit outgroup favoritism and intergroup judgment: The moderating role of stereotypic context. *Social Justice Research*. 21, 4 (Dec. 2008), 490–508.

[5] Bada, M. and Sasse, M.A. 2014. *Cyber Security Awareness Campaigns Why do they fail to change behaviour ?*

[6] Baxter, J. 2002. Jokers in the Pack: Why Boys are More Adept than Girls at Speaking in Public Settings. *Language and Education*. 16, 2 (Jun. 2002), 81–96.

[7] Beautement, A., Becker, I., Parkin, S., Krol, K. and Sasse, M.A. 2016. Productive Security: A scalable methodology for analysing employee security behaviours. *Proceedings of the Symposium On Usable Privacy and Security (SOUPS)* (2016), 1–18.

[8] Beautement, A., Sasse, M.A. and Wonham, M. 2008. The compliance budget. *Proceedings of the 2008 workshop on New security paradigms - NSPW '08* (New York, New York, USA, 2008), 47.

[9] Blythe, J.M., Coventry, L. and Little, L. 2015. Unpacking security policy compliance : The motivators and barriers of employees ' security behaviors. *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)* (2015), 103–122.

[10] Bollmer, J.M., Harris, M.J., Milich, R. and Georgesen, J.C. 2003. Taking Offense: Effects of Personality and Teasing History on Behavioral and Emotional Reactions to Teasing. *Journal of Personality*. 71, 4 (Jun. 2003), 557–603.

[11] Briggs, P., Jeske, D. and Coventry, L. 2017. The design of messages to improve cybersecurity incident reporting. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Jul. 2017), 3–13.

[12] Bursztein, E., Margolis, D., Archer, A., Pitsillidis, A. and Savage, S. 2014. Handcrafted Fraud and Extortion : Manual Account Hijacking in the Wild. *In Proceedings of the 2014 Conference on Internet Measurement Conference* (2014), 347–358.

[13] Carlton, M. and Levy, Y. 2015. Expert assessment of the top platform independent cybersecurity skills for non-IT professionals. *Conference Proceedings - IEEE SOUTHEASTCON* (Apr. 2015), 1–6.

[14] Cheswick, W. 2013. Rethinking Passwords. *Communications of the ACM*. 56, 2 (Feb. 2013), 40–44.

[15] Cooke, R.A. and Kernaghan, J.A. 1987. Estimating the Difference Between Group Versus Individual Performance on Problem-Solving Tasks. *Group & Organization Management*. 12, 3 (Sep. 1987), 319–342.

[16] Dembo, M.H. and McAuliffe, T.J. 1987. Effects of perceived ability and grade status on social interaction and influence in cooperative groups. *Journal of Educational Psychology*. 79, 4 (1987), 415–423.

[17] Dudley, M.G. and Harris, M.J. 2003. To think or not to think: The moderating role of need for cognition in expectancy-consistent impression formation. *Personality and Individual Differences*. 35, 7 (Nov. 2003), 1657–1667.

[18] Egelman, S., Harbach, M. and Peer, E. 2016. Behavior Ever Follows Intention?: A Validation of the Security Behavior Intentions Scale (SeBIS). *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 5257–5261.

[19] Egelman, S. and Peer, E. 2015. Scaling the Security Wall: Developing a Security Behavior Intentions Scale (SeBIS). *Proceedings of the ACM CHI'15 Conference on Human Factors in Computing Systems* (2015), 2873–2882.

[20] Fogg, B.J. 2002. Persuasive technology. *Ubiquity*. 2002, December (Dec. 2002), 2.

[21] Furnell, S. and Thomson, K.L. 2009. From culture to disobedience: Recognising the varying user acceptance of IT security. *Computer Fraud and Security*. 2009, 2 (Feb. 2009), 5–10.

[22] George, J.M. 1992. Extrinsic and Intrinsic Origins of Perceived Social Loafing in Organizations. *Academy of Management Journal*. 35, 1 (Mar. 1992), 191–202.

[23] Giessner, S.R., van Knippenberg, D., van Ginkel, W. and Sleebos, E. 2013. Team-oriented leadership: The interactive effects of leader group prototypicality, accountability, and team identification. *Journal of Applied Psychology*. 98, 4 (2013), 658–667.

[24] Giessner, S.R. and Mummendey, A. 2008. United we win, divided we fail? Effects of cognitive merger representations and performance feedback on merging groups. *European Journal of Social Psychology*. 38, 3 (Apr. 2008), 412–435.

[25] Government Communications Headquarters 2015. *Simplifying Your Approach: Password Guidance.*

[26] Grawemeyer, B. and Johnson, H. 2011. Using and managing multiple passwords: A week to a view. *Interacting with Computers*. 23, 3 (Mar. 2011), 256–267.

[27] Guerin, L. 2007. *Smart Policies for Workplace Technologies: Email, Social Media, Cell Phones & More*. Nolo, Berkeley, CA.

[28] Harris, M.J. and Perkins, R. 1995. Effects of distraction on interpersonal expectancy effects : A social interaction test of the cognitive busyness hypothesis. *Social Cognition*. 13, 2 (Jun. 1995), 163–182.

[29] Herath, T. and Rao, H.R. 2009. Protection motivation and deterrence: a framework for security policy compliance in organisations. *European Journal of Information Systems*. 18, 2 (Apr. 2009), 106–125.

[30] Herley, C. 2009. So long, and no thanks for the externalities. *Proceedings of the 2009 workshop on New security paradigms workshop - NSPW '09* (2009), 133.

[31] Ifinedo, P. 2014. Information systems security policy compliance: An empirical study of the effects of socialisation, influence, and cognition. *Information and Management*. 51, 1 (Jan. 2014), 69–79.

[32] Ion, L., Reeder, R. and Consolvo, S. 2015. "... no one can hack my mind": Comparing Expert and Non-Expert Security Practices. *Symposium on Usable Privacy and Security* (2015), 327–346.

[33] Kiahr, R., Shah, J.N., Sheriffs, P., Rossington, T., Pestell, G., Button, M. and Wang, V. *Cyber Security Breaches Survey 2017*.

[34] Kirlappos, I., Parkin, S. and Sasse, M. 2015. Shadow security as a tool for the learning organization. *ACM SIGCAS Computers and Society*. (2015).

[35] Kirlappos, I., Parkin, S. and Sasse, M.A. 2014. Learning from "Shadow Security": Why understanding non-compliant behaviors provides the basis for effective security. *Usec '14* (2014), 1–10.

[36] Kirlappos, I., Parkin, S. and Sasse, M.A. 2015. "Shadow security" as a tool for the learning organization. *ACM SIGCAS Computers and Society*. 45, 1 (2015), 29–37.

[37] Kolkowska, E. and Dhillon, G. 2013. Organizational power and information security rule compliance. *Computers and Security*. 33, (Mar. 2013), 3–11.

[38] Lafferty, J.C., Eady, P.M. and Elmers, J. 1974. The desert survival problem. *Experimental Learning Methods*. (1974).

[39] Lee, C., Lee, C.C. and Kim, S. 2016. Understanding information security stress: Focusing on the type of information security compliance activity. *Computers and Security*. 59, (Jun. 2016), 60–70.

[40] Li, I., Forlizzi, J., Dey, A. and Kiesler, S. 2007. My agent as myself or another. *Proceedings of the 2007 conference on Designing pleasurable products and interfaces - DPPI '07* (New York, New York, USA, Aug. 2007), 194.

[41] Li, L., Xu, L., He, W., Chen, Y. and Chen, H. 2016. Cyber security awareness and its impact on employee's behavior. *Lecture Notes in Business Information Processing* (Dec. 2016), 103–111.

[42] Littlepage, G., Robison, W. and Reddington, K. 1997. Effects of Task Experience and Group Experience on Group Performance, Member Ability, and Recognition of Expertise. *Organizational Behavior and Human Decision Processes*. 69, 2 (Feb. 1997), 133–147.

[43] McAninch, C.B., Milich, R. and Harris, M.J. 1996. Effects of an Academic Expectancy and Gender on Students' Interactions. *The Journal of Educational Research*. 89, 3 (Jan. 1996), 146–153.

[44] Ohtsubo, Y. and Masuchi, A. 2004. Effects of Status Difference and Group Size in Group Decision Making. *Group Processes & Intergroup Relations*. 7, 2 (Apr. 2004), 161–172.

[45] Ovelgönne, M., Dumitras, T., Prakash, B.A., Subrahmanian, V.S. and Wang, B. 2017. Understanding the Relationship between Human Behavior and Susceptibility to Cyber Attacks. *ACM Transactions on Intelligent Systems and Technology*. 8, 4 (Mar. 2017), 1–25.

[46] Pahnila, S., Siponen, M. and Mahmood, A. 2007. Employees' behavior towards IS security policy compliance. *Proceedings of the Annual Hawaii International Conference on System Sciences* (Jan. 2007), 156b–156b.

[47] Parkin, S. and Krol, K. 2015. Appropriation of security technologies in the workplace. *Presented at Experiences of Technology Appropriation: Unanticipated Users, Usage, Circumstances and Design* (Oslo, Norway, 2015).

[48] Parsons, K., Calic, D., Pattinson, M., Butavicius, M., McCormac, A. and Zwaans, T. 2017. The Human Aspects of Information Security Questionnaire (HAIS-Q): Two further validation studies. *Computers & Security*. 66, (May 2017), 40–51.

[49] Pattabiraman, A., Srinivasan, S. and Swaminathan, K. 2018. Fortifying Corporate Human Wall: A Literature Review of Security Awareness and Training. *Information Technology Risk Management and Compliance in Modern Organizations*. (2018), 142–175.

[50] Phishing Across the Pond: 70% of U.K. Universities Impacted: 2017. *https://duo.com/blog/phishing-across-the-pond-70-percent-of-uk-universities-impacted*.

[51] PwC 2017. *The Global State of Information Security® Survey 2017*.

[52] Rus, D., van Knippenberg, D. and Wisse, B. 2010. Leader power and leader self-serving behavior: The role of effective leadership beliefs and performance information. *Journal of Experimental Social Psychology*. 46, 6 (Nov. 2010), 922–933.

[53] Schneier, B. 2000. *Secrets and Lies*. John Wiley & Sons.

[54] Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L.F. and Downs, J. 2010. Who falls for phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions. *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10* (2010), 373–382.

[55] Stewart, G. and Lacey, D. 2012. Death by a thousand facts. *Information Management & Computer Security*. 20, 1 (Mar. 2012), 29–38.

[56] Stobert, E. and Biddle, R. 2014. The password life cycle: User behaviour in managing passwords. *SOUPS '14: Proceedings of the Tenth Symposium On Usable Privacy and Security* (2014), 243–255.

[57] Thomas-Hunt, M.C. and Phillips, K.W. 2004. When What You Know Is Not Enough: Expertise and Gender Dynamics in Task Groups. *Personality and Social Psychology Bulletin*. 30, 12 (Dec. 2004), 1585–1598.

[58] Toward a Group Facilitation Technique for Project Teams: 2007. *http://gpi.sagepub.com/content/10/3/299.full.pdf*. Accessed: 2016-03-15.

[59] Vaniea, K.E., Rader, E. and Wash, R. 2014. Betrayed by updates. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (New York, New York, USA, 2014), 2671–2674.

[60] Vitale, F., Mcgrenere, J., Tabard, A., Wendy, M.B., Lyon, U., Paris-saclay, U. and Umr, C. 2017. High Costs and Small Benefits : A Field Study of How Users Experience Operating System Upgrades. *CHI 2017* (New York, New York, USA, 2017), 4242–4253.

[61] Wash, R., Rader, E. and Fennell, C. 2017. Can People Self-Report Security Accurately? *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17* (2017), 2228–2232.

[62] Zhang-Kennedy, L., Chiasson, S. and Van Oorschot, P. 2016. Revisiting password rules: Facilitating human management of passwords. *eCrime Researchers Summit, eCrime* (Jun. 2016), 81–90.

APPENDIX

A. Initial set of behaviours as presented to experts in Phase I.

The original list of behaviours was obtained from Ion et al. (2015). Below they are presented unranked as seen by the security experts in Phase I.

Behaviour
Be suspicious of links
Be sceptical of everything
Turn on automatic updates
Save passwords in a file
Clear browser cookies
Use a password manager
Use 2-factor authentication
Check if HTTPS
Look at the URL Bar
Install OS Updates
Don't click on links from unknown people
Use strong passwords
Use unique passwords
Don't write down passwords
Visit only known websites
Don't open email attachments from unknown people
Update applications
Don't enter passwords on links in emails
Use antivirus software

B. Ranked list of behaviours agreed by experts in Phase I.

Our two security experts from Phase I were given the opportunity to add, remove, and rename behaviours from the original list (Appendix A). Below is the final agreed rank list from experts for Phase I.

Ranking	Behaviour
1	Use strong passwords
2	Use antivirus software
3	Turn on auto software updates
4	Check every message is genuine
5	Keep OS up to date
6	Be aware of fake phone calls
7	Use different passwords
8	Be suspicious of links
9	Ask for advice when unsure
10	Check URL bar
11	Check if HTTPS
12	Don't download attachments from unknown senders
13	Don't enter password on website from link
14	Don't click links from unknown senders
15	Update applications
16	Only visit known websites
17	Don't write down passwords
18	Use a password manager
19	Use 2 factor authentication
20	Clear cookies

C. Scatter Plots Comparing Expert and Staff Rankings

Here we present the higher quality versions of the scatter plots from Figure 3. These are omitted from the paper due to space.

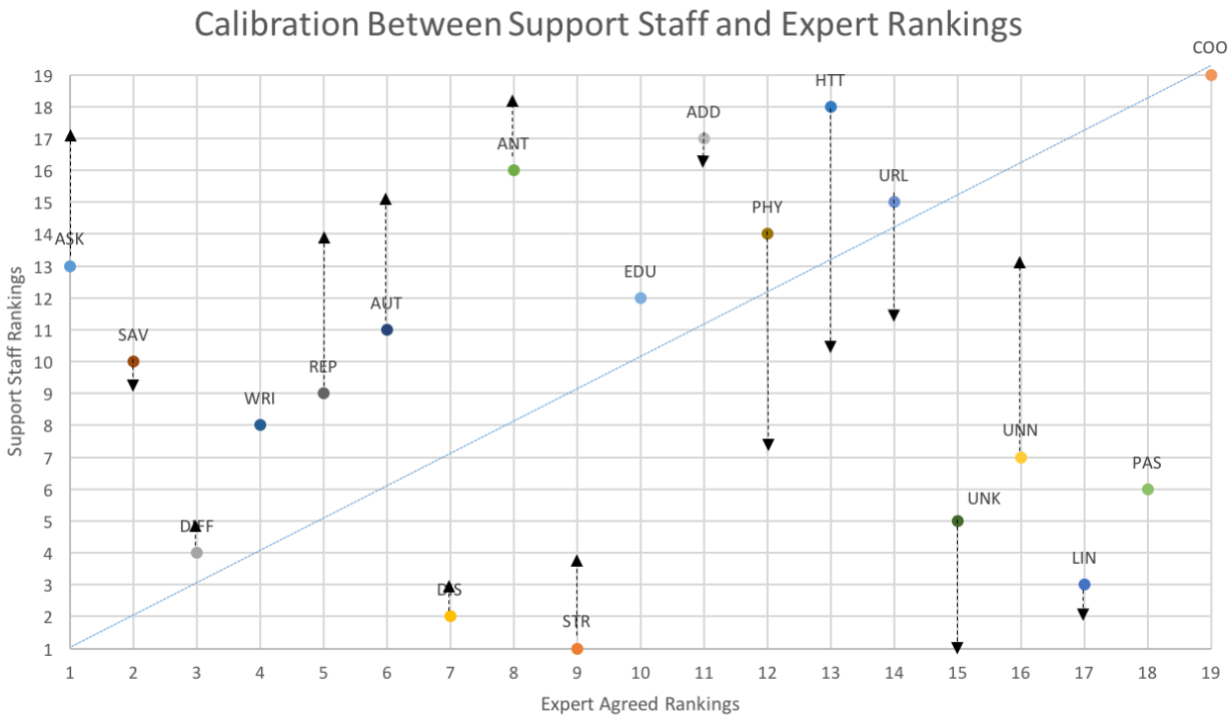


Figure C.1: Scatter Plots comparing the rankings of experts (X-axis) against support staff. Arrows show the shift from mean individual rankings to final group rankings (dots).

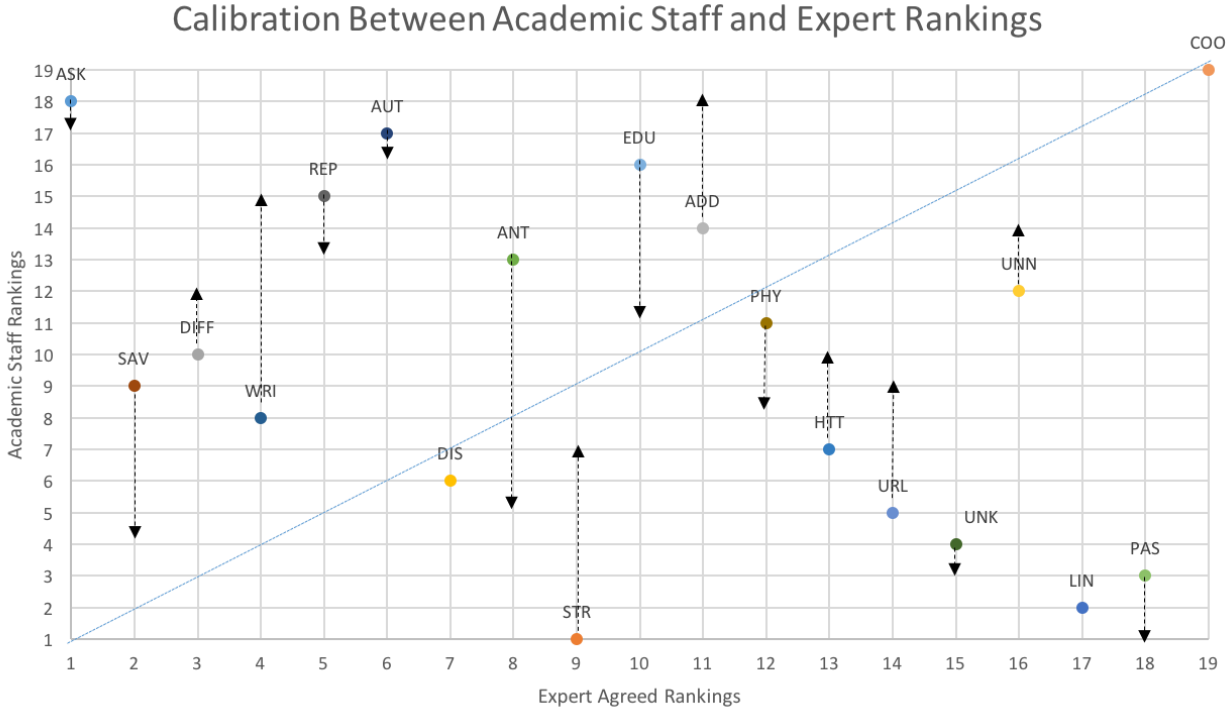


Figure C.1: Scatter Plots comparing the rankings of experts (X-axis) against academic staff. Arrows show the shift from mean individual rankings to final group rankings (dots).