Resolving the Privacy Paradox: Bridging the Behavioral Intention Gap with Risk Communication Theory

Justin Chun Wah Wu
Brigham Young University

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

Part of the Physical Sciences and Mathematics Commons

BYU ScholarsArchive Citation
Wu, Justin Chun Wah, "Resolving the Privacy Paradox: Bridging the Behavioral Intention Gap with Risk Communication Theory" (2019). Theses and Dissertations. 8702.
https://scholarsarchive.byu.edu/etd/8702

This Dissertation is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.
Resolving the Privacy Paradox: Bridging the Behavioral Intention Gap with Risk Communication Theory

Justin Chun-Wah Wu

A dissertation submitted to the faculty of Brigham Young University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Daniel Zappala, Chair
Kent Seamons
Michael Jones
Bryan Morse
Parris Egbert

Department of Computer Science
Brigham Young University

Copyright © 2019 Justin Chun-Wah Wu
All Rights Reserved
ABSTRACT

Resolving the Privacy Paradox: Bridging the Behavioral Intention Gap with Risk Communication Theory

Justin Chun-Wah Wu
Department of Computer Science, BYU
Doctor of Philosophy

The advent of the Internet has led to vastly increased levels of data accessibility to both users and would-be attackers. The privacy paradox is an established phenomenon wherein users express concern about resultant security and privacy threats to their data, but nevertheless fail to enact the host of protective measures that have steadily become available. The precise nature of this phenomenon, however, is not a settled matter. Fortunately, risk communication theory, a discipline devoted to understanding the factors involved in risk-oriented decision-making and founded in years of empirical research in public health and disaster awareness domains, presents an opportunity to seek greater insight into this problem.

In this dissertation, we explore the application of principles and techniques from risk communication theory to the question of factors in the grassroots adoption of secure communication technologies. First, we apply a fundamental first-step technique in risk communication—mental modeling—toward understanding users’ perceptions of the structure, function, and utility of encryption in day-to-day life. Second, we apply principles of risk communication to system design by redesigning the authentication ceremony and its associated messaging in the Signal secure messaging application. Third, we evaluate the applicability of a core decision-making theory—protection motivation theory—toward the problem of secure email adoption, and then use this framework to describe the relative impact of various factors on secure email adoption. Finally, we evaluate perceptions of risk and response with respect to the adoption of secure email features in email scenarios of varying sensitivity levels.

Our work identifies positive outcomes with respect to the impact that risk messaging has on feature adoption, and mixed results with respect to comprehension. We highlight obstacles to users’ mental interactions with encryption, but offer recommendations for progress in the adoption of encryption. We further demonstrate that protection motivation theory, a core behavioral theory underlying many risk communication approaches, has the ability to explain the factors involved in users’ decisions to adopt or not adopt in a way that can at least partially explain the privacy paradox phenomenon. In general, we find that the application of even basic principles and techniques from risk communication theory do indeed produce favorable research outcomes when applied to this domain.

Keywords: privacy, usable security, risk communication, secure communication, security adoption
I would like to first thank my advisor, Daniel Zappala, for being supportive of my research interests and aims, even when they took me in a direction that was unfamiliar to him. His encouragement and assistance, from start to finish, were integral in my becoming who I am today. He has been every bit the mentor you would wish an advisor to be.

I also thank all the students who have aided me in this long process, whether indirectly as friends and colleagues, or more directly, by collaborating with me in my research: Mark O’Neill, Elham Vaziripour, Scott Heidbrink, Jordan Whitehead, Devon Howard, Jake Tyler, Cyrus Gatrell, Calvin Woods, Ammon Mugimu, James Conners, Scott Ruoti, and Ruba Abu-salma.

Special thanks go to my family, who have supported me this long while; my parents, my brothers, and especially my wife.

Finally, I wish to express my gratitude to everyone who has made active effort to accommodate someone who might well be the black sheep of the department: the members of my committee and, particularly, Jen Bonnett, who has never been anything but fiercely supportive of the graduate students in this department.
Table of Contents

List of Figures x

List of Tables xiii

1 Introduction 1

1.1 Understanding and enabling users .......................................................... 2
1.2 Risk communication theory .................................................................... 5
1.3 Applying risk communication principles to the adoption of encrypted communication 8
  1.3.1 Extending the literature .................................................................... 8
  1.3.2 Chapter 1: understanding perceptions of encryption ............................ 9
  1.3.3 Chapter 2: applying principles of risk communication to system design in the Signal secure messaging app ............................... 11
  1.3.4 Chapter 3: evaluating the applicability of protection motivation theory to the adoption of secure email and the relative influence of motivating factors in decisions to adopt ......................................................... 12
  1.3.5 Chapter 4: studying perceptions of risk and response in the adoption of secure email features ......................................................... 14

2 When is a tree really a truck? Exploring mental models of encryption 17

2.1 Introduction ............................................................................................. 18
2.2 Methodology ........................................................................................... 22
  2.2.1 Recruitment ....................................................................................... 22
  2.2.2 Demographics .................................................................................. 23
2.2.3 Study design ................................................. 23
2.2.4 Data analysis ............................................. 27

2.3 Results ........................................................ 28
2.3.1 Impressions about encryption .......................... 28
2.3.2 Mental models of encryption ........................... 29
2.3.3 The role of encryption .................................... 39

2.4 Discussion ..................................................... 44
2.4.1 Encryption is access control with a symmetric model 44
2.4.2 Grassroots adoption is also a public relations battle 46
2.4.3 Confusion about encryption strength ................... 47
2.4.4 Learning how encryption works doesn’t help ........... 48
2.4.5 Consider the context ..................................... 49

2.5 Limitations ..................................................... 50

2.6 Related work .................................................. 51
2.6.1 Encryption usability ..................................... 51
2.6.2 Mental modeling in computer security ................. 53

2.7 Conclusion ...................................................... 53

2.8 Contributions .................................................. 54

2.9 Recommendations ........................................... 56

2.10 Future work ................................................... 57

2.11 Acknowledgments ............................................. 59

3 “Something isn’t secure, but I’m not sure how that translates into a problem”: Promoting autonomy by designing for understanding in Signal 60

3.1 Introduction ................................................... 61

3.2 Related work ................................................... 64
3.2.1 Protection motivation theory ............................ 64
3.2.2 Risk communication ..................................... 65
3.3 Evaluating warnings in Signal ........................................... 66
3.3.1 Cognitive walkthrough ........................................... 67
3.3.2 User study #1: Methodology ................................. 69
3.3.3 User study #1: results ......................................... 73
3.4 Developing improvements ............................................... 77
3.4.1 Visual indicator .................................................. 77
3.4.2 Notification and dialogs .................................... 79
3.4.3 Notification flow .............................................. 81
3.5 Evaluating the effectiveness of our redesign .................. 83
3.5.1 Results .......................................................... 84
3.6 Discussion .......................................................... 87
3.6.1 Risk communication gives users the ability to make personal trade-offs
between perceived risk and response cost. ....................... 87
3.6.2 Users’ strategies for coping with online threats extend beyond the ecosystem
of your app. .......................................................... 88
3.7 Limitations .......................................................... 89
3.8 Conclusion .......................................................... 90
3.9 Acknowledgments .................................................... 91

4 To encrypt or not to encrypt: exploring perceptions of secure email through the lens
of protection motivation theory ........................................ 92
4.1 Introduction .......................................................... 94
4.2 Related work ........................................................ 97
4.3 Methodology ........................................................ 99
4.3.1 Instrument development .................................... 99
4.3.2 Study sample ................................................ 101
4.4 Questionnaire results ............................................. 101
4.4.1 Risk likelihood ............................................... 102
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4.2</td>
<td>Risk severity</td>
<td>103</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Response efficacy</td>
<td>104</td>
</tr>
<tr>
<td>4.4.4</td>
<td>Self-efficacy</td>
<td>107</td>
</tr>
<tr>
<td>4.5</td>
<td>Analysis and modeling</td>
<td>110</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Analysis – approach</td>
<td>110</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Analysis – validating assumptions</td>
<td>111</td>
</tr>
<tr>
<td>4.5.3</td>
<td>Secure email analysis – factorial structure</td>
<td>112</td>
</tr>
<tr>
<td>4.5.4</td>
<td>Analysis – SEM analysis</td>
<td>113</td>
</tr>
<tr>
<td>4.5.5</td>
<td>Self-filtering analysis</td>
<td>116</td>
</tr>
<tr>
<td>4.6</td>
<td>Discussion</td>
<td>116</td>
</tr>
<tr>
<td>4.7</td>
<td>Limitations</td>
<td>119</td>
</tr>
<tr>
<td>4.8</td>
<td>Conclusion</td>
<td>120</td>
</tr>
<tr>
<td>5</td>
<td>Exploring the “secure” in secure email: perceptions of risk and response in the adoption of secure email features</td>
<td>121</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>122</td>
</tr>
<tr>
<td>5.2</td>
<td>Related Work</td>
<td>126</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Adoption of encryption</td>
<td>126</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Risk communication</td>
<td>128</td>
</tr>
<tr>
<td>5.3</td>
<td>System design</td>
<td>129</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Secure email features</td>
<td>131</td>
</tr>
<tr>
<td>5.4</td>
<td>Methodology</td>
<td>133</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Sample population</td>
<td>134</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Study protocol</td>
<td>135</td>
</tr>
<tr>
<td>5.5</td>
<td>Results</td>
<td>138</td>
</tr>
<tr>
<td>5.5.1</td>
<td>Feature use</td>
<td>138</td>
</tr>
<tr>
<td>5.5.2</td>
<td>General email use</td>
<td>139</td>
</tr>
<tr>
<td>5.5.3</td>
<td>Scenario: Inviting coworker to lunch</td>
<td>140</td>
</tr>
</tbody>
</table>
C Study materials for “Exploring the ‘Secure’ in Secure Email: Perceptions of Risk and Response in the Adoption of Secure Email Features”  

C.1 Study coordinator instructions .......................... 169
C.2 Roleplay instructions ....................................... 171
C.3 Post-task questionnaire ................................. 175

References .......................... 179
List of Figures

2.1 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. .................................................. 18

2.2 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. ........................................ 28

2.3 Edward’s barrier diagram, which represents “the symbolism of locking.” The skull is “just blocking, something that’s just blocking the message.” ..................... 31

2.4 Wally’s black box encryption. The data to be encrypted goes through a “layer of encryption” that runs “in the background.” ........................................ 33

2.5 Eva’s watermark diagram. ........................................ 33

2.6 Allen’s ciphered cloud. “The picture is split into different sections or elements and each element turns into a different element” ................................. 36

2.7 Franklin’s iterative text encryption process. ................................. 37

3.1 Signal notifications when safety numbers differ, depending on the internal state of the application. .................................................. 67

3.2 New notification dialogs, framing the authentication ceremony as a privacy check and using risk communication principles. ................................. 80

3.3 New authentication ceremony success and failure dialogs. ..................... 81

4.1 Risk likelihood Likert-items. .................................................. 102
4.2 Risk severity Likert-items. .................................................. 104
4.3 Response efficacy Likert-items for self-filtering. .......................... 105
4.4 Response efficacy Likert-items for secure email. .......................... 106
4.5 Self-efficacy Likert-items for self-filtering. ................................. 108
4.6 Self-efficacy Likert-items for secure email. ................................. 109
4.7 The final SEM model path diagram with unstandardized coefficients. Leftmost values, leading into the items, are error terms representing the variance in each item not attributed to the associated constructs. Values sitting on the reverse arrows leading from the constructs to each item are the factor loadings; the first item for each construct is normalized to 1.0 and the magnitude of other items’ factor loadings are relative to this value. The final values leading from the constructs to “User” are the SEM coefficients. Numbers in parentheticals display the standard error for the corresponding value. .................................................. 114

5.1 Our compose-screen interface. This allows access to both our encryption and self-destructing message features. .......................................................... 130
5.2 The self-destructing feature shown within a thread. The top message has already expired, and its body has been replaced with an indicator, while subsequent messages have active timers. .................................................. 133
5.3 Feature use per participant per scenario. ....................................... 138
5.4 Risk perception of normal email versus the response efficacy of encrypted email and self-destructing messages. ............................................. 145

B.1 Flow diagram depicting the how Signal reacts to a safety number change. . . . . . 164
B.2 Flow diagram depicting how our redesigned Signal reacts to a safety number change. The blue box encloses the elements and choices with analogous equivalents in the original Signal client. The area contained by the dashed lines shows choices, elements, and state changes that we added in our version that are expansions on the authentication ceremony and beyond. ........................................ 164

B.3 Privacy check notification flow ........................................ 166
B.4 Phone call and QR code privacy check results .................... 167
B.5 Conversation log notifications ........................................ 167
B.6 Privacy check indicator ........................................ 167
B.7 Privacy check screen ........................................ 167
List of Tables

3.1 Comparison of icons a 10-point Likert scale ........................................ 78

3.2 Comparison of participant understanding of the authentication ceremony using
    Signal and our redesign. ................................................................. 86

4.1 List of questionnaire items given participants. Each item corresponds to one of the
    four PMT categories. The final column indicates which items were included in the
    final SEM analysis; items excluded from final analysis have the reasons for their
    removal described. ................................................................. 100
Chapter 1

Introduction

The advent of the Internet, the cloud, and mobile devices has led to unprecedented levels of connectivity and data mobility. However, this increased connectivity has not been solely beneficial; data is easier to access than ever and unfortunately, that does not apply just to “intended parties.” The last decade has seen numerous massive data breaches such as the infamous Equifax hack of 2017, an attack that lasted for 76 days and compromised the data of 148 million people.1 Similarly, we have been confronted with a seemingly endless array of reports of innovative ways in which traditional expectations of the privacy of personal data are being subverted: for example, some have interpreted the position of the popular social media platform Facebook as “blam[ing] users for destroying their own right to privacy.”2

It is within this context that consumer-facing privacy enhancing technologies—tools designed to allow users to reclaim control over the security and privacy of their personal data—have similarly become increasingly more commonplace. Nevertheless, the overall adoption rates of privacy tools have remained low.3 Moreover, platforms and services that have been discovered to

---


2Sam Biddle, In Court, Facebook Blames Users for Destroying Their Own Right to Privacy, The Intercept, 14 June 2019, https://theintercept.com/2019/06/14/facebook-privacy-policy-court/

mine and sell the data of their users have continued to thrive despite stated outrage at their practices.\textsuperscript{4} This discrepancy between users’ reported distaste at the ways in which the privacy of their personal data has been eroded, contrasted against their failure to adopt privacy-preserving tools and practices is an established phenomenon, coined the \textit{privacy paradox}.

There have been many attempts to unravel the nature of the privacy paradox [48]. Indeed, some have suggested that the term itself is a misnomer, an oversimplification of a complex issue.\textsuperscript{5} For example, it is possible that users are indeed uncomfortable with perceived invasions of privacy, but that the paradox exists only as an artifact of measuring those sentiments in a vacuum: perhaps users are actually distinctly aware that their data has value—but so do the tools and platforms they use—and the trade-off between utility and privacy is simply a financial transaction like any other [55]. Similarly, perhaps users’ reported concerns accurately capture their true sentiments, i.e., that they would actively use more privacy-enhancing services and technologies – if only they were made aware of them or they were easier to use or cheaper. Or, perhaps the privacy paradox exists exactly as described, and users are simply fickle and their responses unreliable.

Whatever the case, what \textit{is} undeniable is that we need a clearer understanding of users: of the factors that motivate them and of the decisions that manifest as the behaviors that mystify us.

\subsection*{1.1 Understanding and enabling users}

\textit{Usable security} describes the research discipline focused on making the often-complicated technical aspects of computer security and privacy more accessible to normal users. This often describes efforts to improve the interface and system design of security and privacy technologies, and is


\textsuperscript{5}Serge Egelman, Twitter, 26 December 2018, \url{https://twitter.com/v0max/status/1078001421579153408}
analogous to more general usability and UX aims. However, it has also been applied to work of a more human-oriented nature which have focused on characterizing “the person behind the user” as opposed to directly improving users’ interactions with systems and services. For example, efforts to improve users’ ability to effectively use passwords includes both the development of password managers—tools which create, store, and input strong passwords on behalf of users—as well as studies into user notions of what constitutes a strong password and how those perceptions manifest in their personal password policies [85]. Whichever approach is used, however, the goal has often been the same: ensuring positive security and privacy outcomes for users.

In this work, we combine these approaches with a certain alternative philosophy in mind: enabling users to navigate the privacy vs. convenience trade-off as best suits their individual needs, values, and priorities. More specifically, we wish to promote users to actively and consciously make trade-offs between privacy and convenience in ways that align with their personal preferences, whether that be for increased privacy at the cost of decreased convenience or to consciously forego adopting more secure behaviors and tools if that is what is desired. Simply stated, we believe that “usable” security should not focus on “us[ing] design techniques to force users to keep paying attention and push them toward what we deem the secure—and hence better—option” [72]. Instead, we should direct our efforts toward assisting users in determining what is most appropriate for them and empowering them with the ability to personally realize those decisions, whatever they may be. It has, for example, been argued that the practice of computer security would benefit from incorporating Value Sensitive Design [18], an approach which prompts designers to consider the values of all stakeholders in a system, both direct and indirect.

Similarly, rather than operating under a binary “secure” vs. “insecure” paradigm for categorizing user behavior, wherein rejecting the “secure” option is often equated with making a mistake, we believe it valuable to instead:

1. Characterize the spectrum of user preferences and values with respect to the privacy of their online activity in specific contexts.
2. Communicate to users the trade-offs inherent in the possible decisions they may make with respect to adopting secure behaviors and technologies, and the cost of enacting those strategies in these contexts.

3. Design privacy-enhancing technologies in a way that pushes users to consciously and contextually exhibit a range of behaviors presenting different security vs. convenience trade-offs in a way that aligns with their personal values.

The privacy paradox presents a unique opportunity to demonstrate the viability and utility of such an approach.

More specifically, among the set of users whose behavior can be categorized as falling under the privacy paradox—i.e., despite expressing concern about their privacy, they fail to adopt privacy enhancing technologies and behaviors—there likely exists a spectrum of reasons for their behavior: there are likely users for whom this “discrepancy” exists because their preferences are inconsistent with the practical realities of the online ecosystem and users who fall under the paradox because they simply lack the knowledge and/or ability to protect themselves in spite of their intention to do so, and everything in between. Accordingly, in our view, “resolving” the privacy paradox is less about promoting the use of secure technologies and more about demystifying the range of risks and corresponding preventive strategies that surround online activity. In other words, we wish to bridge the gap between intention and behavior not by pushing users to privacy-preserving technologies that will guarantee the security and privacy outcomes they claim they desire, but rather by ensuring that the mapping between user-expressed sentiment and manifested outcomes is one that the user has chosen for themselves with full awareness of the options available to them, regardless of what that respective action may be.

It is to this end that we believe drawing upon the wealth of knowledge and experience in the analogous field of risk communication theory will be of great benefit.
1.2 Risk communication theory

Risk communication theory is a discipline with roots in public health and disaster awareness, whose focus is on establishing effective communication between experts and laymen on the risks and mitigating responses surrounding certain activities or events. [77] The World Health Organization describes risk communication⁶ as follows:

Risk communication refers to the exchange of real-time information, advice and opinions between experts and people facing threats to their health, economic or social well-being. The ultimate purpose of risk communication is to enable people at risk to take informed decisions to protect themselves and their loved ones. Risk communication uses many communications techniques ranging from media and social media communications, mass communications and community engagement. It requires a sound understanding of people’s perceptions, concerns and beliefs as well as their knowledge and practices. It also requires the early identification and management of rumours, misinformation and other challenges.

The literature includes a number of guiding principles to follow in attempting to maximize the effectiveness of risk communication efforts.

First, modern risk communication should be based upon scientific, methodologically-sound evaluation, particularly since intuition is often unreliable and subject to bias. While early risk communication efforts were “forged through decades of applied experience in response to infectious disease outbreaks, industrial accidents, and natural disasters, contemporary public health risk communication principles and practices build upon a foundation of behavioral, health, and social science theory and research.” [17]

Consequently, it is critical that communication about risk should be founded in empirical fact, acknowledge both knowns and unknowns, and most importantly, measured. Indeed, “without

---

monitoring and assessing outcomes on changes to knowledge, behaviour and practice, the activities related to risk communications become mechanical, meaningless, and do not help manage and control a public health emergency. Unmonitored for outcome, risk communications consumes and wastes valuable resources, are ineffective and create a false sense of achievement in those who are responsible for the response.” [29]

Thus, feedback from the intended audience is a critical component of achieving effective risk communication. Measured understanding of how risk communication messages are perceived and interpreted by their audience is necessary to adjust and improve their efficacy. Subsequently, effective risk communication should be the result of an ongoing and consistent dialogue about what is understood and what is misunderstood, what is perceived as important and unimportant, and what is believed and not believed.

Second, the goal of effective risk communication should be effective communication. It ought to encourage a dialogue between the sender and the receiver of the message, and ideal risk communication is “a two-way conversation in which an agency or organization informs, and is informed by, affected community members.”7 Accordingly, information should be conveyed simply and plainly, in language that is common to the intended audience. Similarly, it is important not to overwhelm individuals with too much information, particularly information that they lack the ability to process due to constraints on time or level of understanding; all the moreso in today’s world where people are inundated by a wealth of opinions that can be difficult to parse.

Third, risk communication should be provided in context, such that communication is “customized to meet the specific interests, concerns, and habits of the target audiences. [...] It’s essential to know your clients or your audience and understand who they are, what they care about, and what their personal situation is.”8 In other words, the “correct” action in a given scenario can

---

7Environmental Protection Agency, Risk Communication, https://www.epa.gov/risk/risk-communication
be dependent on the who, the what, the when, the where, as well as the why, as opposed to a single prescriptive response that should be unilaterally enacted. For example, past experience seeking insight into the mindset of individuals who have failed to evacuate before a flood has shown that this can occur even when individuals believe the warning given by officials, for example because they lack the practical ability to follow this order, due to lacking access to transportation and/or temporary lodging.

Thus, an in-depth understanding of the community that risk communication is intended for is absolutely vital. This includes factors such as language, risk perception, trusted sources of information, acceptable forms of communication, etc.\(^9\) Indeed, “\textit{when we move beyond generalizations—and dig a little deeper to discover more distinct audience traits—we can do a better job of reaching more people in ways that are meaningful to them so they can better receive important information.}”\(^10\)

This tailoring of messages refers not just to content, but even to mechanism: channels of communication must be carefully selected in order to ensure that the message is effectively conveyed to those at risk; some are more trusted than others, and some have more visibility than others. Similarly, the timing of messages is important. Again, an in-depth understanding of the target population is a prerequisite to crafting effective risk communication.

Finally, risk communication should communicate \textit{action}. The ultimate value of risk communication is in allowing individuals to appropriately respond to the risks they may encounter in their lives. Simply informing people of the risks they face—without offering recommendations for how to resolve those risks—is counterproductive and engenders feelings of hopelessness and impotence. On the other hand, the ability to take active action to ensure desirable outcomes can leave individuals feeling empowered and in control.

\(^9\)Centers for Disease Control and Prevention, Culture and Communication, https://emergency.cdc.gov/cerc/cerccorner/article_091716.asp

\(^10\)Centers for Disease Control and Prevention, Communicating with Diverse Audiences, https://emergency.cdc.gov/cerc/cerccorner/article_080516.asp
1.3 Applying risk communication principles to the adoption of encrypted communication

The potential benefits of applying of risk communication principles and approaches to information security have not gone unnoticed.

For example, Stewart and Lacey [80] caution against the “technocratic” approach to information security awareness: i.e., one that directly attempts to convey expert-level technical knowledge to users without allowing for cognitive and psychological constraints. In contrast, they present three concepts which they believe will heighten the efficacy of information security communication: bounded rationality, mental modeling, and the Extended Parallel Process Model. Similarly, Nurse et al. [56] conducted a review of the state of effective communication in information security, outlining several hurdles unique to the domain (with respect to traditional risk communication domains), as well as recommendations for improving the efficacy of current communication. Such recommendations include positive framing on messaging, using narratives to frame outcomes, taking caution when communicating risks numerically, and the need to evaluate risk messaging empirically.

1.3.1 Extending the literature

In this dissertation, we seek to extend existing work by evaluating the efficacy and applicability of risk communication principles and theory when applied to the question of the adoption of encrypted communication as a means of resolving privacy concerns expressed by users. This includes, in part, the application of risk messaging as an intervention, but more generally, speaks to our evaluation of the effectiveness of the risk communication approach as a whole when applied to this domain. This is accomplished in a four-part process where we apply various risk communication principles in order to:

1. Take a measured, scientific approach toward better understanding our target demographic.

   We profile perceptions of encryption technology as a whole, and of its role in daily life. We
further evaluate the sentiment of the risks surrounding email use, and of the perceived efficacy of two mitigating strategies: self-filtering and secure email with a statistical approach that gives a model for the factors influencing adoption.

2. Evaluate the applicability of risk communication to the domain of adoption of secure communication tools. We apply principles of risk communication to system design in two scenarios: messaging about the authentication ceremony in secure instant messaging and messaging about secure email. We find that the application of a risk communication approach leads to higher rates of secure feature adoption, with mixed results when it comes to the ability of risk messaging to directly improve comprehension. This leads to greater autonomy on the part of users, as they are able to personally and consciously weigh privacy trade-offs in context.

1.3.2 Chapter 1: understanding perceptions of encryption


The first step in risk communication is to establish what beliefs already exist regarding the risk and corresponding response in a given context. Because we are focusing specifically on applying risk communication theory to the domain of secure (encrypted) technologies, our first step is to profile pre-existing beliefs regarding encryption: what it is, how it works, and what role it plays in daily life. This gives insight into discrepancies between expert and user models of what it means for something to be encrypted—which facilitates communication via common language—as well as highlighting positive and negative perceptions of the utility of encryption, which can have implications for how users weigh its value when balancing risk trade-offs.
To this end, we conducted 19 semi-structured interviews with participants geographically distributed across the United States regarding their beliefs about encryption in an effort to profile 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 diagramming exercise that forced participants to exercise their mental models. The second half of the interview focused on how encryption might be used, and presented participants with three distinct encryption use cases for discussion: mobile device encryption, HTTPS (web encryption), and encrypted instant messaging apps. This data was then processed in a conventional qualitative content analysis approach using thematic analysis and a data matrix to define commonalities and distill discrete mental models.

We found that models of how encryption works essentially reduce to an access control abstraction with a shared secret acting as the encryption mechanism. This implies that communicating about encryption should focus on describing its properties as an access control mechanism and any interaction with cryptographic primitives should in some way align with the notion of a shared secret. We also found that perceptions of the strength of encryption varied greatly, that evaluations of its utility to a normal user were largely negative, that proactive attempts by users to learn more about encryption had only confused them due to poor risk communication approaches, and that careful effort must be made to understand how users interpret communication efforts, as we found that they had backfired in the case of the HTTPS indicator. These findings reaffirmed the importance of careful risk communication that is tailored to users’ beliefs as opposed to experts’ knowledge and also the utility of applying a risk communication approach to computer security. Mental modeling, a standard first step in a risk communication approach, was able to produce new insights.
1.3.3 Chapter 2: applying principles of risk communication to system design in the Signal secure messaging app

Justin Wu, Cyrus Gatrell, Devon Howard, Jake Tyler, Elham Vaziripour, Kent Seamons, Daniel Zappala. “‘Something isn’t secure, but I’m not sure how that translates into a problem’: Promoting autonomy by designing for understanding in Signal.” Symposium on Usable Security and Privacy, 2019.

The next step in exploring the applicability of risk communication theory to usable security is to apply it directly to system design and evaluate its effectiveness in improving user comprehension in an in-vivo environment. Accordingly, we employed principles from risk communication toward a redesign of the authentication ceremony and its messaging in Signal with a goal of promoting greater user autonomy through improving user understanding. This redesign process consisted of three phases: (1) an evaluation of the baseline effectiveness of Signal’s warning messaging, (2) the development and evaluation of improvement candidates, and (3) the implementation and evaluation of our new design.

We first conducted a cognitive walkthrough of Signal’s messaging regarding the authentication ceremony, establishing both what notifications appear to the user as well as identifying accompanying changes in internal state. This process revealed a number of potential obstacles to understanding, such as three distinct control flows that present different state and visual changes in response to the same key change event. To both verify the results of our cognitive walkthrough and identify missed pain points, we then executed a user study evaluating user reactions to these control flows (15 pairs each, for a total of 45 pairs) in a scenario where they were exposed to a simulated man-in-the-middle attack while performing a non-sensitive communication task.
We then designed significant changes to the messaging content using principles from risk communication, such as a focus on ensuring that users’ interaction options clearly described their consequences, making success/failure implications explicit, and conveying response costs upfront. We also designed multiple visual element changes—such as icons for a status indicator and candidates for a unified notification flow—and then had these evaluated by participants on Amazon Mechanical Turk. The “winning” candidates were then implemented into a final redesign.

Finally, we evaluated the effectiveness of our redesign with a third user study that mirrored the conditions of our first. Our redesign evidenced higher rates of participants performing the authentication ceremony and better comprehension of the authentication ceremony and its implications as compared to default Signal. Our findings showed that users do make active judgments about risk and can choose to use or avoid security measures depending on individualized, perceived risk combined with perceived response efficacy and cost. We also found that users rely on a host of mitigating strategies, including those that exist beyond the scope of any app or system, such as relying on alternative communication channels or apps.

1.3.4 Chapter 3: evaluating the applicability of protection motivation theory to the adoption of secure email and the relative influence of motivating factors in decisions to adopt

Justin Wu, Judy Ma, Daniel Zappala. “To encrypt or not to encrypt: exploring perceptions of secure email through the lens of protection motivation theory.” In submission.

Our next step is to evaluate the applicability of a core behavioral theory—protection motivation theory—to the adoption of secure communication (secure email, in this case). More specifically, we evaluate the applicability of protection motivation theory in explaining the differences between a
secure email-using and non-using population and frame the relative impact of several influential factors on users’ decision to either adopt or not adopt secure email. This approach: (a) sheds light on the nature of the privacy paradox with respect to secure email, (b) assesses the applicability of protection motivation theory to a new domain, and (c) focuses the direction for future risk communication efforts by highlighting user perceptions and sentiments with respect to the risks inherent in email use and the efficacy of secure email at mitigating those risks.

To this end, we executed a study where we devised and refined a survey instrument that captures sentiments regarding the risk/response model of secure email. We created a questionnaire of 5-point Likert-items where each item corresponded to one of the four component factors of the PMT model: risk likelihood, risk severity, response efficacy, and self-efficacy. This survey instrument was evaluated qualitatively by external content area experts and then administered to 149 users of a secure email system and 564 non-users of secure email systems. We analyzed the results quantitatively with a structural equation modeling approach.

Our results reveal that users of secure email view the trustworthiness of existing communication channels more negatively than non-users. Furthermore, this difference of view extends to all four PMT factors, validating the applicability of protection motivation theory to the problem of secure email adoption. Finally, our structural equation modeling analysis of these four factors reveals response efficacy the strongest validated predictor of adoption, and self-efficacy likely a similarly strong predictor, although we are unable to demonstrate causal effects with this factor.

Our findings suggest that the privacy paradox is indeed potentially a misnomer, with users’ inaction a failure of response assessment (response- and self-efficacy) rather than a misreporting of risk assessment. This provides direction for future risk communication efforts on this issue. In particular, we found that non-users of secure email were more likely to have negative views of the efficacy of this technology, believing it vulnerable to hackers and the government. Thus, effective communication about secure email will involve combating misconceptions and not just describing the potential benefits that can come from its use. Similarly, we found that confidence
in the ability to find and utilize a new secure email solution was potentially a large determining
difference between secure email-users and non-users, though this finding must be validated by
further research. This reinforces the value of ongoing efforts to improve the usability of secure
email tools, but also highlights the need to bolster users’ confidence in their ability to properly use
such tools, not just their actual ability to do so.

1.3.5 Chapter 4: studying perceptions of risk and response in the adoption of secure email
features

Justin Wu, Devon Howard, Ammon Mugimu, Calvin Woods, Daniel Zappala. “Exploring the
“Secure” in Secure Email: Perceptions of Risk and Response in the Adoption of Secure Email
Features.” In submission.

Previous work has indicated that factors outside of usability have impacted users’ decisions
regarding the adoption of secure email. For example, our own findings, from chapter three of this
work, suggest that self-efficacy and response efficacy failures contribute greatly to this non-adoption
phenomenon. However, these studies have evaluated user sentiment and behavioral intent, rather
than behavior itself.

Our final step involves building upon this foundation to evaluate how users perceive the risks
involved in email use and the utility of two secure email features when engaged in email scenarios
of varying sensitivity. We also evaluate the effect of a tutorial intervention on response efficacy
and behavior when presented prior to task performance. Thus, relative to our earlier Signal study,
we expand the scope of observation to now include multiple scenarios of varying sensitivity (as
opposed to the single, low-sensitivity scenario of our Signal study) and multiple response options
(as opposed to the adoption or non-adoption of a single, prescribed behavior). Additionally, whereas
our Signal study focused on the effects of our authentication ceremony redesign and messaging, this study focuses on characterizing how users reason about the use of secure email features.

We thus conducted a 31-participant user study wherein participants roleplayed standard email tasks according to the described parameters using an email system we provided with two optional protective features: encrypted email and self-destructing messages. More specifically, participants were divided into two groups: a treatment group that received a tutorial-based, messaging intervention prior to task performance and a control group. Participants then roleplayed a sequence of four email scenarios of varying sensitivity presented in random order, with the option of enabling none of the features, either of the features, or both features for any given scenario. They then answered a questionnaire evaluating their response efficacy models for encrypted email and self-destructing emails relative to the perceived vulnerability of normal email. Finally, they participated in a short verbal interview about their perceptions of normal email, their actions during the study, and their reasoning for such.

We found that our participants’ threat models did not align well with current models of secure email—largely equated with end-to-end encrypted email—as they did not perceive their email provider as a threat, but were concerned about hackers with access to compromised credentials. We also found that data permanence, or the persistence of sensitive email content, was a concern for our participants, and a direct cause for a number of them to enable our provided self-destructing message feature.

Moreover, lending credence to our findings from part three of this dissertation, we observed self-efficacy failures from participants who were too intimidated by the notion of encryption to even try using the feature, which had an actual interaction model of a simple password-protected email. Relevantly, we also found participants had a poor understanding of how encryption thwarted would-be attackers, as they simply associated it with greater levels of security, which affirms the findings of part one of this dissertation: i.e., that for most, encryption distills down to an access-control mechanism.
On the positive side, our findings suggest that automated encryption implemented by trusted email providers is a viable option. One of our four scenarios utilized an automated encryption feature that went unnoticed by virtually all our participants, implying that such a model would have minimal impact on usability, while sidestepping grassroots adoption obstacles. While the most straightforward implementation of such an architecture would require trusting email providers with keys, we believe that this is not an issue considering that participants did not view their email providers as adversarial.

Finally, we found that our response-efficacy-focused intervention failed to produce comparably different responses between our treatment and control groups, although we did observe statistically-significantly higher levels of feature usage. Unfortunately, our data is not rich enough to explain this phenomenon; future work is needed to investigate this in greater detail.
Chapter 2

When is a tree really a truck? Exploring mental models of encryption

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.
2.1 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 [14, 66, 67, 76, 83, 95]. 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 [20, 65]. Automation is not always a perfect solution; even the best software at times encounters errors that require user interaction to proceed [93]. 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 [4] and of the authentication ceremony in secure messaging apps [73] 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 [89]. 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 [92]. 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 [2], 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 ???. 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.

2.2 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.

2.2.1 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.
2.2.2 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.

2.2.3 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.

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.

**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.

**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.

**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’ familiarity 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 [2], 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.

2.2.4 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.
Figure 2.2: 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.

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.

2.3 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.

2.3.1 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\textsuperscript{2}, who related encryption with “\textit{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 “\textit{security},” “\textit{safety},” “\textit{protect},” and “\textit{privacy}” in addition to specific contexts where they imagined encryption was used such as “\textit{access control},” “\textit{email},” and “\textit{passwords}.”

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 “[\textit{w}hen they’re big enough to show up in the normal news. I’m not putting much effort into looking for this].”

\subsection*{2.3.2 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. \textbf{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).

\textsuperscript{2}All names used are pseudonyms
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 2.2.

**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
Figure 2.3: Edward’s barrier diagram, which represents “the symbolism of locking.” The skull is “just blocking, something that’s just blocking the message.”
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 2.2, 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 2.3) 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.

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.”

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
Figure 2.4: Wally’s black box encryption. The data to be encrypted goes through a “layer of encryption” that runs “in the background.”

Figure 2.5: Eva’s watermark diagram.
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 2.4.

In another example, Figure 2.5 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 unrandomize the letters.” (A handful of others possessed a similar, if slightly different view: a key was a like translator that guided the reversal.)
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 2.6, 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. “[If 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.”

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.

For example, Figure 2.7 shows Franklin’s depiction of how textual encryption would work:

“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
Figure 2.6: Allen’s ciphered cloud. “The picture is split into different sections or elements and each element turns into a different element”
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.”
Shared secrets

One vital aspect of interaction with encryption software is an understanding of the requirements for decryption. 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.”

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.
2.3.3 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.

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.

**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.”

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.
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.”

2.4 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.

2.4.1 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 [83].

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 [87].

**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.

2.4.2 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.”

2.4.3 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 [91]: “without an understanding of threats, home computer users intentionally choose to ignore advice that they don’t believe will help them. Security education efforts should focus not only on recommending what actions to take, but also emphasize why those actions are necessary.”

### 2.4.4 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” [89]. Indeed, “the ‘broadcast of facts’ approach has been discredited by experts in safety risk communications” [80].

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.

### 2.4.5 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.
2.5 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.

2.6 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.

2.6.1 Encryption usability

Beginning with the seminal work, “Why Johnny Can’t Encrypt” [95], 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 [95] 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. [83] 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. [2] 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. [69] 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.
2.6.2 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 [13, 62, 80].

Volkamer and Renaud [89] 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 [91] 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. [47].) 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. [12] 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. [46] 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.).

2.7 Conclusion

In this dissertation, we explored the application of principles from risk communication theory to the problem of grassroots adoption of secure communication tools. First, we explored mental models of encryption. We found that users’ perception of encryption essentially reduces to an access control abstraction and that its’ utility to the average user was generally perceived negatively. Second, we attempted to apply risk communication to system design, and evaluated its effectiveness in communicating with users about the purpose of the authentication ceremony in the Signal
secure messaging application. Our work demonstrated positive effects on both comprehension and adherence, illustrating the potential benefits of this approach. Third, we tested the applicability of a core theoretical framework of risk communication theory—protection motivation theory—to the domain of secure email adoption, and then used this framework to evaluate the relative impact of its four constituent factors on predicting the adoption of secure email services. We found that all four components of protection motivation theory were statistically significant predictors of adoption, and that of these four, response and self-efficacy were the strongest predictors. Finally, we evaluated perceptions of risk and response of secure email features when presented alongside email tasks of varying sensitivity. We found that end-to-end encryption of email aligns poorly with user threat models for email use, that unpermissioned access of persistent email content by hackers was a major concern, that self-efficacy was indeed an obstacle with respect to encrypted email, that risk messaging does not necessarily produce higher rates of comprehension—although feature adoption did increase, and that automated encryption of emails is a viable option for users.

We now discuss the contributions of our work. We also include both recommendations for improvements as well as directions for future work.

2.8 Contributions

The research contributions of this dissertation are:

1. **First directed effort exploring mental models of encryption.** We conducted the first targeted study of mental models of encryption, identifying four models of encryption, which largely distill down to an access-model abstraction. We found that users’ perceptions of encryption align with symmetric encryption, which helps explain the difficulties users face when interacting with asymmetric encryption mechanisms. We further identified response efficacy issues with perceptions of encryption, such as stigmatized views of its utility in daily life and confusion about attackers’ ability to break encryption in a timely manner.
2. Identified obstacles to users’ understanding of the authentication ceremony in Signal (and other messengers using the same ceremony, such as WhatsApp). We performed a cognitive walkthrough to identify obstacles to understanding of the authentication ceremony and then verified the existence of these obstacles via a user study.

3. Designed, implemented, and evaluated comprehension-focused redesign of the authentication ceremony and surrounding messaging in Signal. We designed, implemented, and evaluated a redesign of the Signal authentication ceremony and associated messaging using risk communication principles in a way that improved both comprehension and adherence levels. Our evaluation further allowed us to collect evidence suggesting that perceptions of risk must be divided into likelihood and severity components, and that the presence of risk is not alone a sufficient condition to predict the adoption of secure behaviors. Relevantly, we also provide evidence that the set of protective responses users consider at their disposal extends beyond the set of actions allowed them by the system.

4. Evaluated the applicability of the protection motivation theory framework to the problem of secure email adoption. We designed and evaluated a questionnaire whose items were composed of Likert-statements corresponding to each of the four components of protection motivation theory. Using a structural equation modeling approach and responses to our questionnaire from both secure email-using and non-using populations, we verified that all four components of PMT are individually predictive of secure email adoption. We also found that self-efficacy and response efficacy were the strongest predictors, suggesting that the privacy paradox can be explained by response assessment failures as opposed to failures by users to accurately report privacy behavioral intentions.

5. Studied risk and response perceptions of two models of secure email as perceived during four email scenarios of varying sensitivity levels. We designed an email interface with two additional optional secure email features—encrypted emails and self-destructing messages—and a risk communication intervention delivered to a treatment group. Participants interacted with
these features as they roleplayed a set of four email scenarios. We found that risk communication is more difficult under the broader parameters in this study as compared to our Signal work; comprehension did not improve, although feature use did. We affirm our previous findings regarding models of encryption and the impact of self-efficacy with respect to secure email use. We also outline favorable conditions for an automated email ecosystem.

2.9 Recommendations

1. **Avoid mentioning encryption as much as possible to users; instead, focus on providing positive framing for the properties it provides.** We found that encryption confuses users; even those who favorably associate it with positive security outcomes do not necessarily have a model for how or why it achieves that purpose. On the contrary, we found that encryption carried negative connotations: participants viewed use of it in daily life as the domain of the paranoid or immoral, and as a feature, complicated and intimidating enough as to fall outside the realms of easy use. However, based on recommendations from risk communication literature and our own experience with redesigning the authentication ceremony in Signal, we believe there to be a positive effect by presenting users with a positive framing on the properties that encryption can give them.

2. **Automated email encryption—and perhaps automated key management in general—seems a promising path toward an ecosystem that is more secure than what we currently have without compromising usability aims.** Our work shows that users are most concerned by the threat of hackers compromising their data, as opposed to service-level entities such as service providers or the government. This suggests that systems which automate key management through a centralized key server—such as the model utilized by Protonmail or Signal—can be employed at greater scale to offer a level of protection to users beyond that which they currently enjoy, even if short of full end-to-end cryptographic guarantees. An automated approach would further bypass certain grassroots adoption obstacles, such as response or self-efficacy failures.
3. **Users need the ability to control the removal of their data.** Our participants and others, such as those interviewed by Ruoti et al. [69] and Delta.Chat developers ³, described concerns regarding data permanence. Relevant to this concern, in our final study, we found that participants were generally fond of the self-destructing message feature we provided, largely in part due to the extent to which it aligns with this data permanence concern. We further noted, as did Kang et al. [46], that users demonstrated a strong sentiment that their data was simply “out there” in the Internet, vulnerable and exposed. These sentiments, combined with low response efficacy perceptions, led participants to rely on external methods for protecting their data, such as self-filtering even when inconvenient.

4. **Utilize risk communication principles in system design—and elsewhere—to more effectively communicate with end users the risks they face online and the responses available to them.** Our work demonstrates the potential efficacy of applying even simple risk communication principles to system design and messaging, achieving higher rates of protective feature usage in both cases we applied it, and higher comprehension levels in one. In particular, our work highlights the critical need for more effective, empirically-tested communication about the range of responses that are available to users, and the efficacy of those responses.

2.10 Future work

While we have presented evidence that risk communication can, and should, be applied to the adoption of secure communication technologies, we also note the need for further work investigating this issue.

- **Study the effect of risk communication on adoption longitudinally.** Our studies were all conducted in one-time sessions, either in the lab or over the phone. Thus, our data did not provide us the ability to evaluate whether or not the effects we observed would persist over time.

³Delta.Chat needfinding report: https://delta.chat/assets/blog/dcneedfindingreport.pdf
It is likely that, as with other things, such messaging would need to be repeated often for it to truly be internalized.

- **Evaluate alternative decision-making/technology adoption models for applicability to the secure communication adoption problem.** In this dissertation, we evaluated the applicability of one model—protection motivation theory—to secure email adoption. There are, however, multiple models and extensions to protection motivation theory that should likewise be evaluated. Similarly, the predictiveness of these models should also be evaluated on other types of privacy-enhancing technologies.

- **Study risk and response perceptions on different populations, particularly those facing different sociopolitical climates.** The participants of our students were almost entirely drawn from a U.S. population, and thus their perceptions of risk are shaped by the sociopolitical climate of this region. It is likely that the threat model of populations from other regions would differ in some ways, perhaps drastic ways; for example, perhaps we might observe that trust in email providers would not extend to other populations. This has impact on both the style and content of risk messaging that must be delivered to such populations as well as the design and architecture of privacy-enhancing technologies for their use.

- **Design and evaluate alternative models for communicating security/privacy risk and response in system interfaces.** We have traditionally provided visual indicators that offer binary measures of safety and protective features that similarly operate under binary on/off conditions. However, due to the many variables involved in a truly accurate risk-response calculation, complicated by diversity in levels of user comprehension, it is not clear that users possess the ability to independently assess the response that most aligns with their values. Thus, rather than straightforwardly exposing users to technical manifestations of risk and response, we believe that it might be more effective to offer an alternative paradigm where users are instead probed for their threat model by the software, whereupon appropriate protective features are suggested and/or enabled by the system.
2.11 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.
Chapter 3

“Something isn’t secure, but I’m not sure how that translates into a problem”: Promoting autonomy by designing for understanding in Signal¹

Abstract

Security designs that presume enacting secure behaviors to be beneficial in all circumstances discount the impact of response cost on users’ lives and assume that all data is equally worth protecting. However, this has the effect of reducing user autonomy by diminishing the role personal values and priorities play in the decision-making process. In this study, we demonstrate an alternative approach that emphasizes users’ comprehension over compliance, with the goal of helping users to make more informed decisions regarding their own security. To this end, we conducted a three-phase redesign of the warning notifications surrounding the authentication ceremony in Signal. Our results show how improved comprehension can be achieved while still promoting favorable privacy outcomes among users. Our experience reaffirms existing arguments that users should be empowered to make personal trade-offs between perceived risk and response cost. We also find that system trust is a major factor in users’ interpretation of system determinations of risk, and that properly communicating risk requires an understanding of user perceptions of the larger security ecosystem in whole.

¹Justin Wu, Cyrus Gatrell, Devon Howard, Jake Tyler, Elham Vaziripour, Kent Seamons, Daniel Zappala. “‘Something isn’t secure, but I’m not sure how that translates into a problem’: Promoting autonomy by designing for understanding in Signal.” Symposium on Usable Security and Privacy, 2019.
3.1 Introduction

The primary goal of usable security and privacy is to empower users to keep themselves safe from threats to their security or privacy. Their ability to do so is reliant on an accurate assessment of the existence and severity of a given risk, the set of available responses, and the cost of enacting those responses. Ideally, users would like to take action only when a threat has been realized and the negative consequences of that threat are severe enough to outweigh the costs of enacting the mitigating measure. In practice, however, it is difficult for users to have a comprehensive view of the situation and thus make informed decisions. Typically developers of secure systems best understand the nature of risks users will encounter and design responses that will mitigate those risks, but it is difficult for them to communicate this knowledge to users who are ultimately responsible for weighing risk severity and response cost trade-offs.

Consequently, the design of many security mechanisms seeks to simplify the threat-mitigation equation by avoiding calculations of risk impact and response cost, either through automating security measures or by pushing users to unilaterally enact protective measures regardless of context. This approach, however, is not without drawbacks. It discounts the impact of response costs on users’ lives by presupposing that the execution of a protective behavior is always a favorable cost-benefit proposition. In reality, however, the “appetite and acceptability of a risk depends on [users’] priorities and values” [44]. Indeed, it has been argued that, “Security that routinely diverts the attention and disrupts the activities of users in pursuit of these goals is thus the antithesis of a user-centered approach” [72].

This approach and its drawbacks is evident in the current design of secure messaging applications. In a typical secure messaging application, an application server registers each user and stores their public key. When a user wishes to send a secure message to someone, the application transparently retrieves the public key of the recipient from the server and uses it to automatically encrypt messages. However, because the server could deceive the user, either willingly or because it has been coerced by a government or hacked by an attacker, communicating parties must verify
one another’s public keys in order to preserve the cryptographic guarantees offered by end-to-end-encryption. The method by which parties verify their public keys has been called the authentication ceremony, and typically involves scanning a contact’s QR code or making a phone call to manually compare key fingerprints.

The usability of the authentication ceremony in secure messaging applications has been studied in recent years, with the general conclusion that users are vulnerable to attacks, struggling to locate or perform the authentication ceremony without sufficient instruction [1, 73, 87]. The root cause of this difficulty is that the designers of these applications do not effectively communicate risks, responses, and costs to users. The automatic encryption “just works” when there is no attack, but the application does not give users enough help to judge risk and response trade-offs when an attack is possible. Prior work [88] applied opinionated design to the Signal authentication ceremony and showed that they could significantly decrease the time to find and perform the authentication ceremony, with strong adherence gains. However, this work assumed that all users should perform the ceremony for every conversation, when many users may not want to incur this cost due to low perceived risk or high response cost.

In this study, we demonstrate an alternative design approach that emphasizes users’ comprehension over compliance, with a goal of empowering users to make more informed decisions that align with their personal values. We employ a design philosophy that might be seen as partway between opinionated and non-opinionated design: our design pushes users to make decisions, but not any decision in particular.

To this end, we conduct a three-phase redesign of the warning notifications surrounding the authentication ceremony in the Signal secure messaging app. We use Signal because the Signal protocol has been the foundation upon which other secure messaging applications have built, and thus many secure messaging applications share its basic design features and have similar authentication ceremonies. Because Signal is open source, we can apply design changes and, if these
changes are successful, influence applications based on Signal, such as WhatsApp and Facebook Messenger.

The authentication ceremony in Signal is a particularly good fit for applying a risk communication approach to design. First, the system has an explicit and timely heuristic for identifying shifts in risk levels: encryption key changes. Moreover, because changes in security state are contingent upon key changes, we need only communicate with users once a potential risk occurs. Furthermore, the available mitigating response to a key change is unambiguous: performing the authentication ceremony. Finally, the authentication ceremony is a mechanism where response cost factors heavily into the equation—users must be synchronously available to perform it—even though most key changes are due to reinstalling the application, not a man-in-the-middle attack.

Our redesign generally follows a standard user-centered design process, but with an explicit focus on enabling users to make more informed decisions. First, we measured the baseline effectiveness of Signal’s man-in-the-middle warning notifications with a cognitive walkthrough and a lab-based user study. Next, we designed a set of candidate improvements and evaluated their effectiveness by having participants on Amazon’s Mechanical Turk platform interact with and rate design mockups. Lastly, we implemented selected improvements into the Signal app and evaluated our redesign with a user study that repeated the conditions of the first study.

We make the following contributions:

- **Identify obstacles to user understanding of the authentication ceremony in Signal.** We performed a cognitive walkthrough of Signal’s authentication ceremony and associated notifications, highlighting barriers to understanding its purpose and implications. We followed up on our findings with a user study exposing participants to a simulated attack scenario, which allowed us to evaluate the effectiveness of these warnings in practice.

- **Perform a comprehension-focused redesign of the authentication ceremony with an aim at empowering users to balance risk-response trade-offs in a manner concordant with their personal priorities.** Building on the findings of our cognitive walkthrough and user study, we
redesigned the authentication ceremony and associated messaging with a focus on empowering users to make more informed decisions. Candidate designs were evaluated by users on Amazon Mechanical Turk with a final redesign evaluated in a user study. Our redesign results in higher rates of both comprehension and adherence as compared to Signal’s default design.

- **Show that risk communication empowers users to decide that not enacting protective behaviors is the right choice for them.** We find evidence that making users aware of the presence of an active threat to their data privacy is insufficient to produce secure behaviors. Users instead weigh the perceived impact of negative outcomes against the cost of enacting the response. Because “worst-case harm and actual harm are not the same” [38], this balancing of trade-offs can weigh unfavorably against performing protective measures.

- **Show that users’ strategies for mitigating perceived threats are dependent on their perception of the larger security ecosystem as a whole.** Despite our redesign prompting a greater share of users to perform the authentication ceremony, and producing greater understanding of the purpose thereof, participants’ preferred strategies for mitigating the perceived interception risk did not change substantially. Instead, it is apparent that users have developed an array of protective behaviors they rely upon to ensure positive security and privacy outcomes that exist beyond the ecosystem of any given app or system.

**Artifacts:** A companion website at [https://signal.internet.byu.edu](https://signal.internet.byu.edu) provides study materials, source code, and anonymized data.

### 3.2 Related work

#### 3.2.1 Protection motivation theory

We base our work on protection motivation theory (PMT), which tries to explain the cognitive process that humans use to change their behavior when faced with a threat [51, 64]. The theory posits that humans assess the likelihood and severity of a potential threat, appraise the efficacy and
cost of a proposed action that can counter the threat, and consider their own efficacy in being able to carry out that action.

Recently, PMT has been applied to a variety of security behaviors. Much of the work in this area is limited to studying the intention of individuals to adopt security practices, such as the intention to install or update antivirus software, a firewall, or use strong passwords [49, 97]. However, psychological research has demonstrated there is a gap between intention and behavior [74, 75], similar to the gap reported between self-reported security behaviors and practice [94]. A few studies have used objective measures of security behavior to study connections to PMT, such as compliance with corporate security policies [97], adoption of home wireless security [96], and secure navigation of an e-commerce website [86].

3.2.2 Risk communication

We are interested in studying how application design can be modified to help users assess risk and thus make more informed choices. We thus draw upon the wide variety of work in usable security that has focused on the design of warnings given to users.

Microsoft developed the NEAT guidelines for security warnings [61], emphasizing that warnings should only be used when absolutely necessary, should explain the decision the user needs to make, should be actionable, and should be tested before being deployed. Browser security warnings, in particular, have had a long history of lessons learned, including eliminating warnings in benign situations [81], removing confusing terms [10], and following the NEAT guidelines [26]. Phishing warnings are recommended to interrupt the primary task and provide clear choices [22]. Other work has recommended that software present security behaviors as a gain and use a positive affect to avoid undue anxiety [32].

We also draw upon risk communication, a discipline focused on meeting the need of governments to communicate with citizens regarding public health and safety concerns [15]. Nurse et al. provide a summary of how risk communication can be applied to online security risks [56].
Their recommendations include focusing on reducing the cognitive effort by individuals, presenting clear and consistent directions for action, and presenting messages as close as possible to the risk situation or attack. One noteworthy effort used a risk communication framework to redesign warnings for firewall software [60]. Their results show that the warnings improved comprehension and better communicated risk and consequences. However, the focus of this study, as with many others, was on greater compliance with recommended safe behaviors.

In contrast, we feel that risk communication provides a greater benefit in usable security when it enables users to make rational decisions based on their values, as opposed to compliance with a prescriptive behavior that experts believe is correct. For example, Herley has emphasized the rationality of users’ rejection of security advice, by explaining that users understand risks better than security experts, that worst-case harm is not the same as actual harm, and that user effort is not free [38]. Sasse has likewise warned against scaring or bullying people into doing the “right” thing [72]. Indeed, recent work on what motivates users to follow (or not follow) computer security advice indicates that differences in behavior stem from differences in perceptions of risk, benefits, and costs [25].

As stated by the National Academies, “citizens are well informed with regard to personal choices if they have enough understanding to identify those courses of action in their personal lives that provide the greatest protection for what they value at the least cost in terms of those values” [15]. Success is measured in terms of the information available to decision makers, and need not result in consensus or uniform behavior due to differences in what individuals value or perceive in terms of risks or costs of action.

3.3 Evaluating warnings in Signal

Signal uses the phrase safety number to describe a numeric representation of the key fingerprints for each participant in a conversation, warning users when this safety number changes. A safety number change occurs either when someone reinstalls the app (which generates new keys), or if a
man-in-the-middle attack is conducted, with an attacker substituting their own key for an existing one. The authentication ceremony in Signal is referred to as verifying safety numbers; matching safety numbers rules out an attack. To evaluate the effectiveness of notifications that Signal currently uses we conducted both a cognitive walkthrough and a lab user study.

3.3.1 Cognitive walkthrough

We performed a cognitive walkthrough of the notifications presented to users when a key change occurs and the authentication ceremony. The walkthrough was conducted by four of the authors, with a range of experience—a professor and a graduate student with substantial prior HCI and Signal research and two undergraduate students with no prior experience with HCI or with Signal. Our walkthrough consisted of exposing the user to every possible scenario leading to a safety number change, documenting all notifications and messages that are presented to the user and mapping the flow of decisions the user can make at each point. In addition, we analyzed Signal’s code base to establish how internal state accompanied each warning notification and the effects of user actions on these states.
Our cognitive walkthrough revealed that, depending on the internal state of the system prior to a key change, Signal will react in one of three different ways to a key change event, as depicted in Figure B.1 in Appendix B.1:

- **Message not delivered (top path in Figure B.1):** This path is activated when the user has not previously verified safety numbers, is still on the conversation screen, and attempts to send a message. Sent messages will show up in the conversation log, accompanied by a notification informing the user that they were “not delivered” and that they may tap for more details. Doing so brings up another screen which clarifies that there is a “new safety number” alongside a “view” button. Tapping the button generates a dialog (Figure 3.1a) with a succinct message about safety number changes and several options for proceeding, including one that leads to the authentication ceremony screen and one that clears the warning state.

- **Message delivered (bottom path in Figure B.1):** This path is activated when the user has not previously verified safety numbers and has either left the conversation screen or received a message. Signal will insert a notification into the conversation log informing the user of a safety number change, using a shield icon to mark the notification (Figure 3.1b). Tapping this dialog will take the user to the authentication ceremony screen. The shield and message appear in all three flows, but this is the only notification given to users in this flow; no other changes occur.

- **Message blocked (middle path in Figure B.1):** This path is activated when the user has previously verified safety numbers and has either left the conversation screen or received a message. This scenario places a blue banner at the top of the conversation log, warning users that their “safety number has changed and is no longer verified”. Tapping this banner takes users to the authentication ceremony. If the user attempts to send a message while in this state, Signal will prevent the message from being sent, and a dialog will be shown (Figure 3.1c). This dialog informs users that the safety number has changed and asks whether they wish to send the message or not. The user has three ways to clear the warning state in this scenario. They may select the
“send” option at the dialog, mark the contact as verified on the authentication ceremony screen, or tap the “x” on the blue banner.

Our cognitive walkthrough identified numerous issues that may be confusing and that contradict recommendations on effective warning design:

- **Unclear risk communication.** It may not be clear to users what the term “safety number” means, nor what it means that these have changed.

- **Inconsistency of choice across dialogs.** Although the message-not-delivered and message-blocked flows show dialogs that convey nearly identical messaging, they present users with different choices for interaction (Figures 3.1a and 3.1c respectively).

- **The consequences of user actions are not clear beforehand.** For example, in the message-not-delivered flow, it is likely for the user to send multiple messages that are blocked from delivery before noticing and attempting to resolve the error. If the user selects “Accept” at the ensuing dialog, this will automatically re-send all failed messages; not just the one selected for inspection. Conceivably, should one or more of those failed messages contain sensitive information, this might be undesirable behavior.

- **The implications of success or failure of the authentication are unclear.** In the event of a failed safety number match—the identification of which is the entire reason for the authentication ceremony—no recommendations for subsequent action are made to the user.

- **Does not communicate response cost.** The costs and requirements for performing the authentication ceremony are not made clear before users are brought to the authentication ceremony screen.

### 3.3.2 User study #1: Methodology

The following study, and all others in this work, were approved by our Institutional Review Board.
We designed a between-subjects user study to evaluate the effectiveness of each of these three notification flows at informing users of the potential risks they face and the responses available to them when exposed to a man-in-the-middle attack scenario. To control environmental conditions all participants used a Huawei Mate SE Android phone that we supplied.

For each of the three notification flows we discovered in our cognitive walkthrough, 15 pairs of participants (for a total of 45 pairs) conducted two simple conversation tasks. A simulated man-in-the-middle attack was triggered between the first and second tasks, causing the corresponding warning notifications to appear for each participant at the start of their second task. We simulated the attack by modifying the Signal source code to contact a server we operate and then change the encryption keys on demand. Participant reactions were recorded with video and a post-task questionnaire.

Our choice of tasks differs from previous work that asked participants to transmit sensitive information. Instead, we had participants communicate non-sensitive information, because this has the potential to reveal more diverse behaviors when faced with a risk of interception. For example, some users may be unconcerned by interception or unwilling to incur the cost of conducting the authentication ceremony if they perceive a conversation with non-sensitive information to be low risk. Others, on the other hand, may still find a potential attack to be unsettling and thus assess the risk to be more severe and/or the cost to be more worthwhile. A scenario with sensitive information could interfere with this dynamic.

We performed the studies for each treatment type—each notification flow—in succession, such that the first 15 pairs all experienced the message-not-delivered flow, the next 15 pairs saw only the message-delivered flow, and the final 15 pairs were exposed to the message-blocked flow.
Recruitment and Demographics

We recruited participants by posting flyers in buildings on our university campus. The flyer instructed participants to bring a partner to the study. Participants were each compensated $15, for a total of $30 per pair. Studies lasted approximately 40 minutes.

Our sample population skewed young, with 92.2% (n=83) of our participants aged between 18-24. Our population also skewed female (61.1%, n=55). A skills-based, self-reported assessment of technical familiarity revealed a normal distribution with most participants familiar with using technology.

Study design

When participants arrived, they were randomly assigned to an A or B roleplay condition (with a coin flip). Participants were then escorted to separate rooms, where they were presented with a packet of instructions, with one page per task.

Participants were first directed to register the Signal app pre-installed on the phones, granting all permissions the app sought in the process. Once both participants had finished registration, they were directed to begin their first task: to coordinate a lunch appointment using Signal. This task was designed to familiarize our participants with the operation of Signal. Exchanging messages is also necessary for Signal to establish safety numbers that could then be changed as part of the man-in-the-middle-scenario.

Next, participant B’s roleplay informed them that participant A had gone to Hawaii on vacation, and to hand their phone to their study coordinator to simulate this communication disconnect. Participant A’s roleplay provided similar information, including the instruction to hand their phone to their study coordinator, but additionally provided a half-page description of their “trip”.
Study coordinators took this opportunity to manipulate Signal into the conditions necessary for the associated treatment as well as triggering the simulated man-in-the-middle attack. Finally, phones were handed back to participants, and they were instructed to continue on to their final task.

Finally, participants were instructed to discuss and share photos of participant A’s trip to Hawaii, which had been preloaded onto participant A’s phone. With the simulated attack active, participants were now exposed to the warning notifications corresponding to their treatment group. These final instructions explicitly stated that participants were finished with this task whenever they believed they were, to avoid biasing participants toward any particular action in the event of a failed authentication ceremony.

Once both participants declared the task complete, they were given the post-task questionnaire. This questionnaire asked them if, within the context of their roleplay, they had perceived a risk to their privacy. They were then asked how they might mitigate this risk, and to describe how effective they believe their strategy would be. Finally, participants were shown each of the warning notification elements in turn, and asked: (1) whether or not they had seen them, (2) what message they believed the notification was attempting to convey, and (3) what effects they believed the associated interactive elements would produce.

Upon completion of the questionnaire, participants were read a short debrief, informing them that the attack had only been simulated, that Signal employs multiple features intended to both prevent and identify interception, and that no such attacks have ever been reported in the wild.

Data analysis

All open-ended questionnaire responses were coded by two of the authors in joint coding sessions using a conventional content analysis approach [40].
3.3.3 User study #1: results

Risk perception and mitigation

Roughly half of groups 1 and 3, the treatment groups whose messages either failed to send or were blocked, perceived a risk during the study scenario (13/30 and 16/30 participants respectively). In stark contrast, however, only a small fraction of the participants in group 2 (4/30), whose workflow was not interrupted, felt that they had encountered a risk. In explaining the nature and properties of the risk they perceived, participant responses generally fell in one of three categories: (1) a security risk of an unknown nature, (2) a risk of interception, or (3) a risk of an insecure communication channel. Perceptions of how to mitigate such a risk generally fell under one of three categories: self-filtering (avoiding communicating sensitive information), use of an alternative communication channel such as another app, and verifying a contact.

Shield message

The shield message in the conversation log, “Your safety number with \texttt{contact} has changed”, confused a number of participants. While many participants correctly associated this message with a change in security status, a number interpreted it to mean precisely the opposite of its actual meaning—that it conveyed improved security levels. As one participant explained following our post-study debrief, “I thought that it was improving security—that every once in a while, you change the safety number so it refreshes and makes it harder for people to hack into. So, I was like, ‘Oh, it’s doing its job.’ Apparently, it wasn’t!”

Next, as our cognitive walkthrough predicted, participants were confused by what, precisely, it was that had changed, offering numerous different explanations. Examples include: phone number, connection, safety number, safety code, “something technical”, settings, security code, and verification code. As one participant remarked, “Some sort of safety code changed. Or his actual phone number, I was a little confused.”
Participants acted on this message all cited the importance of ensuring privacy/security outcomes. Those who did not act on it did so because: (1) they did not see it as an actionable message, (2) they explicitly expressed having been habituated against such notifications, (3) the information they were communicating was seen as non-sensitive, or (4) they perceived it to be a part of the study task.

Notably, perceptions of the non-sensitivity of the conversation were critical in putting participants at ease even if they had found the notification alarming, as exemplified by one participant response: “I felt that it was important because of the nature of the app and whenever a safety anything is changed that usually is noteworthy. I would have put that it was extremely important if I had felt like there was an actual risk of someone actually trying to read our conversation.”

**Message-not-delivered dialog**

Only participants in treatment group 1 were exposed to the message-not-delivered dialog. Participants were asked to describe what they believed would happen if they were to tap the three interactive elements in this dialog: the “Accept” and “Cancel” buttons and the link embedded in the text.

Participants generally understood that “Cancel” would leave the system state unchanged. Similarly, most participants understood that “Accept” would unblock their messages and allow them to communicate once more. Perception of the link, however was more confused. 9 of the 14 participants who responded to this question responded that they had believed it would have taken them to a screen explaining more about the situation. This is in contrast to what it really does, which is to redirect users to the authentication ceremony, as noted by one participant who expected it to lead “to an ‘About’ or ‘Info’ page, but it ended up taking me to the verification.”
Blue banner & message-blocked dialog

Understanding of the options presented by the message-blocked dialog—“Send” and “Cancel”—were high. However, unlike the message-not-delivered dialog, the message-blocked dialog does not present a method to reach the authentication ceremony—instead accessible via the blue banner.

Understanding of the blue banner was mixed among those participants of group 3 who reported having seen it. Only roughly half understood that it was a privacy-related warning. Others were either entirely at a loss to explain its purpose or believed that it was a system error notification. Those who were confused by its meaning or believed it to be a system error did not feel it warranted action. Of the five participants who correctly interpreted the blue banner as a warning, two did not feel they were at risk, and thus did not feel like action was warranted.

Authentication ceremony

Participants who reported having seen the authentication ceremony screen were asked about the significance of verifying safety numbers (and whether or not they matched) as well as about the verification toggle. Participants may have seen, and even interacted with, the authentication ceremony screen without necessarily having performed the authentication ceremony. In total, 5 pairs of participants conducted the authentication ceremony while 27 participants reported having seen the screen.

As predicted in our cognitive walkthrough, participants were confused about what a safety number was or why it had changed. For instance, one participant explained that “I honestly wasn’t sure what it meant. I didn’t know that I had a safety number with them in the first place so I was unaware that it could change.” We also noted occasions where participants entered the authentication ceremony screen only to back out without completing it. This may be due to poor communication regarding response cost—both conversation partners should either be in the same
physical location to execute the QR-code ceremony or be willing to verify safety numbers over
another medium (such as a phone call).

Also as predicted, the verification toggle confused participants. Of the 11 participants who
reported having flipped the toggle, not one participant correctly intuited its use. 7 of these 11 toggled
it purely as an exploratory action, unaware that doing so would inadvertently and incorrectly clear
the warning state.

When asked to characterize the purpose of the authentication ceremony, participants did
generally associate it with verification, although their model for what it verifies was often incorrect.
Table 3.2 shows a qualitative analysis of participant responses when asked the purpose of the
ceremony, and the meaning of a matching or non-matching result, with responses coded and then
categorized as correct, partially correct, or incorrect. Only a few participants understood that the
purpose of the authentication ceremony is to verify the confidentiality of the conversation. Instead, a
number of participants mistakenly believed that it was about verifying the identity of the individual,
i.e., that “it makes sure the other person is who you think they are”, as one participant explained.
This threat model does not account for a different type of attacker the authentication ceremony
is intended to detect: a passive man-in-the-middle who simply decrypts and forwards messages
without interfering in the conversation.

These misconceptions naturally carried forward into responses about the significance of
matching and non-matching safety numbers. Perceptions of non-matching safety numbers correctly
assessed this result as indicative of interception occurring, but again, participants often believed
that this meant that they had detected an impersonator, as with one participant who remarked that,
“Someone using another phone could be posing as my brother, I guess.” Participants did almost
unilaterally understand that matching safety numbers were indicative of a positive security/privacy
outcome, although several participants misinterpreted the role of the authentication ceremony as a
mechanism that would actively prevent interception, as opposed to detecting it.
3.4 Developing improvements

Based on the results of our cognitive walkthrough and subsequent user study, we concluded that there were three main areas for improvement worthy of focus: (1) the need for an accessible, persistent visual indicator for verification state, (2) the messaging used in warning notifications and dialogs, and (3) the notification flow and all associated UI elements.

3.4.1 Visual indicator

Visual indicators, or icons, are important both as an accessible measure for communicating security state to users with a single glance as well as for enhancing the consistency of warning notifications. While the authentication ceremony screen in the original version of Signal does have a (somewhat hidden) lasting representation of verification state, the verified toggle switch, we believe that this indicator is inadequate because it represents only two states (verified and unverified) and because it confused users in our lab study who believed that toggling the switch would verify their partner.

We decided to create a set of icons that would properly reflect all three verification states: (1) the default, assumed-safe state of the conversation prior to a safety number change, (2) a verified state that reflects matching key fingerprints, and (3) an unsafe state that reflects having found non-matching fingerprints in the authentication ceremony. Ideally, the icon for the default state could have a small modification to represent the other two states. By adding this visual indicator onto the action bar, it becomes both an accessible indicator of state as well as a shortcut to the authentication ceremony.

We began by designing a neutral icon to represent the default state. Our goal was to select an icon that would be intuitively associated with privacy, and that would not evoke unwarranted feelings of concern, since this state does not signal a cause for concern. We selected a blank shield icon for this purpose. We then created variants of this icon, as shown in Table 3.1 to represent the success and failure states post-authentication ceremony.
Table 3.1: Comparison of icons a 10-point Likert scale

<table>
<thead>
<tr>
<th>Icon</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive – association with privacy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.50</td>
<td>2.21</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>6.54</td>
<td>2.82</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>5.74</td>
<td>2.56</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>7.52</td>
<td>2.43</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>Negative – association with worry</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.08</td>
<td>2.36</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>4.95</td>
<td>2.36</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>5.11</td>
<td>2.64</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>4.17</td>
<td>2.51</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>4.95</td>
<td>2.49</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>4.52</td>
<td>2.53</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>5.56</td>
<td>2.53</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>5.05</td>
<td>2.51</td>
<td>62</td>
<td></td>
</tr>
</tbody>
</table>

We evaluated our designs on Amazon’s Mechanical Turk platform, with each icon being shown to at least 50 participants. Each icon was shown occupying a position on the action bar in a screenshot of Signal’s interface, next to the call button. For positive-valenced icons we asked participants to rate how strongly they associated the icon with privacy on a scale from 1-10. For negative-valenced icons we asked participants to rate how worried they would feel if they saw the associated icon. We asked both questions for the blank shield icon.

As shown in Table 3.1, the blank shield has a moderate association with privacy and a low association with worry, making it a good fit for a default icon. We discounted any icons using a lock because it is used elsewhere in the app to represent encryption, and we wished to avoid conflating meanings. We chose the shield with a checkmark enclosed by a circle for the positive icon because of the remaining choices it had the strongest association with privacy. Surprisingly, no negative icon evoked strongly negative associations. We chose the shield with an exclamation mark because it had the strongest negative associations, and if the privacy check fails we do want users to be alarmed.
Appendix B.3 shows how these indicators are used in our design.

3.4.2 Notification and dialogs

We revised notifications and dialogs concerning safety numbers throughout Signal by following recommendations for warning design and risk communication. The principles we followed are (a) interrupt the primary task, (b) present messages close to the risk situation, (c) reduce cognitive effort (d) use a positive affect, (e), explain the decision the user needs to make, and (f) present clear and consistent directions for action. In particular, we designed the following changes, with screenshots shown in Appendix B.2:

• Positive framing for the authentication ceremony. We framed the authentication ceremony as a “privacy check”, which emphasizes the role it plays rather than the primitives or actions involved, which will be unfamiliar to users. Notifications of changed safety numbers (what we refer to as the “shield message”) instead report that Signal recommends a privacy check, turning what was sometimes seen as a routine system notification into an explicitly actionable recommendation. We also frame the consequences of performing the privacy check in a positive manner such that both positive and negative results present benefits for the user: a positive match guarantees conversation privacy and a failed match reveals ongoing message interception. In this way, even if fingerprints do end up matching—by far the most common case—users need not feel that it was a waste of their time to engage in the verification process.

• Communicating response cost and providing users with alternatives to the privacy check. Our dialogs inform participants up-front what executing the privacy check will require (Figure 3.2a), with a “Not now” option that generates a reminder dialog for participants who are uninterested (Figure 3.2b). The reminder dialog includes a recommendation to not communicate sensitive information until the privacy check is first completed, with an option be reminded at a later time and a description of how to access this functionality at any time. Thus, participants’ options are framed as clear choices with defined costs.
Figure 3.2: New notification dialogs, framing the authentication ceremony as a privacy check and using risk communication principles.

- **Safety number labeling and interaction changed.** To promote better understanding of the safety numbers and their role, we divided the safety numbers into their constituent halves, relabeled them as *device identifiers*, and explained that they are used in encrypting the conversation.\(^2\) Prior work indicated users dislike how long the safety number is [87]. Thus, we also rearranged groupings from 5 digits in a set to 3. This aligns more naturally with the standard process for grouping numbers, where numbers larger than 999 are grouped into sets of three known as periods. This does not reduce the actual count of numbers, but does reduce cognitive load.

We provide two options for performing the authentication ceremony, an in-person QR-code scan and a phone call comparison, as recommended in [88]. In the phone call version, we removed the confusing toggle element, which we replaced with two buttons explicitly labeled “Match” and “No match”.

- **Addition of success and failure messaging when the privacy check is completed.** We added dialogs after the privacy check that inform users of the implications of success and failure (Figures 3.3a and 3.3b respectively). We also designed a dialog shown before the authentication ceremony. If the privacy check has already been completed, it will show the current state and its implications for the user; if it has not, then it will instead explain what the privacy check entails and provide access to our authentication ceremony options.

\(^2\)This is not technically accurate, as they are key fingerprints and not keys themselves, but our goal is to have participants associate the comparison task with the preservation of a secure conversation and not to overwhelm them with details of the encryption process.
3.4.3 Notification flow

As described earlier, based on the system state prior to a key change, Signal diverges into one of three different notification flows. In order to provide a consistent user experience, we decided to instead use a single, unified flow every time a key change occurs. We eliminated the non-interrupting flow from consideration because in our first study it was ineffective at promoting either adherence or comprehension. This left us with the two interrupting flows, the message-not-delivered and message-blocked flows, which produced similar comprehension levels in our user study. We hypothesized there might be a difference between these because the timing of interruptions can have an impact on decision-making [3, 7, 52, 79]. We further identified two additional UI elements that might contribute to our aims of increased comprehension: (1) an introduction screen showing the privacy check icons after registration and (2) the blue banner element that accompanies the message-blocked flow in the original Signal.
To evaluate the relative effectiveness of these elements we designed a website containing a simulated Signal experience using mockups of our candidate flows and had users of Amazon’s Mechanical Turk platform interact with the simulation using a between-subjects comparison. Participants were, as in our user study, presented with two simple communication tasks involving non-sensitive information and a man-in-the-middle occurring between the first and second task. Unlike in our user study, users selected from a set of predefined messages, although they otherwise interacted with the interface normally, and their interactions were recorded in a database. For example, participants that wished to proceed with the privacy check were shown the results of the check as if they performed the authentication ceremony and asked to choose a response from the resulting dialog. After they were done, participants were given a tailored questionnaire which asked their perception of the notifications they had seen as well as why they had chosen the options they did. A total of 223 participants interacted with mockups and explained their actions via a post-task questionnaire.

We separated the elements to be evaluated into three rounds. The first round compared our delivery mechanisms: the message-not-delivered and message-blocked flows. The winner of the first round was then evaluated against a version that also included the blue banner element. Finally, the winner of the second round was then evaluated against a version that added an introductory screen.

To test for the difference between the message-not-delivered and the message-blocked flows we measured how many participants chose to start the authentication ceremony. We observed no significant difference (35/50 vs 31/50). We opted to use the message-blocked control flow because the message-not-delivered flow complicates the user’s task when they must resolve multiple failed messages. To test whether the blue banner message had an improvement we again measured how many participants chose to start the authentication ceremony and observed no difference (31/50).

To test whether the introductory screen had a difference we qualitatively measured comprehension. To do this we used participant responses to a question asking them what the privacy check
notification meant. Several authors coded each response and then determined whether the participant understood that this notification meant interception of their conversation could be happening and found a slight improvement with the introductory screen (30/50 vs 23/50). However, we chose to leave the introductory screen out of our final design because the effect was not large and could have been exaggerated due to the short-term nature of the simulation.

Qualitative analysis of participant responses regarding their decision to perform (or not perform) the privacy check showed participants weighed risks with response costs and made reasoned choices. Roughly 60% of all groups opted to perform the privacy check, with the remainder choosing the other option, “not now”. Participants who opted to perform the privacy check typically stated having done so out of a desire to verify the existence of the risk, because they believed it better to be safe than sorry, or out of curiosity. Participants who chose “not now” had either determined the risk to be of minimal severity or because they felt executing the privacy check would be inconvenient. Those who felt it would be inconvenient described it as such either because the current timing was seen as inappropriate or because of the synchronization cost (needing both members of the pair to execute the privacy check at the same time).

3.5 Evaluating the effectiveness of our redesign

We conducted a lab study to evaluate the effectiveness of our changes. Appendix B.2 shows the control flow we used and screenshots of the new notifications and UI elements, and Appendix B.3 shows the new indicators. We maintained the same study design used in our first lab study, with some minor modifications. Since our redesign has just one control path, we ran this study with just one treatment group of 15 pairs. Because we included additional screens that have no analogous equivalent in the original version of Signal, the post-task questionnaire in this study is not fully comparable with that from our first.
3.5.1 Results

Risk perception

Two-thirds (20/30) of our final user study participants reported having perceived a threat within the context of their roleplay. Qualitative responses indicated participants largely correctly perceived an interception risk, while a handful, interestingly, believed that Signal was itself the risk; this view seemed to be fueled by the number of permissions that Signal asks for in short succession. Participant notions of how they might mitigate perceived risks virtually mirrored those from the first user study—self-filtering, using alternate communication channels, and verifying contacts—along with restricting app permissions. Notably, only a small fraction of open responses (2/20) mentioned the privacy check as their mitigating strategy of choice despite, as we describe shortly, improved adherence and comprehension rates.

“Signal recommends a privacy check”

Qualitative responses indicate nearly all participants associated this notification with security, although as with our first study, there were a few who misinterpreted it as an increase in security. Due to our removal of Signal’s original messaging regarding a change, participants of our final user study were not confused about what had changed as the first groups had been.

Those who felt it important to act upon this message generally explained that they felt ensuring privacy outcomes to be important, as with one participant who explained, “I hear a lot about data breaches and such, so seeing that the app was giving a warning notification showed to me that it was something important that I should act on.” Importantly, those who did not feel that the notification was cause for concern typically felt that way because the information they were communicating was perceived as non-sensitive in nature.
**Privacy check dialog**

Qualitative responses indicate participants generally understood the dialog was informing them of a potential threat, although perceptions of the nature of that threat and of the likelihood of that threat were more varied. For example, while most participants correctly perceived that the dialog informs them only of a “potential” threat, a couple participants misinterpreted this notification as informing them of a confirmed threat, as with one participant who believed that “someone was hacking my account”.

Qualitative analysis of participant responses regarding their decision to perform (or not perform) the privacy check showed participants weighed risks with response costs and made reasoned choices. These results roughly match those of the Mechanical Turk participants who evaluated our candidate designs. Participants who felt performing the privacy check was important reported that this stemmed out of a desire to confirm the validity of the reported risk or because they believed it better to be safe than sorry. Those who did not feel the privacy check worth doing, on the other hand, had either deemed the risk minimal or decided that conducting the privacy check was too inconvenient.

**Privacy check**

As with the authentication ceremony in the first user study, participants in our final user study may have seen and interacted with the privacy check screen without having conducted the privacy check itself. 17 of our 30 participants reported having seen the privacy check screen, with 3 participants unsure. 6 participant pairs fully performed the privacy check, while 3 participant pairs partially performed the check (one participant in each pair incorrectly informed their partner that they had already matched the identifiers and that they thus did not need to complete the full process). This is in contrast with the 5 pairs (out of 45) who performed the authentication ceremony in our first study.
Table 3.2: Comparison of participant understanding of the authentication ceremony using Signal and our redesign.

<table>
<thead>
<tr>
<th>Auth. cerem.</th>
<th>Signal</th>
<th>Redesign</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>Verifies security</td>
<td>Verifies security</td>
</tr>
<tr>
<td></td>
<td>(conversation)</td>
<td>(conversation)</td>
</tr>
<tr>
<td></td>
<td>Verifies device</td>
<td>Verifies device</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>4 (16%)</td>
<td>8 (50%)</td>
</tr>
<tr>
<td>Partially correct</td>
<td>Verifies person (not impersonator)</td>
<td>Verifies person (not impersonator)</td>
</tr>
<tr>
<td></td>
<td>Verifies security</td>
<td>Verifies security</td>
</tr>
<tr>
<td></td>
<td>(connection)</td>
<td>(connection)</td>
</tr>
<tr>
<td></td>
<td>Prevents interception</td>
<td>Prevents interception</td>
</tr>
<tr>
<td></td>
<td>(connection)</td>
<td>(connection)</td>
</tr>
<tr>
<td></td>
<td>Improved security</td>
<td>Improved security</td>
</tr>
<tr>
<td></td>
<td>(conversation)</td>
<td>(conversation)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>8 [56.3%]</td>
<td>9 [56.3%]</td>
</tr>
<tr>
<td>Incorrect</td>
<td>Verifies phone number</td>
<td>Verifies phone number</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Makes the contact trusted</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10 (40%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

Matching

<table>
<thead>
<tr>
<th>Matching</th>
<th>Signal</th>
<th>Redesign</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>Interception not possible</td>
<td>Interception not possible</td>
</tr>
<tr>
<td></td>
<td>Verifies device</td>
<td>Verifies device</td>
</tr>
<tr>
<td></td>
<td>Verifies security</td>
<td>Verifies security</td>
</tr>
<tr>
<td></td>
<td>(conversation)</td>
<td>(conversation)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>7 (26.9%)</td>
<td>9 (56.3%)</td>
</tr>
<tr>
<td>Partially correct</td>
<td>Verifies person (not impersonator)</td>
<td>Verifies person (not impersonator)</td>
</tr>
<tr>
<td></td>
<td>Improved security</td>
<td>Improved security</td>
</tr>
<tr>
<td></td>
<td>Prevents interception</td>
<td>Prevents interception</td>
</tr>
<tr>
<td></td>
<td>(connection)</td>
<td>(connection)</td>
</tr>
<tr>
<td></td>
<td>Prevents interception</td>
<td>Prevents interception</td>
</tr>
<tr>
<td></td>
<td>(connection)</td>
<td>(connection)</td>
</tr>
<tr>
<td></td>
<td>Verifies security</td>
<td>Verifies security</td>
</tr>
<tr>
<td></td>
<td>(conversation)</td>
<td>(conversation)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>17 (65.4%)</td>
<td>3 (18.9%)</td>
</tr>
<tr>
<td>Incorrect</td>
<td>Don’t know</td>
<td>Don’t know</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Confusion</td>
<td>Confusion</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2 (7.7%)</td>
<td>4 (25%)</td>
</tr>
</tbody>
</table>

Non-matching

<table>
<thead>
<tr>
<th>Non-matching</th>
<th>Signal</th>
<th>Redesign</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>Interception occurring</td>
<td>Interception occurring</td>
</tr>
<tr>
<td></td>
<td>(MITM)</td>
<td>(MITM)</td>
</tr>
<tr>
<td></td>
<td>Conversation not secure</td>
<td>Conversation not secure</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>7 (28%)</td>
<td>9 (56.3%)</td>
</tr>
<tr>
<td>Partially correct</td>
<td>Connection not secure</td>
<td>Connection not secure</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>11 (44%)</td>
<td>4 (25%)</td>
</tr>
<tr>
<td>Incorrect</td>
<td>Don’t know</td>
<td>Don’t know</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7 (28%)</td>
<td>3 (18.9%)</td>
</tr>
</tbody>
</table>

These three “successful” misunderstandings had the same root cause—our design was not robust against false positives. Our design pops up an informative success dialog when a user taps the “Match” button. Unfortunately, this confused these participants who had mistakenly tapped the “Match” button. More specifically, one participant assumed that the “Match” button would activate an automated mechanism that would perform the verification for them. When the success dialog popped up in response, this participant assumed that the result had been in response to this
“automated process”. The other mistaken participants accidentally tapped the “Match” button and were similarly misled by the resulting success dialog.

Table 3.2 shows a qualitative analysis of participant responses when asked the purpose of the privacy check, and the meaning of a matching or non-matching result, with responses coded and then categorized as correct, partially correct, and incorrect. This table reveals that comprehension of the purpose of the authentication ceremony and of the significance of matching and non-matching numbers visibly improved with our redesign. While far from perfect, these results are promising given the context: a non-sensitive task scenario, no accompanying instruction or tutorials, and no incentive. Risk communication was limited to the messages contained within the application.

For all categories and for both user studies, partially correct responses center on the same few misconceptions: believing the verification process itself to be an active prevention mechanism, believing the “connection” and not the conversation to be the entity to be secured, and believing that the verification process verified the contact’s identity, and not their device.

3.6 Discussion

3.6.1 Risk communication gives users the ability to make personal trade-offs between perceived risk and response cost.

Simply knowing that a negative outcome is likely to happen is not a sufficient reason to take action to prevent it: it must also be negative enough. As the participant quoted in the title of this work so eloquently stated, sometimes “something isn’t secure, but I’m not sure how that translates into a problem.” Indeed, this view was shared by a number of participants of our studies. We observed numerous instances where participants did not believe that conducting the authentication ceremony was a worthy use of their time, whether because they perceived their communications as non-sensitive and thus unworthy of protecting, or because they felt that performing the authentication ceremony would be too inconvenient. One shared response captures both these sentiments perfectly,
“If it was easy enough I would be happy to secure my conversation, but at the same time, how necessary is it?”

While lowering response cost seems a natural way forward, particularly with automation, the deeply personal way in which calculations of risk function are made suggests obstacles ahead. Perceptions of risk severity in common scenarios will differ from person to person as a function of personal priorities and values. To wit, while many participants viewed communicating about our toy scenario as inherently non-sensitive, some participants were nevertheless uncomfortable at the realization that interception was “occurring”. One such participant, commenting on the thought of an interceptor eavesdropping on their discussion of a fictitious Hawaii trip, remarked, “Even though it’s only about fish, that’s not really cool with me.” We thus observe differences in risk assessment from different individuals although both the type of information being communicated and the nature of the threat itself were identical in all cases.

For these reasons, it is our position that enabling users to truly make informed decisions requires properly communicating the nature and likelihood of the risk and the cost of recommended protective measures, and then giving them the freedom to determine that not actively protecting themselves is actually the decision most in line with their interests.

3.6.2 Users’ strategies for coping with online threats extend beyond the ecosystem of your app.

Although our redesign evidenced both higher rates of participants conducting the authentication ceremony as well as comprehension, participants’ responses of how they might mitigate the risks they had perceived did not change in any notable fashion. Despite both having been made aware of a protective measure (in the privacy check) and also having understood its purpose, participants ultimately did not find it a reliable measure for mitigating a perceived interception risk should they encounter a similar situation in the wild. Rather, participants mentioned self-filtering, restricting
app permissions, and using alternative apps or channels of communication as key strategies for dealing with the interception threat introduced in the study.

This appears to be due to varying ideas about the source of the risk; in-app mitigating measures can only be depended on to do so much. Because the privacy check and associated messaging only informed users that conversation confidentiality had been violated, but not how that interception had been accomplished, users completed the process of threat assessment with personal interpretations of the origin of the interception risk. Relevantly, if the source of the risk is perceived to be outside the scope of the app—or even the app itself—it seems imprudent to rely on mitigating strategies that fall within the domain of the app.

System trust, perhaps unsurprisingly, appears to play a key role in this calculation. One participant response as to how they might better protect themselves is particularly ironic—they would forego use of Signal and “use [a] secure messaging app like Facebook Messenger”. Facebook Messenger does not protect conversations with end-to-end encryption by default, unlike Signal. However, due to unfamiliarity with Signal, and trust in Facebook, this participant’s preferred strategy would be to move from a secure messaging platform to a less secure one.

Future work could examine whether additional risk communication regarding the source of the threat could lead to improved understanding of the efficacy of the privacy check. System designers should also consider that users choose, to varying extents, appropriate responses to perceived threats, and that these include viable methods above and beyond what the system itself offers.

### 3.7 Limitations

Our cognitive walkthrough was thorough but limited to the expertise of the authors who participated in it. We ameliorate this by having a variety of backgrounds among those who participated, but a walkthrough performed by other experts or novices may find different issues with Signal’s
notifications. Our Mechanical Turk studies of icons and notification flows are limited to a simulated experience and thus may not match what users would feel or choose when interacting directly with the application. Our lab studies were limited to a young, college student population and may not generalize to a larger or more diverse population. Our Mechanical Turk results from the simulation provide some evidence that the results of the second lab study generalize to a larger, more diverse population. It would be helpful to study populations with different risk-cost trade-offs, such as immigrants or dissidents, and to ascertain that risk communication translates well to other cultures and languages. Our lab studies are also limited because users may act differently due to the Hawthorne effect [78]. Several participants made comments indicating this limitation was present, such as “while it is very concerning to me that someone could be intercepting my conversation, I thought that it was just because it was in a study.” However, because the focus of our study was on comprehension as opposed to behavior, this effect may be less impactful in our study.

Aside from these more common issues, we also observed a bug in Signal’s phone call functionality. The first time a Signal user makes an outgoing phone call, the caller is unable to hear audio although the recipient can hear clearly. Participants in our study simply redialed their partner when this occurred, typically chalking the issue up to a spotty wireless connection. This error, however, was present in the user studies evaluating both the original version of Signal and our redesign, so if this bug did have an effect, it likely existed in both cases, and thus is unlikely to have caused discrepancies in our results.

3.8 Conclusion

In this paper, we present the results of our experience redesigning the risk communication surrounding Signal’s authentication ceremony for comprehension. Our three-part process reveals significant obstacles to understanding in Signal’s current design, and demonstrates the effectiveness of applying risk communication principles to system design. Our user studies, which deliberately employ a non-sensitive communication task, provide evidence that users’ decisions not to enact protective
behaviors are actually conscious, informed decisions that are the product of balancing response cost and risk assessment. We further find that users rely on a host of protective behaviors that exist beyond the scope of any particular app or system, and that, consequently, responses to perceived threats may similarly exist outside of system designers’ control.

3.9 Acknowledgments

The authors thank the reviewers and our shepherd for their helpful feedback. This material is based upon work supported by the National Science Foundation under Grants No. CNS-1528022 and CNS-1816929, and by the Department of Homeland Security (DHS) Science and Technology Directorate, Cyber Security Division (DHS S&T/CSD) via contract number HHSP233201600046C.
Chapter 4

To encrypt or not to encrypt: exploring perceptions of secure email through the lens of protection motivation theory

Abstract

The privacy paradox is a noted phenomenon characterizing an apparent inconsistency between users’ expressed privacy concerns and their inaction when it comes to protecting their data. In this work, we explore a subproblem of the privacy paradox – the adoption of secure email software – using protection motivation theory, which characterizes the enacting of protective behaviors as dependent upon four component factors. We designed a survey instrument comprised of Likert-items corresponding to each of these four component factors as they relate to secure email. This instrument was distributed to an online population (N=564) of users and non-users of secure email services, and their responses were evaluated using a structural equation modeling approach. We find that the four factors of the protection motivation theory model – risk likelihood, risk severity, response efficacy, and self-efficacy – are all statistically significant predictors of secure email adoption, although we caution that we are unable to demonstrate causality for self-efficacy due to constraints on our sample population. Of these, self-efficacy and response efficacy have the greatest effect, which suggests that low adoption rates for secure email are not because users’ stated concerns about privacy and security are untruthful, but rather a consequence of a lack of confidence in both technology and, perhaps, their own ability to effectively mitigate those risks. This gives
credence previous efforts to improve the usability of secure email tools, highlights a critical need for better risk communication in conveying to users the effectiveness of secure email technologies to protect them, and potentially, the need to actively promote user confidence in using those tools.
4.1 Introduction

The so-called privacy paradox has been a subject of much discussion in recent years [8]. This phenomenon is characterized by an apparent inconsistency in how little users act to protect their data despite strong stated concerns about its privacy. The factors at play have been a topic of much debate, including assertions that the “paradox” of mismatched sentiment vs. behavior is actually a misnomer. That is to say, it has been suggested that this phenomenon occurs not because there is a contradiction between what users say and what they do, but rather because we have failed to capture a full picture of the factors at play in users’ decisions not to act. For example, one prominent usable security expert offered this commentary in response to a New York Times article on why people allow companies access to their data: “People might *want* to behave a certain way, but if they don’t believe they have control, often they will ultimately fail to act. This doesn’t mean their intentions weren’t true.”²

In this work, we examine one area where the privacy paradox appears to apply—the adoption of secure email software. The vast majority of the roughly 3.9 billion email users [35] do not use secure email software, despite standard email not offering strong privacy protections. However, there are some users who have chosen to use secure email services, with sites such as ProtonMail³ claiming 5 million users. This provides us an opportunity to understand differences in the perceptions of email risks and secure email utility between those who have adopted this technology and those who have not. In particular, secure email presents an opportunity to evaluate whether the gap between user sentiment and behavior is indeed a manifestation of feelings of resignation and lack of control, as has been suggested.

To frame the issue of secure email adoption, we draw upon protection motivation theory (PMT), a behavioral model that can help explain why people engage in practices that incur risk [51, 64]. The PMT model contains four factors—risk likelihood, risk severity, response efficacy, and

²Serge Egelman, Twitter, December 2018, https://twitter.com/v0max/status/1078001421579153408
³https://protonmail.com/
self-efficacy—and proposes that users will adopt a behavior if they perceive the risk is likely and
severe and that the protective behavior is both effective and one they can easily adopt. This model is
particularly appropriate for our purposes because it explicitly divides decision-making into risk and
response components, thus allowing us to independently assess users’ privacy concerns and beliefs
about their ability to control that privacy—and whether the latter is truly a factor in the privacy
paradox.

PMT was originally created to understand the effect of “fear appeals” for health attitudes
and behaviors. A fear appeal is a strategy that motivates individuals to action by exposing them to a
potential risk, such as injuries from auto accidents, and then offering recommendations that can
avoid the danger, such as wearing a seat belt. PMT has been used with some success in explaining
health-related intentions and behaviors, with self-efficacy generally having the strongest effect [27].

In contrast, individual adoption of secure email exists in a context where there is not a
prescribed, unilaterally “correct” decision and the cost-benefit trade-off is heavily dependent on
individual priorities and values. Thus, while wearing a seat belt is generally considered to be the
right decision for most everyone, adopting secure email is not necessarily the right choice for all
people in all situations, given their perception of risks and costs. Consequently, the scenario in
question is no longer about how to promote compliance, i.e., why those who fail to act are making a
“mistake,” but rather one in which users are faced with a non-binary decision where all actions may
be justifiable in specific contexts. This complicates how developers build secure email features and
communicate about them to users, since there is no clear policy prescription.

Applying PMT to the study of secure email gives us the potential to evaluate the applicability
of this model in a domain with different properties—secure email—compared to more traditional
domains of usage. Furthermore, should PMT’s component factors be shown to hold in this domain,
it will also serve to allow us to further explore the relative influence these factors hold in users’
decisions to adopt or not adopt this technology.
To study adoption of secure email in this context, we conducted a cross-sectional study of concurrent behavior to relate the dependent variable of secure email use with the variables of PMT constructs. We created a survey instrument composed of Likert-items corresponding to the four factors of the PMT model, as applied to secure email adoption. We distributed this instrument to an online population including both users who have already adopted secure email and those who have not and compared their responses. We conducted factor analysis to remove problematic items and finalize a model before performing a structural equation modeling (SEM) analysis to determine the plausibility of PMT factors as causal assumptions for predicting secure email use as well as to evaluate the relative role these four PMT factors play in differentiating secure email users against the general public.

Our analysis finds that all four components of the PMT model are statistically-significant predictors of secure email usage, as well as plausible causal factors for accounting for the adoption difference between the secure email-using and non-using populations, although we caution that we are unable to evaluate the causality assumption for the self-efficacy factor. This suggests that the PMT model is at least partially-applicable to this new domain, and likely wholly so. Our analysis of the relative significance of these four components finds that response assessment is the dominant differentiating factor, with self-efficacy showing the largest effect, followed by response efficacy. We thus offer evidence that, in this case, the privacy paradox does exist, but that its origins do not not lay in users lying about their privacy concerns; failure to adopt secure email services is perhaps instead because they lack confidence in technology’s ability to resolve those concerns and, potentially, also their personal ability to protect themselves effectively.

Because our secure-email using population are already users of secure email, we are unable to establish whether their confidence in their ability to execute predated their adoption of these services or exists because they have already successfully demonstrated such ability.
4.2 Related work

Our study seeks to examine the factors that influence secure email adoption in hopes of gaining insight into the larger Privacy Paradox problem. To this end, we utilize protection motivation theory (PMT) [51, 64] as a theoretical framework for profiling perceptions of secure email. More specifically, this theory seeks to characterize the cognitive process that individuals use to evaluate fear appeals, strategies that use fear to promote particular behaviors. Concordant with our purposes, protection motivation theory divides the binary decision of whether or not to adopt a given protective behavior into a set of four conditions that must be individually satisfied before people will act.

It posits that humans assess the likelihood and severity of a potential threat, appraise the effectiveness of a proposed action that can counter the threat, and consider their own efficacy in being able to effect that action. In deciding whether to take a proposed course of action, a person balances these factors, resulting in the person adopting either an action that can mitigate the threat or maladaptive behaviors that may be unproductive or even harmful. Importantly, PMT has conventionally been applied in personal health contexts, and has been used with some success in predicting health-related intentions and behaviors in areas such as smoking, alcohol consumption, and nutrition [27, 53]. This suggests that applying protection motivation theory to an information security context – such as secure email adoption – should grant valuable insight into the nature of the issue.

Indeed, the application of protection motivation theory has recently spread to information security domains. However, much of the work in this area is subject to a couple limitations. First, it is limited to studying the intention of individuals to adopt security practices, such as: the intention to install or update antivirus software, a firewall, or use strong passwords, the intention to adopt security actions recommended by a university, or the intention of teenagers to disclose personal information to websites [45, 49, 97, 101]. Previous research has, however, demonstrated gaps between intention and behavior, and particularly a gap between self-reported security behaviors and practice [74, 75, 94].
Secondly, many previous efforts have focused on the effectiveness of using fear appeals to promote compliance with institutional security policies in the workplace [37, 43, 45, 49, 58, 59]. Less prevalent have been efforts to study fear appeals and the adoption of secure behaviors in the wild, where users must self-diagnose and prescribe solutions to their data security concerns, although a few studies have attempted to tie objective measures of security behavior to a PMT-framing, such as with compliance with corporate security policies and adoption of home wireless security, and secure navigation of an e-commerce website [86, 96, 97].

In this study, however, we apply protection motivation theory to a domain – grassroots secure email adoption – where there is no organizational pressure to promote compliance, and where there is not a prescribed “correct” action that is unilaterally beneficial for users to adopt. Furthermore, we use this framework to study existing behavior, and not intention; our participants are differentiated by a binary dependent variable: they either already use secure email or they do not. Finally, the direct applicability of this theory to this new domain has yet to be evaluated, and so we make it a primary goal of this work to evaluate this.

To this end, we devise a survey instrument whose items correspond to the four components of the PMT model and seek to evaluate how responses differ between participants who have and have not adopted secure email on a personal basis. This approach has been previously employed in several instances. Notably, Egelman and Peer previously developed and validated a Security Behavior Intentions Scale (SeBIS) [21, 23]. This scale was later correlated against demographic features by Gratian et al [34]. Similarly, Hadlington evaluated the suitability of this approach in general [36]. Hadlington compared several scales measuring different constructs against actual behaviors, finding that several measures were indeed strong predictors of the associated behavior, reinforcing the validity of this approach. Finally, Thompson et al. used an extended PMT model to study predictors of various home computer and mobile device security behaviors [82]. They found that risk likelihood and self-efficacy affected both, risk severity only affected mobile security behaviors, and response-efficacy affected neither.
4.3 Methodology

Our IRB-approved study utilizes a cross-sectional design relating secure-email usage to the PMT factors. This involved the development of a new survey instrument designed to: (1) measure the four constructs of the protection motivation theory model as applied to secure email on two distinct populations – users of secure email and non-users (2) and evaluate how the differences in their responses correspond with the difference in their use or non-use of secure email. This instrument was validated both qualitatively, by content area experts, and quantitatively, with a structural equation modeling approach involving both exploratory and confirmatory factor analysis.

4.3.1 Instrument development

We first surveyed existing end user-facing online content discussing the benefits of secure email as well as privacy assessments such as Pew’s privacy survey in designing an initial survey instrument. Based on this survey, we created a questionnaire to measure the four components of protection motivation theory (PMT), i.e., risk likelihood, risk severity, response efficacy, and self-efficacy. We designed questionnaire items to elicit responses reflecting respondents’ perceptions of different aspects each of the components on a 5-point Likert scale.

We avoided the use of any technical jargon because we were aware that participants might not all understand what secure or encrypted email is, or different participants might interpret the phrase “secure email” differently. Instead, response assessment items described the properties granted by potential secure email mechanisms; for example, one item which describes expiring messages is phrased as follows: “We could better protect our privacy if email systems let us control when our email messages expired.” Additionally, to allow for greater expressivity of the spectrum of user sentiment, all categories include both positive and negative items with respect to the corresponding PMT category (agreement with the former indicating that the participant satisfies the conditions of the respective PMT factor, whereas agreement with the latter indicates the opposite).
While risk assessment items apply generally to privacy risks encountered with email, response assessment is more specific, relating directly to a particular coping strategy; there may be multiple coping strategies corresponding to the same risks. Accordingly, we thus designed two separate response assessment blocks, which covered both response efficacy and self-efficacy items for two potential defensive strategies: self-filtering and secure email.

We presented this preliminary instrument to three external content area experts and we instituted recommended changes. After content-area review, our instrument now comprised a total

Table 4.1: List of questionnaire items given participants. Each item corresponds to one of the four PMT categories. The final column indicates which items were included in the final SEM analysis; items excluded from final analysis have the reasons for their removal described.
of 7 risk likelihood items, 6 risk severity items, 9 response efficacy items (4 for self-filtering, 5 for secure email), and 8 self-efficacy items (4 for self-filtering and 4 for secure email). A final item was added to profile our binary dependent variable: whether or not respondents currently use secure email services, and if so, which ones. A complete list of these items can be found in Table 4.1.

4.3.2 Study sample

With our instrument prepared, we began recruitment for two distinct populations differentiated by our dependent variable: whether or not they already use secure email. We specified that secure email use must be with a personal account to avoid users who adopted secure email simply for compliance reasons at work. This recruitment was carried out on two platforms: Amazon’s Mechanical Turk and a Reddit subreddit devoted to privacy tools (r/privacytoolsIO), with the latter specifically intended to recruit a secure email-using population. In total, we recruited 513 participants from Mechanical Turk, of which 44 were users of secure email. We recruited 115 participants from Reddit, of which 10 were not users of secure email. From this data, we merged participants into two new groupings: users of secure email (N = 149) and non-users (N = 564).

Mechanical Turk users were compensated $1 for their participation, while Reddit participants contributed purely on a voluntary basis.

4.4 Questionnaire results

We report the arithmetic mean and 95% $t$-distribution confidence interval for each questionnaire item. We also report Cohen’s $d$ effect sizes for comparing the differences between the responses from the secure email-using and non-using groups.
4.4.1 Risk likelihood

Risk likelihood is the first of two risk assessment factors in the protection motivation theory model. It assesses the extent to which individuals believe that risks will or will not actually happen to them. Thus, even if an event is perceived as having extremely negative impact, if the perceived likelihood is minimal, people will not act to mitigate those risks.

Risk likelihood items are displayed in Figure 4.1. The first four items assess the likelihood of third-party access to respondents’ emails. Four potential third parties were described: email providers, advertisers, the government, and hackers. Because access by a government or a hacker may not be ongoing, we used a timeframe of within “the next several months.”
There are quite a few differences between users and non-users of secure email. Interestingly, two of the strongest differences are regarding their email provider and advertisers, with users of secure email leaning toward unlikely for those two third party types, whereas non-users lean in the opposite direction, toward likely. However, a caveat must be noted here. Due to a failure on our part to clarify, some respondents who use secure email services may have believed that items #1 and #2 referred to their secure email providers as opposed to generic email services, and thus responses to these two items are inconsistent. For this reason, we exclude items #1 and #2 from our later analysis.

Neither users of secure email nor those who use more mainstream services feel that either the government or hackers are likely to access their emails in this timeframe. But secure email users believe more strongly that it is unlikely that hackers will gain access to their emails, with a medium effect (Cohen’s $d = 0.65$). Secure email users are somewhat more worried that their email messages could be read by anyone (Cohen’s $d = 0.22$). Both groups agree that email is inherently insecure, with secure email users expressing this sentiment moderately more strongly (Cohen’s $d = 0.41$).

### 4.4.2 Risk severity

Risk severity is the second of the two risk assessment factors in PMT. This characterizes the extent to which potential risks, if realized, would negatively impact one’s life. It plays a critical role in risk calculations because people may choose to ignore those risks that they perceive as having minor negative impact, even if the perceived likelihood of those events is relatively high.

Risk severity items are shown in Figure 4.2. These items mostly ask participants to assess how concerned they would feel about email leakage across a handful of scenarios, with one negative item for participants who do not feel that actual harm would be realized by such scenarios.

There are again numerous differences between users and non-users of secure email systems. The population that uses secure email consistently rated the severity of the first four items higher than the non-user population, resulting in Cohen’s $d$ scores of 0.60, 0.52, 0.76, and 0.56 respectively.
Indeed, the secure email-using population averaged scores higher than 4.0 (out of 5) for the three scenarios involving advertisers and hackers. Also worthy of note are the ratings leaning toward “agree” for both demographics on item #12, which characterizes the email provider itself as a cause for concern. This suggests there might be perceived utility in end-to-end encryption as opposed to a more centralized system controlled by the provider. Finally, responses to item #13 revealed our non-user participants to be more likely to characterize a separation between discomfort and actual harm when it comes to unpermissioned message access (Cohen’s $d = 0.38$).

### 4.4.3 Response efficacy

Response efficacy is the first of the two response assessment factors in the PMT model. It describes the perceived ability or inability of a given protective behavior to mitigate a specific risk. If an individual does not perceive the utility of a given protective action to be high, then even if they
An effective strategy to safeguard our privacy is to never send any sensitive information when we use our personal email accounts.

Everybody would be safer if they avoided sending personal financial or medical information via email.

Keeping sensitive information out of our email messages altogether is not realistic. The responsibility should be on email providers to keep our email messages private.

People shouldn’t have to watch what they say when they send email. Meaningful conversations sometimes require sharing sensitive information.

Meaningful conversations sometimes require sharing sensitive information.

satisfy the risk assessment conditions of likelihood and severity, they are still unlikely to execute that action as a response to the risk.

Self-filtering

The self-filtering response efficacy items are shown in Figure 4.3. The first two items correspond to the effectiveness of self-filtering while the latter two characterize potential limitations of this strategy.

There are several areas where users and non-users of secure email differ in this category. First, our population that does not use secure email seems to consider the act of self-filtering to be quite effective, averaging over 4.0 on the first two items. The secure email-using demographic, by contrast, expresses similar sentiment, but to a lesser extent, and these differences have a small to medium effect (Cohen’s $d$ of 0.29 and 0.44). It is worth noting, however, that the latter two items
show fairly good agreement, indicating that even if self-filtering is considered an effective strategy, participants nevertheless believe that it cannot be an all-encompassing strategy for protecting the privacy of email. More specifically, the third item states that it is “not realistic” to keep sensitive information out of messages entirely, and that thus, email providers should enable privacy outcomes. Similarly, the fourth and final item specifies that there are occasions when the situation simply demands the sharing of sensitive information, precluding self-filtering as an option in certain scenarios.

**Secure email**

Previous work focusing on the privacy needs of chat messaging found that users placed great importance on message expiration/deletion, prioritizing it over encryption. We wished to evaluate this finding within the context of secure email, and thus item #22 corresponds to controlled
expiration, while item #23 describes the access control properties of encryption. Items #24-26 negatively characterize the effectiveness of secure email as failing to guarantee security outcomes.

Interestingly, item #23 generated substantially more agreement than item #22. Indeed, participants strongly agreed with item #23 in general, indicating that the properties encryption can potentially offer them are indeed something that users desire. Users of secure email systems agree even more strongly that emails should only be read by the recipient (Cohen’s $d = 0.42$).

Users and non-users of secure email disagree significantly on items #24-26. The non-user group appears to evidence some degree of hopelessness about the utility of technology to protect their data against would-be attackers. The secure email-using group, on the other hand, seems to place a certain degree of trust in the technology, particularly against hackers. This discrepancy in views can be seen in the Cohen’s $d$ values for items #24 and #25: 1.01 and 0.88, respectively.

### 4.4.4 Self-efficacy

Self-efficacy is the final component of the PMT model, and the second of the two response assessment factors. It characterizes an individual’s confidence in their personal ability to successfully execute a protective action. Thus, even if someone satisfies risk assessment conditions and believes that a particular action will mitigate said risk, if they do not believe they have the ability to adopt the corresponding protective behavior, they are still unlikely to do so.

**Self-filtering**

Figure 4.5 displays the items for self-efficacy of self-filtering. Agreement with the first two items indicates confidence in the ability to execute this strategy, while agreement with the latter two items indicates inability to utilize self-filtering as a strategy in all conditions.

These items generated the most neutral sentiment of the categories we presented participants and, with the exception of item #19, the sentiments of both survey populations mostly agreed. This
item describes the ability to control whether or not sensitive information is *received* as opposed to sent; secure email-users did not feel confident they could achieve this aim, whereas non-users did, albeit very weakly (Cohen’s $d = 0.52$). Secure email users also feel mildly more strongly that there would be situations where they must email sensitive information, despite the risk (Cohen’s $d = 0.22$).

**Secure email**

Figure 4.6 shows the items for self-efficacy of secure email. The first three items correspond to confidence in the ability to choose and use secure email services, while the last item describes a potential cost of using such a service.

This category generated very different responses between our two survey populations, with Cohen’s $d$ values of over 0.70 for three of the four items. Users of secure email systems feel very
strongly that they can choose a secure email system that protects their needs and learn its features, and they disagree that changing their email address is an impediment to using a new service. Non-users share this sentiment, but are less confident. These differences are perhaps unsurprising since the first two of these items characterize the ability to find and use a secure email service—something these participants have already successfully done, and thus know they are capable of. Users of secure email systems are somewhat more pessimistic about their ability to convince family and friends to use new privacy features in an email system (Cohen’s $d = 0.37$).
4.5 Analysis and modeling

4.5.1 Analysis – approach

We utilize a structural equation modeling approach to answer our primary research questions: (1) the applicability of protection motivation theory to a secure email adoption domain and (2) the relative weighting of the four PMT factors in predicting secure email adoption. Because this analysis was focused exclusively on the secure email scenario, we excluded the two blocks of self-filtering items from subsequent analysis.

Structural equation modeling is a modeling technique commonly employed in the social sciences to evaluate the relationship between variables when the system involves unobservable, latent constructs. This process involves the measurement of observed variables which are then used to impute the relationships between latent variables. For example, this approach has been used to investigate the relationship among components of second language motivation [16] and the relationship between media use and belief in global warming [39].

Interpreting the results of an SEM analysis can be confusing, particularly as this approach is explicitly intended to investigate causal relationships, and not just the correlation between dependent and independent variables, as with a regression analysis.

We stress that an SEM analysis does not demonstrate the presence, or lack thereof, of causal relationships; i.e., a significant coefficient does not demonstrate a causal relationship and a non-significant coefficient does not demonstrate the lack of one. Instead, “SEM is an inference engine that takes in two inputs, qualitative causal assumptions and empirical data, and derives two logical consequences of these inputs: quantitative causal conclusions and statistical measures of fit for the testable implications of the assumptions. [...] Fitting the data does not ‘prove’ the causal assumptions, but it makes them tentatively more plausible. Any such positive results need to be replicated and to withstand the criticisms of researchers who suggest other models for the same data” [11].
Consequently, an SEM analysis goes beyond regression analysis with respect to the degree in which it provides evidence about assumed causal relationships between variables, but falls short of demonstrating the existence of such a relationship, which must instead be validated by further study.

There are two primary reasons for our use of structural equation modeling in this study. First, its ability to evaluate the viability of causal assumptions as explanatory factors on a dependent variable, which matches our use case of assessing the suitability of PMT as an explanatory theory for predicting secure email adoption. Second, SEM was preferred to a regression model because of its ability to handle latent (unobserved) constructs. This is appropriate for answering the second research question of our study, wherein we have direct measures of sentiments which correlate with the four underlying PMT constructs and an explicit dependent variable, and we wish to link the former with the latter. Moreover, SEM accounts for the measurement error of individual items with respect to the larger constructs, providing some more granular insight.

4.5.2 Analysis – validating assumptions

Prior to beginning analysis, we first evaluated our data to assess whether it satisfied certain assumptions necessary for the following statistical tests. More specifically, we checked: coverage (whether all possible response values were observed for each item), multicollinearity (the extent to which independent variables can be predicted by one another, making them redundant), normality of data (whether data is normally distributed), and missing data percentage (whether per-item or per-respondent missing item percentages are sufficiently small).

First, data coverage. The set of responses for all items spanned the full spectrum of 1-5 on a 5-point Likert-scale, so we decided to analyze this data as continuous [63]. Next, a bivariate correlation table measuring the strength of the relationship between all pairs of independent variables did not reveal any extreme multicollinearity values (i.e., where correlation \( \geq 0.85 \)). However, items #12, 13, 22, 23, and 30 did not correlate with any others at all (correlation \( \leq 0.3 \)), so we removed
these items from subsequent analysis. (Because each item is intended to be one of a set measuring a PMT construct, items that do not correlate with any other do not capture the same construct and thus do not achieve the intended purpose.)

This removal left us with 7 risk likelihood questions, 4 risk severity questions, 3 response efficacy questions, and 3 self-efficacy questions. Next, we evaluated the normality assumption for our data. We randomly selected three pairs of items and conducted a fat pencil test for each. We performed curve estimation on each pair, comparing the first-order (linear) fit against the second-order (quadratic) fit; the normality assumption was shown not to be an issue. Finally, participant responses missed at most between 1-3 items per response. Because we observed data coverage and the missing data percentage per-respondent was low, maximum likelihood estimation was used for later analysis.

4.5.3 Secure email analysis – factorial structure

The first step in the SEM process is to characterize the relationship between instrument items and the latent constructs you desire to measure. Because our instrument was newly created for the purpose of this study and thus had not been previously validated, we began by conducting an exploratory factor analysis (EFA) [90]. EFA takes as input a number of factors (constructs) and the response data. Rather than impose a structure on the data, it allows all items to measure the full number of specified constructs and evaluates the strength of each item to measure the respective construct. Items in an EFA may thus measure multiple constructs, an undesirable phenomenon known as cross-loading.

Our EFA models revealed two items (#4 and 29) which cross-loaded across 4-, 5-, and 6-factor models, so we removed these problematic items in accordance with the guidelines outlined by Worthington and Whittaker [98]. Moreover, the first two risk likelihood items (#1 and 2) either consistently identified as their own construct or cross-loaded in the 4-factor model—likely a result of
the inconsistency in how participants interpreted these items, as mentioned earlier—so we similarly removed these two items from consideration.

We next conducted confirmatory factor analysis (CFA). This assesses whether the imposed factorial structure, constrained such that items measure only the associated construct, still fits the data. We evaluated potential models with three fit statistics: root mean square error of approximation (RMSEA), the comparative fit index (CFI [9]), and the Tucker-Lewis index (TLI [84]).

Achieving “good” fit at this stage is critical because it verifies that the instrument functions as intended. Hu and Bentler described an RMSEA of under 0.06 as a good fit [42], while Wang and Wang suggest a value of 0.08 as a cutoff threshold [90]. Hu and Bentler also characterized a CFI value of 0.95 or higher as a good fit [41, 42], while Wang and Wang deem CFI/TLI values of 0.90 and above as indicative of a good fit [90].

Based on the reported fit statistics, we eventually settled on the 4-factor SEM model whose items are listed in Table 4.1. This model produced fit values of: RMSEA = 0.068, CFI = 0.931, TLI = 0.909, indicating that our model fits the data well. Finally, correlations between these four factors ranged from -0.240 to 0.385, indicating that the four factors measure distinct constructs.

4.5.4 Analysis – SEM analysis

The final step of the SEM approach is to perform the SEM analysis itself. This is analogous to conducting a logistic regression, wherein a dichotomous variable (indicating whether a participant is a secure email user or not) is regressed against the four components of our model, although here it describes the plausibility with which these components are causal factors of the dependent variable.

The final SEM model, with unstandardized (non-normalized) coefficients, can be seen in Figure 4.7. Interpreting this diagram takes some care. The factor loadings shown on the arrows from each construct to a question show the relative weight of the question in measuring the associated construct. The coefficient shown between the latent constructs and the user indicate the influence of
Figure 4.7: The final SEM model path diagram with unstandardized coefficients. Leftmost values, leading into the items, are error terms representing the variance in each item not attributed to the associated constructs. Values sitting on the reverse arrows leading from the constructs to each item are the factor loadings; the first item for each construct is normalized to 1.0 and the magnitude of other items’ factor loadings are relative to this value. The final values leading from the constructs to “User” are the SEM coefficients. Numbers in parentheticals display the standard error for the corresponding value.
that construct on the user’s decision to adopt a secure email system. However, the weights shown are not standardized, meaning we cannot compare the relative weight of each PMT construct on the user’s choice. Instead, to answer that question, we convert the unstandardized coefficients into standardized coefficients.

Our SEM analysis revealed that all four components of the PMT model were statistically-significant predictors of the dependent variable of secure email use ($p < 0.001$ for risk likelihood, response efficacy, and self-efficacy and $p < 0.01$ for risk severity). The associated standardized regression coefficients were:

- Risk likelihood: 0.239, SE: 0.071
- Risk severity: 0.208, SE: 0.065
- Response efficacy: -0.472, SE: 0.059
- Self-efficacy: 0.599, SE: 0.051

Note that the response efficacy coefficient is negatively valenced because the response efficacy items remaining in our final model were “anti-”response efficacy items. Thus, agreement with these “anti-”response efficacy items negatively predicts secure email usage.

This result suggests that the PMT model is a plausible explanatory theory for predicting secure email adoption, although due to the nature of our secure email-using population, we again caution that we are unable to remark on the plausibility of the self-efficacy factor as a causal assumption. The higher a participant scored on any of the components (except the negatively valenced response-efficacy), the more likely they were to be users of secure email. As with regression, the relative weighting of these coefficients determines the extent to which they predict the dependent variable relative to one another. Accordingly, self-efficacy was thus the strongest correlate of secure email usage when contrasting users of secure email with non-users, with response efficacy the strongest plausible causal predictor.
4.5.5 Self-filtering analysis

We conducted a similar analysis utilizing the same approach as described above for the response of self-filtering.

Risk likelihood and severity items are not differentiated by response, and so the appropriate items for these two categories were used again in this model. For response efficacy, items #14-15 were used, and items #16-17 were excluded due to cross-loading. For self-efficacy, all four items (#18-21) were included in our model. This self-filtering model resulted in the following CFA fit statistics: RMSEA = 0.064, CFI = 0.930, TLI = 0.911.

Running SEM on this model suggests that two of the four PMT factors are viable causal assumptions: response severity and response efficacy. More specifically, the associated coefficients were: 0.406 (risk severity) and -0.220 (response efficacy). Because the included response efficacy items (#14-15) characterize the effectiveness of self-filtering as a strategy, its negative coefficient indicates that high agreement with these two items—i.e., high belief in the utility of self-filtering—negatively predicts secure email usage. This suggests that when participants perceive high data sensitivity and have strong belief in the effectiveness of self-filtering as a protective behavior, they are less likely to adopt secure email as a response.

4.6 Discussion

We conducted this study to answer two questions:

- Does protection motivation theory apply to the domain of secure email adoption?
- How much influence do the protection motivation theory factors carry with respect to predicting secure email adoption.

Our results indicate that the survey instrument we developed can be used to apply protection motivation theory to the domain of secure email adoption. Our four-factor model, with each factor
distinctly corresponding with one of the PMT constructs, fits the data well. This suggests that protection motivation theory itself is likely a good fit for explaining the factors involved in predicting secure email adoption, although given the limitations of our approach and our sample, future work is needed to both experimentally validate the model and evaluate the fit of alternative explanatory models. In particular, given the limitations of a study design that evaluates concurrent behavior such as ours to evaluate self-efficacy’s plausibility as a causal factor, self-efficacy-focused interventions should be designed and tested for their ability to promote secure email adoption.

Our results suggest a focus for directing future efforts where they will be most effective. More specifically, the strongest factors in our SEM model were response assessment factors (response efficacy and self-efficacy). Though risk assessment (likelihood and severity) also plays a role, this indicates that perhaps users fail to adopt secure email technology because of a lack of confidence that: secure email can protect them adequately (Items #24-26) or that they possess the ability to choose among competing systems (Item #27) or learn a new system (Item #28). Consequently, efforts to improve response efficacy and, likely, self-efficacy will have a larger bang-for-the-buck relative to efforts to increase awareness of risk likelihood and/or severity.

Regarding response efficacy specifically, users who don’t currently use secure email have a significantly lower belief in the efficacy of a secure email system, perhaps due to differing beliefs about the resources available to attackers (hackers may find a way around it, the government is too powerful). They also have high doubt that any secure email system would help them feel comfortable sending sensitive information. Subsequently, increasing response efficacy for these users may be difficult.

On one hand, there is some hope that with better publicity about the strength of encryption, feelings of powerlessness with respect to hackers and the government can be changed. Prior work on mental models of encryption indicates that most users do not understand how difficult it is even for a well-sourced government agency to break encryption [99]. Thus, perhaps a better understanding of its strength might lead to greater confidence in this technology.
On the other hand, however, even with strong encryption, there are still vulnerabilities deriving from implementation flaws [57]. Similarly, encryption can be bypassed with stolen credentials that give access to decrypted data. Thus, users are correct in their perception that no system can be perfect. For example, there is little ability to control information once it has left the domain of a user’s control—it is hard to ensure a recipient’s email has been deleted, and even then recipients can preserve the content in question with screenshots. One possible way forward is to help users understand that encryption puts up a significant hurdle for an attacker to overcome. While no system is likely to be perfect, perhaps they can be convinced an obstacle of this level is nevertheless sufficient to protect them from being personally victimized.

Regarding self-efficacy, the two items corresponding to this factor in our model characterize confidence in being able to choose an appropriate system and confidence in being able to learn new privacy-preserving features. Although our analysis finds this factor to be a strong correlate of secure email use, because our study design sampled users who have already accomplished adoption of secure email, we are consequently unable to remark on its applicability as a causal factor. However, given previous findings regarding the role of self-efficacy, it is likely that it is indeed a predictor. Should this be the case, several implications follow.

If the first item holds as a predictor, encouraging widespread adoption of encrypted email so that it is readily available in existing systems seems beneficial, as it would obviate users’ need to research and select new systems. If the second item holds as a predictor, refinement of existing research seeking to improve the usability of secure email will pay dividends, such as by automating encryption [70] so that even novice users can be successful. Indeed, since automating encryption reduces response cost, it may increase self-efficacy to the extent that it changes the cost-benefit trade-off for users. Consequently, easy-to-use encryption may make it possible for users to adopt secure email features even when risk likelihood or severity factors are low. It bears noting, however, that self-efficacy goes beyond just usability (a measure of actual ability) to include measures of user confidence. Thus, if self-efficacy is indeed validated as having causal effect on adoption,
the community will likely wish to direct effort toward understanding how to increase confidence, particularly among less tech-savvy populations and those under duress.

We believe this work argues for a holistic understanding of risk communication [15, 56] as it pertains to secure email. Risk communication can help users to understand the particular risks that are involved with sending sensitive information via email (including both likelihood and severity), explain to users what properties are offered by a given solution (such as encryption), and give them confidence that a given software system will be easy for them to use. This could enable users to make an informed and reasoned decision about whether they should adopt a new email system or choose to use a privacy-enhancing feature in their current email system.

4.7 Limitations

Our study is subject to several limitations. First, our secure email user sample was largely recruited by sampling from a specific location (a privacy-oriented subreddit) as opposed to subsampled from a larger population; only approximately 30% of our sample was obtained in the latter fashion. Similarly, a majority of our secure email user sample is composed of users of one secure email platform in particular: Protonmail, although there is some use of other systems such as Tutanota, Enigmail, GPG, and Posteo. Thus, they may not be representative of secure email users in general. Additionally, there may be cultural differences in our two populations that are not captured here: our Mechanical Turk population is constituted of U.S. residents, but our Reddit population had no regional restrictions on participation.

Furthermore, two statistical limitations arise from the small sample size of our secure email-using population. First, it is typical in the EFA/CFA process to build an EFA model with half the data and validate the CFA with the other half. However, due to our small sample size, building the models required our full data set. Secondly, with an instrument that has not been previously validated, as with ours, it is customary to evaluate the instrument for multi-group measurement invariance – the extent to which distinct populations use or interpret an instrument in the same
way. Unfortunately, once again, our sample was too small to make such an assessment, and so we assumed measurement invariance when performing our analysis.

4.8 Conclusion

In this paper, we present our findings drawn from an online survey of users of secure email and those who do not in an attempt to understand the differentiating factors that explain their respective adoption and non-adoption of secure email. To this end, we both utilize and evaluate the applicability of protection motivation theory in explaining this phenomenon. We devised a cross-sectional study relating concurrent behavior with PMT constructs. We created a survey instrument composed of Likert-items corresponding to each one of the four components of the protection motivation theory model: risk likelihood, risk severity, response efficacy, and self-efficacy. This instrument was evaluated using a structural equation modeling approach.

We find that protection motivation theory is a plausible model for predicting secure email adoption, with all four components showing statistically significant predictive ability, although we are unable to remark on the self-efficacy factor’s suitability as a causal assumption. Of these, self-efficacy and response efficacy demonstrated the greatest predictive ability, affirming hypotheses that the privacy paradox is not the consequence of misreported intentions, but rather captures users’ sense of futility with respect to protecting the privacy of their data. Improved risk communication regarding the efficacy of secure email technologies thus could have the greatest impact on adoption of privacy-enhancing email software. We end with a call for further work to validate these findings in both empirically validating the impact of the PMT factors as well as evaluating the applicability of alternative explanatory models.
Chapter 5

Exploring the “secure” in secure email: perceptions of risk and response in the adoption of secure email features

Abstract

Previous work has indicated that factors other than poor usability have impacted users’ decisions of whether or not to adopt secure email. These studies, however, have focused on email sentiment and behavioral intent, and have been isolated from contexts where users face decisions regarding actual use. In this study, we examine perceptions of the risks of email use and utility of two secure email features in a user study where participants interacted with two optional secure email mechanisms—encryption and self-destructing messages—in roleplaying scenarios of varying sensitivity. We explored participants’ interactions with, and reasoning about, these features. We find that current models of secure email—largely associated with encrypted email—do not align well with the threats perceived most vividly by our participants: hackers with access to compromised account credentials and data permanence. Similarly, some participants were so confused and intimidated by the thought of encryption to the extent of being unwilling to even attempt enabling the feature, despite the simplistic interaction model we provided of password-based access control. Nevertheless, we do find evidence suggesting a way forward for automated encryption, particularly since participants did not view their email providers as a threat, thus opening the door for provider-automated key management.

1Justin Wu, Devon Howard, Ammon Mugimu, Calvin Woods, Daniel Zappala. “Exploring the “secure” in secure email: perceptions of risk and response in the adoption of secure email features.” In submission.
5.1 Introduction

Users’ decisions to adopt or forego privacy enhancing technologies seem difficult to understand. The so-called privacy paradox, for example, characterizes a gap between user-expressed sentiment and exhibited behavior: users generally forego adopting more secure technologies despite stated concerns about their privacy. There have been disagreements about whether the apparent inconsistency in user behavior described by the privacy paradox truly exists, since the true problem may instead be that users lack knowledge of, or confidence in, current privacy-enhancing technologies, or may simply have no alternative. For example, many people may choose to use Facebook despite its poor privacy track record because they have no other privacy-preserving choice that provides similar access to neighborhood groups, local politicians, community events, and classified ads.

In our work we focus on the domain of privacy-preserving email features, often referred to collectively as secure email. This is an area of interest because the vast majority of people still use standard email accounts, which offer little in the way of privacy. Emails are sent in plaintext, meaning not just email providers but also various ISPs along the path of delivery can access email contents\(^2\). A number of options are currently available to users, such as PGP in some email clients and secure webmail providers like ProtonMail. However, PGP is generally considered to demonstrate prohibitively low usability [70] and adopting secure webmail requires users to switch email accounts.

Even if users switch to a secure webmail provider, where emails are automatically encrypted to recipients using the same provider, they must still decide on a per email basis whether to use password-based encryption when the recipient uses a different provider. Users must also decide whether to use automatic expiration for each email sent. Thus, even users who have already adopted secure email services are continually confronted with scenarios where they must consider

\(^2\)STARTTLS provides some confidentiality during email transfer, and DKIM and SPF provide some authentication and integrity, but adoption is low among the long tail of email providers and STARTTLS can be easily circumvented [19, 28].
potential risks, assess the effectiveness of an available privacy feature, and weigh the response cost in determining whether or not to enable particular secure email features.

Prior work on secure email indicates that factors other than usability may play a role in the adoption of privacy-preserving email features. One previous study, by Renaud et al. [62], outlined potential reasons for lack of adoption among both lay users and computer science students, finding support for four non-usability explanations for why users might fail to adopt encrypted email. An earlier survey of a larger group of lay users by Garfinkel et al. [31] found similar reasons, such as lack of understanding about how to use secure email, not being concerned about potential risks, or not feeling they would be a target. Similarly, prior work used a theoretical framework known as protection motivation theory to examine this issue, comparing the risk-response perceptions of secure email users vs. non-secure email users [100]. This work found that lack of confidence in secure email technology and potentially, users’ lack of confidence in their ability to employ those technologies effectively, play a critical role in differentiating these two groups, though they were unable to demonstrate causality in the latter case.

Notably, however, all of these studies involved assessing external measures of sentiment regarding email and were isolated from contexts involving actual use. Previous work has shown that self-reported security behaviors are reported to have, at best, a moderate correlation with actual behavior, provided the behavior involves choices that have visible effects (e.g. installing ad-block software, including a special character in a password) [94]). Since the factors surrounding the use of encrypted email have only been studied with respect to behavioral intent, there is an outstanding need to study how users make determinations regarding the use or non-use of secure email features in the context of actual email tasks.

Consequently, in this work, we build upon the foundation of these studies of sentiment and behavioral intent by studying how 38 participants perceived and interacted with two secure email features—encrypted emails and self-destructing messages—in the context of roleplayed email use. Participants were presented with a webmail system whose interface closely mimics Gmail, with
the exception of the two added features. They were asked to roleplay a set of four email scenarios exhibiting varying levels of sensitivity and given the option of enabling these features during each scenario if they so desired. Finally, they were queried about their perceptions and experience, with a focus on exploring the reasoning behind their actions.

Accordingly, our work is further differentiated from previous work that similarly profiles users’ decision-making processes due to the multi-response condition participants were presented with in our study. In prior work, a scenario typically involves a single, often-prescriptive response to a given risk, and users are faced with a binary decision of whether or not to enact the corresponding preventive measure. In contrast, in our study, participants were presented with four distinct options for action each time they sent an email: (1) doing nothing, (2) using only encryption, (3) using only a self-destructing message, and (4) using both features.

Our goal in conducting this study was to better understand how individuals reason about the risks of email and the use of secure email features. As a result, participants were explicitly informed beforehand that we were not expecting them to execute any particular behavior, but were rather interested in the reasoning behind their decisions. Because participants were exposed to these new features within the context of roleplayed email use, we were able to evaluate how their perceptions of risk and response affected their evidenced behavior.

Our findings, taken from analysis of participant actions during the study and post-task interviews regarding their experience, are as follows:

- **To users, “secure” means secure against hackers, not against email providers.** Our participants were largely unconcerned by the “threat” of their email provider or government having access to their data, and instead much more concerned by hackers gaining access to their accounts at some later date. This indicates a mismatch between the threat model users have and the model underlying secure email services that use end-to-end encryption.
End-to-end encryption protects email data in transit, but does not defend against attackers who have logged in to a user’s account with stolen credentials—a threat that our participants were explicitly wary of.

- **Data permanence concerns users.** We found that data permanence was a significant concern for our participants, with many participants enabling the self-destructing message feature. Participants frequently described how data, once sent over the Internet, simply remains “out there”. This reasoning had prompted many participants to avoid sending sensitive information over email in their past experience, preferring the loss of convenience to a risk to their information that could not later be rectified. Relevantly, we emphasize the need for the existence of protective responses that specifically resolve this concern.

- **Just the word encryption can be intimidating.** We also find evidence affirming one finding from previous work [100]—that self-efficacy is potentially an obstacle to the adoption of secure email—with a handful of participants explaining an unwillingness to even try our encrypted email feature because it seemed intimidating to them. This is despite the fact that the interaction model presented to users was simply that of password-protected emails; aside from describing the feature as encryption, no other encryption primitives or phrasing was used and participants neither saw encrypted text nor interacted with keys. These participants’ lack of confidence (low self-efficacy) was thus an obstacle to their ability to take advantage of this feature—they had already given up before even trying.

- **What encryption protects you against is poorly understood.** Participants had poor understanding of the threat encryption is meant to protect against, instead simply associating it with greater levels of security in general. In contrast, with self-destructing messages participants understood that this protected them from credentialed access by hackers at a later date or recipient indiscretion (bystanders gaining access to an email account left open on a computer).

- **Automated encryption by trusted providers opens new possibilities.** One of our four scenarios involved automated encryption: messages from participants to the study coordinator recipient
were automatically encrypted by the system. This occurrence went unnoticed by nearly all of our participants, despite visual indicators showing the change, suggesting a minimal usability impact. This, in combination with our finding that participants did not view their email providers as a threat, suggests that an automated encryption system where key management is transparently executed by email providers on users’ behalf is very much a viable path forward. Because an automated encryption system of this nature could be unilaterally employed by the email provider, other user-originating obstacles to adoption could be avoided. We outline several possibilities in this direction.

- **Habituation is a major obstacle that must be overcome.** The majority of our participants never enabled either secure email feature despite typically labeling at least one of the scenarios as sensitive and having been made aware from the beginning that the optional use of these features was a key component of the study. When asked to explain this discrepancy, participants often explained that habit had kicked in, and they had simply sent the task emails without additional consideration.

### 5.2 Related Work

#### 5.2.1 Adoption of encryption

Two studies have been conducted to understand why people don’t use encrypted email. The most in depth study was conducted by Renaud et al. [62], who conducted interviews with lay people and a survey of computer science students. They use an ad hoc model of 7 potential reasons why people might not to use encrypted email, finding strong support for four of them – lack of concern, misconceptions about threats, not perceiving a significant threat, and not knowing how to protect themselves. Lack of concern has a number of sources, including a feeling of not having anything to hide, not perceiving any harm if emails are compromised, not feeling important enough to be a target, and assuming the email provider is handling security problems. An earlier study of this
same issue was done by Garfinkel et al. [31], with a survey that asked users why they didn’t use encrypted email. Half of the 400+ respondents indicated they didn’t because they didn’t know how, while those who knew how indicated they didn’t think it was necessary, didn’t care, or thought the effort would be wasted. Overall, work in this area demonstrates that usability is not the only obstacle to adoption, and that (non-)perception of risk, lack of knowledge about effective responses, and poor self-efficacy all play a role in non-adoption of secure email.

Abu-Salma et al. conducted a more general study of why users don’t adopt secure communication tools (primarily secure messaging apps but with some emphasis on secure email as well) [2]. They found that sensitivity of information does not drive adoption and that usability is not the primary obstacle. Instead, there are a variety of other factors that affect adoption, such as having a large user base, interoperability, the context in which the tool is being used, the services offered, the quality of service, and cost. Participants used voice calls and obfuscation techniques to transfer sensitive information. They also had a perception that any secure communication tool could be breached by intelligence agencies, service providers, or skilled hackers.

Several papers have studied the use of secure email within specific organizations. Gaw et al. studied use of PGP within an organization of activists [33]. Users considered everyday use of encryption paranoid, but this was largely influenced by the system they used, which flagged all encrypted messages as urgent. They recommend that tailored solutions for users with different needs and recommend making security invisible. Lerner et al. designed a secure email system called Confidante that is based on PGP and uses Keybase to exchange public keys [50]. They tested its use with lawyers and journalists, finding different threat models and use cases. While journalists have a need to protect sources and thus even metadata (who is talking to whom) is important to keep private, lawyers may not be allowed to protect metadata due to legal requirements and especially need to rely on search, which can be hindered by end-to-end encryption. Thus they also find it is unlikely that a single solution will satisfy all types of users.
Our study is the first to study adoption of individual security and privacy features with ordinary users. Most other studies of secure email have examined one or several secure email systems from the perspective of usability. Typical studies ask users to perform some tasks and then see if they make any “mistakes” with respect to some assumed-correct choice [5, 30, 67, 68]. Stated in a risk communication framework, this work assumes that users will perceive a risk and perceive that the prescribed response is effective in mitigating that risk; failures to enact the corresponding response are deemed usability (self-efficacy) failures. Our work instead examines perceptions of risk, response efficacy, and self-efficacy as separate factors.

5.2.2 Risk communication

Our work studies adoption of secure email features from the perspective of risk communication theory; the application of this paradigm to security and privacy design was first advocated by Camp [13]. Risk communication originated as a discipline focused on meeting the need of governments to communicate with citizens regarding public health and safety concerns [71]. As stated by the National Academies, “citizens are well informed with regard to personal choices if they have enough understanding to identify those courses of action in their personal lives that provide the greatest protection for what they value at the least cost in terms of those values” [15]. Success is measured in terms of the information available to decision makers, and need not result in consensus or uniform behavior due to personal differences in what individuals value or perceive in terms of risks or costs of action.

Nurse et al. discuss how risk communication principles can be applied to online security risks [56], and they recommend reducing cognitive effort, presenting clear and consistent directions for action, and presenting messages as close as possible to the risk situation or attack. Raja et al. used a risk communication framework to redesign warnings for firewall software [60], showing that their warnings improved comprehension and better communicated risk and consequences. However,
the focus of this study, as with many others, was on effecting greater compliance with certain prescribed behaviors.

In contrast, we feel that risk communication theory provides a greater benefit in usable security when it enables users to make informed decisions that align with their values and priorities, as opposed to encouraging compliance with a prescriptive behavior that experts believe is correct. This is grounded in work by Herley that emphasizes the rationality of users in making security decisions [38]. He explains that users can be rational when they reject security advice, sometimes understanding risks better than security experts. He makes important observations that worst-case harm is not the same as actual harm and that user effort is not free. Sasse has likewise warned against scaring or bullying people into doing the “right” thing [72]. Fagan et al. used a survey of 290 participants to show evidence that users are rational when they choose to follow (or not follow) computer security advice based on perception of risks, benefits and costs [25].

5.3 System design

The purpose of this study was as outlined in the title of this work and as explained to our study participants: “to understand how people balance the benefits that come from using improved security/privacy mechanisms against the potential inconveniences they can cause,” with a focus on understanding “the reasoning behind [their] decision, whatever that may be.” Accordingly, we were interested in better understanding the experience of the average user when interacting with existing secure email features. More specifically, how do they decide when the adoption of a secure email feature is warranted? What risks do they believe they are intended to resolve? How do they perceive the utility of these features, what costs do they associate with their use, and how do they balance these properties?

To this end, we created a webmail interface for use in this study, specifically designed to tease out answers to our research questions with respect to two secure email features currently available to users in existing secure email systems—encrypted email and self-destructing messages.
We started with an exact clone of the Gmail interface, and then added these two features, as shown in Figure 5.1. This webmail interface provided core functionality for sending and receiving messages, including attachments and threading, and was hosted at a \texttt{2040mail.com}, which we chose to connote a generic email provider.

We chose to add these features to the Gmail interface, instead of having users interact with existing systems, because we wanted an interface that was familiar to most users. We also wanted the self-destructing messages feature to inform recipients when expiration will occur (ProtonMail does not do this) and to automatically delete messages from the sender’s account (Gmail does not do this). Our goal was to minimize the impact of unfamiliarity as much as possible to avoid pulling participants out of their comfort zone—we wanted them to be roleplaying real life as much as possible.

We emphasize that the purpose of our study was on understanding the user \textit{experience} of interacting with a secure email system, and not on the design and architecture of such a system. Accordingly, we presented our participants with a mockup whose interaction model is comparable with that of current systems, but whose functions were not implemented as presented at an
infrastructural level. Although we did not implement these two protective features to a degree that they could be deployed right now by a given email service, our experiment has ecological validity because we ensured that the user experience we provided our participants is what they would have were those features completely implemented, and what they get now when using existing systems like Protonmail and Gmail’s confidential mode. Specifically, the bodies of encrypted emails were transformed into unreadable characters in inbox view, password-encrypted emails (to recipients not using a 2040mail.com address) required a password to read, and self-destructing emails did indeed destroy themselves at the appointed time. For this reason, though they were interacting with a mock-up, and not a fully-operational system, how participants interacted with these features and how they reasoned about these features’ use is unchanged. In daily usage, users also have little to inform them about whether the systems they currently use really do what they say behind-the-scenes either. Thus the trust model users have with our system is the same as with a real-life system.

5.3.1 Secure email features

For our study, we chose to present participants with two secure email features for optional use: encrypted email and self-destructing messages. These two features were chosen because they are two features provided right now by existing secure email providers. Encrypted email is often equated with secure email, and the value of self-destructing messages, in addition to their presence in existing services, has basis in existing literature [24, 54, 69]. One of these studies, focusing specifically on the design and evaluation of a proposed self-destructing message utility using ephemeral keys, found that over 70% of their participants described wanting a tool that would “make [email] messages unreadable after a certain period of time” [54].

**Encrypted email**

The secure email interface we presented participants included two models of encrypted email: automated encryption to contacts using a 2040mail.com address, and an optional, password-
based encryption feature. This design is modeled after current secure email system designs used by services such as Protonmail and Tutanota.

Current encrypted email systems automate encryption between contacts using their service, while typically using a password-protected mechanism for out-of-provider contacts. Emails that are encrypted to contacts within the same provider are end-to-end encrypted using browser-generated keys that are bound to a password created by each user. Emails to contacts outside the provider are also end-to-end encrypted with user-created passwords, although these passwords are chosen by users at the time an email is composed, and not otherwise linked to their account.

Our interface presented this feature to participants in the compose screen. Encryption status was shown as an additional line in the compose screen, defaulting to “Encrypted” with a dropdown option of “No Encryption” if the recipient had an email address in the 2040mail.com domain. If the recipient address did not use this domain, the default option was instead “Non-encrypted” with a dropdown option of “Password-based encryption.” If password-based encryption was selected, an additional line was inserted below that allowed the participant to enter a password of their choice.

**Self-destructing messages**

The interaction model for self-destructing messages that we presented our participants was modeled after Gmail’s confidential email mode\(^3\). In this model, participants simply select an expiration time and send an otherwise normal email. This email is hosted by the provider (a local database in our case, Google in the case of Gmail’s confidential mode), and recipients are sent a link to read the email; Gmail clients automatically translate the contents of this link into a normal email. This email is deleted off Gmail servers at the appropriate time.

While the exact experience for self-destructing emails in Gmail’s confidential mode on the recipient’s end differs depending on whether the recipient uses Gmail or not, participants in our

---

\(^{3}\)Protonmail also has a self-destructing message feature, but it requires the choice of a password for recipients with non-Protonmail addresses; Gmail’s confidential mode has no such restrictions
study never received self-destructing messages, they only sent them; thus, the interaction model participants experienced was identical to that provided by Gmail’s confidential mode.

In our webmail interface, access to the self-destructing message feature is given via a dropdown at the top-right of the compose screen. Participants selected from a series of options for minutes, hours, and days. After expiration, a message continues to exist in the thread from which it belongs, but its body is replaced by a <Message timed out> indicator. This is shown in Figure 5.2. After a message has been sent, a visual indicator appears on the message in inbox view to alert both sender and recipient that the message will destroy itself after a specific period of time.

5.4 Methodology

Our IRB-approved study utilized a user study with both in-lab (N=23) and online (N=15) populations to investigate risk perceptions of normal email as well as the risk-response perceptions of two models of “secure” email when presented with four scenarios of varying sensitivity.
This user study involved two phases. In the first phase, participants interacted with our email system to roleplay standard email tasks—sending and receiving emails—in a set of four scenarios, presented in random order each time. In the second phase, participants first answered a short, post-task questionnaire gathering demographic data and assessing participant threat models and perceptions of response efficacy for normal email, encrypted email, and self-destructing email with respect to a handful of specific parties. Finally, participants engaged in a roughly 5-minute long interview with the study coordinator, who asked them about their risk perceptions of normal email and asked to review their thinking and behavior during the four scenarios. With the latter, they were asked to touch upon two aspects of each scenario: whether or not they considered the scenario “sensitive” and whether or not they had employed either (or both) of the protective features, and their reasoning for doing so.

In total, each individual study lasted approximately 30 minutes, for which participants were compensated $15 USD.

5.4.1 Sample population

For our in-lab population, we posted recruitment flyers on our local university campus, with a link to a scheduling page. Thus, our sample population—and demographics—are limited to those of a standard university population. Accordingly, our population skewed young, with 17 of our 23 participants (73.9%) between the ages of 18 and 24, and the remainder falling between the ages of 25 and 34. Our participants were also mostly female (N=17, 73.9%). The majority of our participants ranked themselves as “intermediate” on a 5-point, skills-oriented measure of technical ability (N=19, 82.6%), which corresponds with familiarity with spreadsheets and word processing, but excludes knowledge of how to program or use a database. 2 participants self-reported as having less expertise than intermediate, while 2 participants ranked themselves higher than intermediate (i.e., they had some degree of programming ability).
Our online population was recruited via the Prolific Academic platform. All participants were required to reside within the United States and speak English as a first-language. In contrast with our in-lab population, our online population skewed male (N=9, 60.0%). These participants also skewed young, although a bit older than our in-lab population, with 6 of 15 participants between the ages of 18 and 24 (40.0%), 8 between 25 and 34 (53.3%), and 1 between 45 and 54. Finally, as with our in-lab population, the majority of participants self-identified as “intermediate” on our 5-point measure of technical ability, with 1 participant rating themselves below 3, and 5 above.

5.4.2 Study protocol

Upon beginning the study, we first asked participants to read aloud an introduction to the study they were about to perform. This introduction informed them that they were to roleplay four standard email tasks and to act throughout as though they were in a real situation. They were further informed that two new features would be made available to them by our email system—that they could optionally choose to “encrypt and/or set emails to self-destruct”—but that their choice of whether to do so or not was up to them; there were no “right” answers and we were interested only in their reasoning for doing so.

Participants were then led through a series of four email scenarios, assigned in random order. These four scenarios consist of a two-by-two combination of “high” and “low” sensitivity email content as well as contact with an authority figure or a peer. We chose these two variables because we perceived they may be factors in participants’ determinations regarding scenario sensitivity. In addition to data sensitivity, whose impact on scenario sensitivity is straightforward, we also use authority figure vs. peer because it’s a common variable that we perceived might have some bearing on sensitivity and on response cost—it’s one thing to send a friend an email with some new feature enabled and a different thing entirely to send that to your boss or a new employer. The four scenarios we used were:

- Inviting a coworker to lunch. (Low sensitivity x non-authority figure)
• Emailing your boss to let her know you’re sick, and won’t be coming in for the day. (Low sensitivity x authority figure)

• Getting hired by a new employer, where you must send your mailing address and attach a pre-background check form with information such as Social Security number and address and employment history. (High sensitivity x authority figure)

• Having a conversation where you explain some emotional or mental issue, such as depression or substance abuse. (High sensitivity x non-authority figure)\(^4\)

Once participants had completed the four scenarios, they were directed to an 8-item questionnaire. The first 5 items were for demographics purposes, while the final three items correspond to an assessment of their understanding of risks for plaintext email and response efficacy for the two features. Participants were asked the following questions:

• *Suppose you send an email from your Gmail account, without turning on any encryption or expiration features. Who do you think can read the contents of this email? (Mark all that apply.)*

• *Suppose you send an email from your Gmail account with the optional encryption option enabled. Who do you think can read the contents of this email? (Mark all that apply.)*

• *Suppose Google added a feature to your Gmail account that provided timed expiration, like what you used today. If you sent an email with the message expiration feature enabled, who do you think could read the contents of this email after the email’s expiration time has passed? (Mark all that apply.)*

For each of these items, participants were given the same list of possible parties to choose from:

• You

• Your ISP (Internet Service Provider)

\(^4\)To avoid a scenario where participants might include sensitive information from their personal lives, we provided participants with a templated conversation and a list of issues to select from.
Finally, upon completing the questionnaire, the study coordinator conducted a short verbal interview with the participant. Each participant was asked to explain whether or not they had sent sensitive emails before, whether that act had concerned them, and whether or not there was sensitive content that would be convenient to send over email, but that they had been uncomfortable with doing so. They were then asked to review each of the four scenarios and explain whether they considered the scenario sensitive, whether or not they enacted the new protective features, and their reasoning for doing so.

The authors later reviewed the audio of these interviews together for analysis purposes. Each interview was broken down into its constituent parts to aid in analysis, and thus, for each participant, we identified their threat model with respect to general email, whether or not they considered a particular scenario sensitive, and whether or not they enabled a protective feature for that scenario and why. We then executed a thematic analysis on this data using a conventional qualitative content analysis approach, as described by Hsieh and Shannon [40].
5.5 Results

5.5.1 Feature use

Figure 5.3 shows the aggregate breakdown of which features participants enabled for each of the four scenarios (differences in usage rates between our in-lab and online populations were not statistically significant). Encryption numbers for scenario 1 are excluded from consideration because encryption was automatically enabled in that scenario. As can be seen, overall rates for feature usage were low, with only 12 out of 38 participants (31.6%) enabling a protective feature even in the scenario that was nearly unilaterally perceived as sensitive by our participants (the employment paperwork scenario).
5.5.2 General email use

We asked participants several questions regarding their past experience with normal email. More specifically, we first asked if they had ever previously sent sensitive content over email, and if so, whether or not they were worried about that content leaking. If they were, we then asked them to describe what parties they were concerned might gain access to that data. Finally, we asked them if there were occasions when it might have been convenient to send something over email, but where they had not done so due to concerns about email, and if so, what the nature of those concerns were.

The majority of our participants explained that they had indeed sent some sensitive content via email, but largely did not trust email with what they considered the most sensitive data—information that they perceived could be used to perpetrate identity theft. Instead, they chose to convey this information via other channels, often phone or in-person. Interestingly, the party they were most concerned about accessing their data was “hackers” with access to compromised account credentials.

While our participants had not previously adopted secure email systems, it was clear that they were instead adopting a preventive behavior to protect their data: self-filtering. For example, participant 3 explained how she would send sensitive information over email, but actively attempted to restrict it to that which she had previously sent already, limiting the scope of her exposure.

Concordant with this behavior, participants described two relevant concerns: a fear of data permanence, and a sense that once data is sent over the Internet, control over its access is now out of their hands. Participant 4 regularly used Snapchat to convey sensitive information because of its auto-deletion properties, explaining that with email, it “stays there forever even if you delete it.” Participant 16 remarked, “it feels like an inevitability: that if someone wants access, they’ll get it eventually.”

Indeed, several participants explicitly characterized a power imbalance between themselves and would-be attackers. For example, when queried about who she was concerned might have access
to her emails, participant 9 recounted a story about her brother and how his skill with computers had made her distinctly aware of the ease with which hackers might gain access to her account; by contrast, she felt her lack of ability in this domain made her even more vulnerable. Participant 10 was even more convinced of the futility of attempting to protect oneself from hackers. He mentioned a story (reported last October, by Bloomberg News) of how China had managed to plant microchips on server hardware and thus gain access to protected devices. Thus, he felt that even if software was trustworthy, there was no way of knowing if hardware was equally so. For this reason, this participant explained that he would only convey highly sensitive information, such as his Social Security Number, in person.

5.5.3 Scenario: Inviting coworker to lunch

This scenario involved participants initiating the conversation by sending an email to their coworker inviting them to lunch. Because this coworker’s email address fell within the 2040mail.com domain, our system enabled automatic encryption for this scenario.

While ostensibly the “lowest” sensitivity scenario of our four scenarios—the contact is a peer and the conversation content is casual—we nevertheless observed two interesting behaviors. First, as can be seen in Figure 5.3, one participant enabled the self-destructing message feature for this scenario. However, as might be expected, this was not done because this participant viewed this scenario as a sensitive one. Instead, this participant enabled self-destructing messages as a method for enforcing time-sensitivity: if their contact did not see the message in time, they did not want it to live in their inbox and potentially confuse the recipient.

As mentioned earlier, our automated encryption mechanism was enabled for this scenario. Surprisingly, only 4 of our 38 participants even noticed that this had occurred, despite the “Non-encrypted” visual indicator on the compose screen switching to “Encrypted” once the email address was fully typed out, i.e., before the email was sent. Of these, one participant misinterpreted what had occurred, believing that she had mistakenly enabled it.
5.5.4 Scenario: Calling in sick to work

This scenario tasked participants with initiating a conversation by sending an email to their boss to let them know they were too sick to come in to work. Compared with the previous scenario, this scenario introduces the additional components of communication with an authority figure and also a work environment. We were curious to see if the addition of these properties would alter perceptions of sensitivity in what is otherwise similarly a low-sensitivity conversation. However, no participants enabled a protective feature for this scenario for sensitivity reasons; the one exception did so for the same reason observed in the previous scenario; i.e., to enforce time-sensitivity.

5.5.5 Scenario: New employment information

In this scenario, participants first received an (expected) email from a new employer, asking for a mailing address. Upon responding, they were then sent a second email asking them to attach a PDF of a pre-employment background check document containing personally identifying information (PII). Examples of information that appears on the form includes Social Security number, driver’s license number, date of birth, address history, and employment history. This was the highest-sensitivity of our constructed scenarios.

As mentioned earlier, as participants were most concerned by the threat of a hacker seeking to engage in identity theft, the document was seen as extremely sensitive. Interestingly, while nearly all of our 38 participants agreed that this information was sensitive, over two-thirds (26 participants) nevertheless did not enact either protective feature for this scenario.

In total, 12 participants chose to enable at least one of the two protective features, with 4 participants enabling both. The password-based encryption feature was enabled by 6 participants, and self-destructing messages were used by 10 participants. These mechanisms were generally used to protect only the document, as participants explained that while a mailing address might be considered sensitive information, it was also considered generally available information.
As the scenario that participants perceived to be the most sensitive, participant responses regarding this scenario were also the most enlightening.

**Encryption**

A number of participants were intimidated by our use of the term “encryption.” The encryption mechanism provided for participants was either transparently automated encryption (which required no interaction on the participant’s part) or that of a simple password-protected email. Nevertheless, despite all of our participants having familiarity with a username and password paradigm—and thus ostensibly being capable of using the feature as presented—7 of our 38 participants explicitly described not even having tried the encryption feature because they did not believe they would be able to get it to work. Participant 3 explained this sentiment as follows:

“If it was available, that would be awesome, but I’d have to have some training and learn how to use it. I’m a pretty savvy computer user, but I’m not a programmer or anything like that, so I’d have to learn. [...] Encryption sounds scary to me, because I’m not a coder, so just that title alone is something that would terrify me. But if they relabeled it and made it approachable, I’d probably use encryption more.”

This affirms a suggestion in previous work [100] that self-efficacy is a critical component of users’ willingness to adopt secure email: if users perhaps do not believe themselves capable of using a feature or system, they will not use it, regardless of what the truth may be. With the model of encryption we presented users, usability was not an issue; however, participants’ lack of confidence in their ability to use this feature prevented them from even trying. Participants found self-destructing messages, on the other hand, to be fairly intuitive, although several participants remarked that it would be dependent on social convention for the handling of these messages being established in advance.

Similarly, it was apparent that participants had a vague sense of the threat model encryption is meant to resolve. Participants who enabled this function mostly saw it as generally associated
with higher levels of security, as opposed to preventing access from specific parties such as email providers, government surveillance, or a man-in-the-middle-attack. This aligns with a finding from our previous work [99] that users’ mental models of encryption is that of access control—they know it makes it harder for undesired parties to access their data, but not necessarily who or why.

Finally, the response cost of password-based encryption—needing to find a way to securely communicate the password to the recipient—was troubling to some of our participants. However, a handful of participants—all of the 5 who enabled both features in a single scenario—creatively overcame this response cost by using the self-destructing message function to transmit the password for password-encrypted emails in a secondary, plaintext email with the expiration timer set for a short amount of time. This allowed convenient communication of a password to a recipient without creating concern about the password living past its immediate consumption.

**Self-destructing messages**

Participants largely understood that self-destructing messages counter the threat of data permanence or of recipients being careless about their email security.

Some participants opted to enable the self-destructing message mechanism because they were explicitly worried by the threat of data permanence—sensitive information surviving longer than they wanted it to. The comments made by Participant 26 illustrate this point well:

“I think self-destructing is a kind of ’scorched earth’ method, where you make sure that the person that you want to see it sees it, and they get as much time as they need to see that information [...] and then after that set amount of time, it’s just gone.”

Other participants who enabled the feature did not necessarily explain a specific threat model they expected self-destructing messages to resolve, either using it because they wanted additional security but perceived encryption as too difficult to use, or without specifying any reason at all beyond wanting to defend against hackers.
Two participants felt there were times when data permanence is required, which makes the self-destruction functionality inappropriate. Specifically, they felt it could cause problems with workplace retention policies or if the information needs to again be accessed at a later date.

5.5.6 Scenario: Mental/emotional health conversation

Our final scenario involves a conversation with a peer that is sensitive for social reasons, rather than because it involves PII. This scenario had participants take part in a conversation about mental/emotional health. Interestingly, this different aspect of sensitivity teased out differences in both reasoning and behavior from participants. While most participants agreed that this scenario was sensitive, they did not see it as “compromising,” believing that even if a hacker could gain access to this information, a hacker would not possess the relevant social context required to be able to exploit such information. However, as Figure 5.3 shows, there were nevertheless participants who felt that this information did warrant protection. In total, 7 participants enabled at least one of the two features; 2 participants enabled encryption and 6 participants enabled self-destructing messages, with 1 participant enabling both.

Participants who enabled this feature were either generally uncomfortable with the thought of this information “floating around” on the Internet and being seen by others or were people who, while not concerned about a specific threat, simply felt more comfortable ensuring the information remained private. For example, participant 15 explained that they did not want someone to see it while “going through” their email, and so enabled encryption to “keep that more of a private conversation.” On the other hand, participant 32 was not concerned by hackers, but was instead concerned by socially-sensitive information, and thus enabled encryption only for this particular scenario.
5.5.7 Perceptions of risk and response efficacy

We asked each participant to explain which parties they believed they could read their emails in three cases: normal email, an encrypted email, and a self-destructed email after its timer had expired. Their responses for normal email characterize a threat model for email. Their responses for the two secure email features indicate the perceived response efficacy for these features. This data can be found in Figure 5.4.

Normal email: Participant 24 appears to have been confused by the task, as they did not mark ‘You’ or ‘The recipient’ as having access in the normal email case; all other participants responded normally.
We expected that an informed user would understand that plaintext email is accessible by Gmail employees and by hackers breaking into Gmail; we consider it somewhat surprising that significant numbers of participants did not understand this.

Given the easy downgradability of STARTTLS [19, 28], it is technically straightforward for governments to be able to read email, especially when sent to an address belonging to another provider, but we did not expect users to know this. Rather, this sentiment captures participant beliefs about the power of their own government (the United States) as well as its capabilities relative to those of other governments. Finally, roughly half of our participants did not understand that by using HTTPS to connect to Gmail they are preventing their ISP from viewing the contents of even plaintext mail.

**Encrypted email:** Participants generally viewed the response efficacy of encrypted email positively, believing it capable of denying access to most parties. Interestingly, however, many participants (24 out of 38, 63.2%) did not realize that encrypted email would not deny access to a hacker with access to their login credentials. Considering hackers were the threat most participants were concerned with, this is a worrying result.

**Self-destructing messages:** Although we did not explain this feature to participants, roughly half were nevertheless able to intuit its purpose, with 17 of 38 (44.7%) of participants marking that ‘Nobody’ would be able to access the contents of a message once it had self-destructed. It is, however, unclear to what extent the remainder of our participants simply did not believe that truly removing an email from the Internet is feasible. Promisingly, however, for the use of this feature as a response catering to participants’ concerns, only 6 participants believed that a hacker (breaking either into Gmail or their account) would still have access to a self-destructed message.
5.6 Discussion

The goal of our study was to understand how users reason about risk and response when presented with email scenarios of varying sensitivity and given two optional mechanisms for securing their messages. We now discuss the larger themes of our findings.

5.6.1 What does the “secure” in secure email mean?

“Secure email” is often equated with end-to-end encrypted email, which encrypts messages as they are in transit and while at rest. The security community recommended using PGP for many years, and in recent years secure webmail providers such as Protonmail and Tutanota have become popular. Both of these webmail services advertise end-to-end encryption as a primary feature, with advertising touting its protection against any party who may attempt to read users’ email messages.

Interestingly, however, end-to-end encryption does not appear to align with the threat models elaborated upon by our participants, who were not generally concerned by the thought of the government, a third party, or their email provider accessing their emails without their permission. Rather, participants were instead wary of hackers who might later gain access to their account via compromised login credentials. End-to-end encryption does not protect users in this case.

This is particularly interesting in light of secure webmail services, such as those listed above, that automatically encrypt messages between users of their system. Since these internal messages likely never actually leave the provider’s databases to travel across external email servers, protection in “transit” doesn’t seem to have practical meaning outside of protecting against the email provider itself. It may be possible that users of such systems operate under a different threat model from that described by our users, or that users have adopted use of secure webmail services without an awareness of the explicit risk-response model. This bears further investigation.

Given the different threat model, and our users’ strong reaction to the term “encryption”, it is possible that secure email providers could see higher adoption by instead labeling their product
as “private” email. A private inbox could hold messages with sensitive information and require an additional password or PIN protecting that inbox in case an attacker is able to learn their primary account credentials. The private inbox could of course be augmented with encryption to provide additional privacy protection for those users who do have strong concerns about their email provider or other organizations having access to their email.

Finally, we note that our work provides evidence that it is important to distinguish between the likelihood and severity of email risks. Many participants described the mental/emotional health scenario as being sensitive in nature, but were unconcerned about the information within being leaked because it was not seen as particularly compromising. Thus despite any likelihood of access by a hacker, the risk was not severe. Risk communication could help users to distinguish between these concepts, so they could take appropriate action when they consider both likelihood and severity of a risk to be high.

5.6.2 Data permanence worries users

Data permanence was a real concern for our participants. When characterizing past email use and subsequent self-filtering behaviors—i.e., a reluctance to send certain types of sensitive information over email—participants explained that once information is on the Internet, it remains “out there” to stay. This finding matches earlier reporting from Ruoti et al. [69], that found that participants were concerned by the threat of data permanence of even encrypted emails. Deleting emails manually was not an effective response in this scenario, as these emails would continue to live in the other party’s inbox. Participants who employed our self-destructing message feature felt it was a good match for the threat of a hacker compromising their account at some later time.

There were many times when participants chose to use the self-destructing messages feature but did not encrypt the data. This may be an opportunity to help users understand that encrypting messages protects them from some forms of unauthorized access prior to the time that they expire,
and also provides some assurance in case their service provider does not delete their data from their servers as promised.

The popularity of this feature also indicates the need for more service providers to enable this feature in a way that operates seamlessly across providers. The most natural way to do this is by *not* sending the email to the recipient’s email provider (where deletion cannot be controlled) and instead requiring the recipient to access it on the originating email provider’s server. This is how ProtonMail, for example, offers this feature, though they also require that the email be password-protected if the recipient is with a different provider. There is no technical obstacle to providing self-destructing emails as a separate feature from encryption.

5.6.3 *Automated encryption by trusted providers opens new possibilities*

We presented participants with two models of encryption, although only one model explicitly required interaction from our participants. Participants 1, 5, 10, and 27 remarked that they saw the response cost of password-based encryption as prohibitive. This mirrors results from Ruoti et al. [70], where users disliked having to find a secure method for communicating the password to the recipient and doubted people sending email to them would choose strong enough passwords.

However, our results indicate that automated encryption is a promising and viable option for further securing email without requiring users to proactively adopt the technology. During the first scenario, when emails were automatically encrypted, almost none of our participants noticed that this had occurred. This suggests that the usability impact of automated encryption was minimal, although it should be noted that we did not expose our participants to other aspects of encryption’s potential response costs; search and spam filtering are difficult for a provider to offer when emails are encrypted. This echoes findings from a number of previous studies [5, 6, 50, 70] that all used various forms of automated key servers to help users encrypt emails.

Automating encryption has been used to a high degree of success in secure messaging applications, such as Signal and WhatsApp, where each application is its own walled garden.
However, implementing automatic encryption across a generic, heterogeneous Internet of email servers is substantially more complicated. We can envision at least two practical ways to accomplish this.

First, because participants were not concerned by their email providers as part of their threat models, email providers could create and manage their own public and private keypairs. If the system allows automatic lookup of a recipient provider’s public key, then emails could be automatically encrypted provider-to-provider, while still allowing email providers to decrypt and inspect email for spam and malware. This would provide significantly more protection than STARTTLS (content-based encryption would not be susceptible to encrypted connection-based vulnerabilities such as a downgrade attack).

Second, this system could be easily extended to offer true end-to-end encryption if providers stored private keys for users that were encrypted with a password. The system could allow automatic lookup of a recipient’s public key through their email provider. It could also allow email clients to automatically retrieve private keys for users when they logged in. These two features combined would result in transparent end-to-end encryption. Private keys could potentially be discovered by providers using a brute force attack on weak user passwords, but this would at least offer a significant step up toward automatic encryption, and our participants generally expressed trust of their email providers. Note that in this case, providers could not scan incoming encrypted messages for spam or malware. One easy fix could be to only encrypt messages to recipients where a person responds to a plaintext message. Since users typically do not respond to spam or malware originators, this would provide a simple whitelist for end-to-end encryption, where recipients are automatically placed on the whitelist if the user’s behavior indicates they are not spammers or miscreants.

A world where email providers are not the enemy opens the door to ecosystems where automated encryption may not solve all concerns for all users, but would improve the security of the average user without negatively impacting their usability experience.
5.6.4 Habituation is powerful

The majority of our participants failed to enact any protective feature despite most viewing at least one scenario as sensitive. The most common reason they gave for this discrepancy was that of habituation—they were used to simply sending emails without thinking about it. Indeed, even one particular participant who is a secure email (Protonmail) user explained that habituation had gotten to them as well; because they use Gmail and Protonmail side-by-side, when presented with a Gmail-like interface, they defaulted to their standard Gmail behavior—sending an email without enabling additional protection. This has bearing on the addition of secure email features to popular existing webmail services: would users even notice—let alone enable—secure email features that have been added to services they are intimately familiar with?

5.7 Limitations

Our study is subject to several limitations. First, the threat models exhibited by our population are likely intimately intertwined with the cultural and political climate of the United States; it is quite likely that participants from another area would manifest different threat models and resulting behavior. Second, examined behavior was within the context of a lab study, and is subject to the corresponding effects, although we did attempt to mitigate effects such as the social desirability bias by explaining to participants at the start of the study that their reasoning—and not any particular behavior—was our object of interest. Indeed, the majority of our participants (22 out of 38) did not enable any protective features in any of the four scenarios and we observed strong habituation effects; it is possible, however, that observed ratios might be different in a real life scenario.

5.8 Conclusion

In this paper, we present findings from a user study we conducted to seek further insight into users’ decisions of whether or not to protect emails when presented with scenarios of varying sensitivity.
We designed an email system with two security mechanisms: encryption and self-destructing messages. 38 participants used this system to roleplay a set of four email scenarios with the option of enabling or not enabling these features as they saw necessary, and then explained their reasoning for doing so.

We find a mismatch between the protections offered by current secure email services and the threat models that our participants presented. Our participants were not concerned by the threat of email providers, but rather by hackers compromising their accounts and of data permanence. Manually-enabled encryption confused and intimidated some users, although our findings suggest that automated encryption is a viable path forward, particularly if email providers can be a trusted party.

Future work should focus on new models of what secure email could implement that would meet the privacy needs of users. Based on our work, a combination of self-destructing messages, a private inbox with a separate password or PIN, and automatic encryption could meet most needs. More work on interoperability of these features across providers is needed so that they can be made seamless for users.
Chapter 6

Conclusion

In this dissertation, we explored the application of principles from risk communication theory to the problem of grassroots adoption of secure communication tools. First, we explored mental models of encryption. We found that users’ perception of encryption essentially reduces to an access control abstraction and that its’ utility to the average user was generally perceived negatively. Second, we attempted to apply risk communication to system design, and evaluated its effectiveness in communicating with users about the purpose of the authentication ceremony in the Signal secure messaging application. Our work demonstrated positive effects on both comprehension and adherence, illustrating the potential benefits of this approach. Third, we tested the applicability of a core theoretical framework of risk communication theory—protection motivation theory—to the domain of secure email adoption, and then used this framework to evaluate the relative impact of its four constituent factors on predicting the adoption of secure email services. We found that all four components of protection motivation theory were statistically significant predictors of adoption, and that of these four, response and self-efficacy were the strongest predictors, although we can not demonstrate causality with self-efficacy due to limitations on our sample population. Finally, we evaluated perceptions of risk and response of secure email features when presented alongside email tasks of varying sensitivity. We found that end-to-end encryption of email aligns poorly with user threat models for email use, that unpermissioned access of persistent email content by hackers was a major concern, that self-efficacy was indeed an obstacle with respect to encrypted email, and that automated encryption of emails is a viable option for users.
We now discuss the contributions of our work. We also include both recommendations for improvements as well as directions for future work.

6.1 Contributions

The research contributions of this dissertation are:

1. **First directed effort exploring mental models of encryption.** We conducted the first targeted study of mental models of encryption, identifying four models of encryption, which largely distill down to an access-model abstraction. We found that users’ perceptions of encryption align with symmetric encryption, which helps explain the difficulties users face when interacting with asymmetric encryption mechanisms. We further identified response efficacy issues with perceptions of encryption, such as stigmatized views of its utility in daily life and confusion about attackers’ ability to break encryption in a timely manner.

2. **Identified obstacles to users’ understanding of the authentication ceremony in Signal (and other messengers using the same ceremony, such as WhatsApp).** We performed a cognitive walkthrough to identify obstacles to understanding of the authentication ceremony and then verified the existence of these obstacles via a user study.

3. **Designed, implemented, and evaluated comprehension-focused redesign of the authentication ceremony and surrounding messaging in Signal.** We designed, implemented, and evaluated a redesign of the Signal authentication ceremony and associated messaging using risk communication principles in a way that improved both comprehension and adherence levels. Our evaluation further allowed us to collect evidence suggesting that perceptions of risk must be divided into likelihood and severity components, and that the presence of risk is not alone a sufficient condition to predict the adoption of secure behaviors. Relevantly, we also provide evidence that the set of protective responses users consider at their disposal extends beyond the set of actions allowed them by the system.
4. **Evaluated the applicability of the protection motivation theory framework to the problem of secure email adoption.** We designed and evaluated a questionnaire whose items were composed of Likert-statements corresponding to each of the four components of protection motivation theory. Using a structural equation modeling approach and responses to our questionnaire from both secure email-using and non-using populations, we verified that all four components of PMT are individually predictive of secure email adoption, although we are unable to demonstrate causal effects for the factor of self-efficacy. We also found that response efficacy was the strongest predictor, with self-efficacy the strongest correlate, suggesting that the privacy paradox can be explained by response assessment failures as opposed to failures by users to accurately report privacy behavioral intentions.

5. **Studied risk and response perceptions of two models of secure email as perceived during four email scenarios of varying sensitivity levels.** We designed an email interface with two additional optional secure email features—encrypted emails and self-destructing messages—and a risk communication intervention delivered to a treatment group. Participants interacted with these features as they roleplayed a set of four email scenarios. We affirm our previous findings regarding models of encryption and the impact of self-efficacy with respect to secure email use. We also outline favorable conditions for an automated email ecosystem.

### 6.2 Recommendations

1. **Direct effort toward increasing response efficacy perceptions of secure communication technologies.** We found that participants generally did genuinely possess privacy concerns, but that they had a lack of knowledge of, and trust in, technology designed to resolve those concerns. Consequently, we found that participants often turned to analog channels for the communication of sensitive information, preferring to convey highly sensitive information over the phone or in-person. We recommend that associated messaging must focus on the ability of these tools to
make them safer than they already are as opposed to guaranteeing levels of security—guarantees that users often do not trust.

2. **Utilize risk communication principles in system design—and elsewhere—to more effectively communicate with end users the risks they face online and the responses available to them.** Our work demonstrates the potential efficacy of applying even simple risk communication principles to system design and messaging, achieving higher rates of protective feature usage as well as higher comprehension levels. In particular, our work highlights the critical need for more effective, empirically-tested communication about the range of responses that are available to users, and the efficacy of those responses. We also recommend that the applicability of risk communication principles and messaging be applied to a larger range of security/privacy features, such as password managers.

3. **Avoid mentioning encryption as much as possible to users; instead, focus on providing positive framing for the properties it provides.** We found that encryption confuses users; even those who favorably associate it with positive security outcomes do not necessarily have a model for how or why it achieves that purpose. On the contrary, we found that encryption carried negative connotations: participants viewed use of it in daily life as the domain of the paranoid or immoral and as a feature, complicated and intimidating enough as to fall outside the realms of easy use.

However, based on recommendations from risk communication literature and our own experience with redesigning the authentication ceremony in Signal, we believe there to be a positive effect by presenting users with a positive framing on the properties that encryption can give them in terms they can contextualize. Necessarily, the discussion of such beneficial properties must be tailored to each recipient populations, as the perceived value of privacy-enhancing technologies are largely impacted by the sociopolitical climate of the user. For example, our U.S.-based populations were largely unconcerned by thoughts of surveillance by the government; such views might not translate to populations living under more restrictive regimes.
4. **Automated email encryption—and perhaps automated key management in general—seems a viable path forward inasmuch as email providers are considered trustworthy.** A fully-automated approach would further bypass certain grassroots adoption obstacles, such as response or self-efficacy failures. Our work shows that users are most concerned by the threat of hackers compromising their data, as opposed to service-level entities such as service providers or the government. Thus, for our users, this opens the door for two styles of interoperable, automated encryption ecosystems that make them safer without negative impacts on their usability experience. First, email providers could implement automated provider-to-provider encryption using a federated key server in the style of the centralized key server utilized by services such as Protonmail or Signal. Second, providers could manage the distribution of end-to-end encryption keys on users’ behalf and automate interactions with keys, while locking access to these keys via a user-controlled password, such as in Protonmail or Lastpass.

5. **Users need the ability to control the removal of their data.** Our participants and others, such as those interviewed by Ruoti et al. [69] and Delta.Chat developers¹, described concerns regarding data permanence. Relevant to this concern, in our final study, we found that participants were generally fond of the self-destructing message feature we provided, largely in part due to the extent to which it aligns with this data permanence concern. We further noted, as did Kang et al. [46], that users demonstrated a strong sentiment that their data was simply “out there” in the Internet, vulnerable and exposed. These sentiments, combined with low response efficacy perceptions, led participants to rely on external methods for protecting their data, such as self-filtering even when inconvenient. Consequently, providers should design mechanisms—such as the self-destructing message abstraction we provided users in chapter four of this work—that allows their users to control the remote deletion of their data from shared domains, not just personal ones.

¹Delta.Chat needfinding report: [https://delta.chat/assets/blog/dcneedfindingreport.pdf](https://delta.chat/assets/blog/dcneedfindingreport.pdf)
6.3 Future work

While we have presented evidence that risk communication can, and should, be applied to the adoption of secure communication technologies, we also note the need for further work investigating this issue.

- **Study the effect of risk communication on adoption longitudinally.** Our studies were all conducted in one-time sessions, either in the lab or over the phone. Thus, our data did not provide us the ability to evaluate whether or not the effects we observed would persist over time. It is likely that, as with other things, such messaging would need to be repeated often for it to truly be internalized.

- **Evaluate alternative decision-making/technology adoption models for applicability to the secure communication adoption problem.** In this dissertation, we evaluated the applicability of one model—protection motivation theory—to secure email adoption. There are, however, multiple models and extensions to protection motivation theory that should likewise be evaluated. Similarly, the predictiveness of these models should also be evaluated on other types of privacy-enhancing technologies.

- **Study risk and response perceptions on different populations, particularly those facing different sociopolitical climates.** The participants of our students were almost entirely drawn from a U.S. population, and thus their perceptions of risk are shaped by the sociopolitical climate of this region. It is likely that the threat model of populations from other regions would differ in some ways, perhaps drastic ways; for example, perhaps we might observe that trust in email providers would not extend to other populations. This has impact on both the style and content of risk messaging that must be delivered to such populations as well as the design and architecture of privacy-enhancing technologies for their use.

- **Design and evaluate alternative models for communicating security/privacy risk and response in system interfaces.** We have traditionally provided visual indicators that offer binary
measures of safety and protective features that similarly operate under binary on/off conditions. However, due to the many variables involved in a truly accurate risk-response calculation, complicated by diversity in levels of user comprehension, it is not clear that users possess the ability to independently assess the response that most aligns with their values. Thus, rather than straightforwardly exposing users to technical manifestations of risk and response, we believe that it might be more effective to offer an alternative paradigm where users are instead probed for their threat model by the software, whereupon appropriate protective features are suggested and/or enabled by the system.

• **More thoroughly evaluate the effect of self-efficacy—confidence, as well as usability—on adoption.** In chapter three of this work, and to some extent—chapter four, we identified self-efficacy as a potential obstacle to the adoption of secure communication technologies. However, due to limitations of our sample population in chapter three, we were unable to demonstrate that self-efficacy has a direct negative impact on adoption. This link must be investigated further. If it is indeed an issue, then efforts to improve the usability of privacy-enhancing technologies are not enough; we must further target user confidence in the usability of such tools. How to improve confidence in the usability of a tool, however, is also a question requiring further research.
Appendix A

Study materials for “When is a tree really a truck? Exploring mental models of encryption”

A.1 Interview guide

A.1.1 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?

A.1.2 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?

A.1.3 Examples of encryption

1. Okay, we’re 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?
Appendix B

Study materials for “‘Something isn’t secure, but I’m not sure how that translates into a problem’: Promoting autonomy by designing for understanding in Signal”

B.1 Signal authentication flow

Figure B.1 shows a flow diagram of different screens in Signal when the encryption key changes for a contact, along with the transitions between screens based on user input. The top path is the “message not delivered” flow, which appears to send a message but shows a status indicating that the send failed. The bottom path is the “message delivered” flow, which only shows a notification but otherwise proceeds normally. The middle path is the “message blocked flow”, which prevents the user from sending a message initially.

B.2 Redesigned Signal authentication flow

Figure B.2 shows the redesigned authentication flow. There is only a single path, using a blocked message dialog along with a shield message in the conversation log.

Figure B.3 shows the new notifications that correspond to this flow. If the user attempts to send a message after the encryption keys have changed, the message is blocked and a privacy check dialog is shown (upper left). From here, if the user taps “Get Started”, they proceed to the privacy check screen (top middle). They can use either the phone call (top right) or QR code scanner (bottom right). They can choose “Not Now” from either the privacy check dialog or the privacy
Figure B.1: Flow diagram depicting how Signal reacts to a safety number change.

Figure B.2: Flow diagram depicting how our redesigned Signal reacts to a safety number change. The blue box encloses the elements and choices with analogous equivalents in the original Signal client. The area contained by the dashed lines shows choices, elements, and state changes that we added in our version that are expansions on the authentication ceremony and beyond.
check screen, and they will proceed to the reminder dialog (bottom left). The result of the privacy check (failure or success) is shown in Figure B.4.

Figure B.5 shows the notifications in the conversation log. First, when encryption keys change, a notification is displayed that recommends a privacy check (Figure B.5a). Later, if the user completes the privacy check, a different notification is shown if the identifiers match (Figure B.5b) or don’t match (Figure B.5c). These notifications scroll as new messages are added to the conversation.

B.3 Privacy check indicators

Figure B.6 shows the new privacy check indicators. Tapping on the indicator brings up the corresponding privacy check screen, depending on the current state of the conversation, as shown in Figure B.7. These same screens are accessed if a user taps of any the conversation log notifications.
Figure B.3: Privacy check notification flow

Signal assigns identifiers to each user's device. These identifiers are used to encrypt your conversations.

You and Alex can compare your copy of each other's identifiers to make sure they match.

If they match, you know that this conversation can only be read on Alex's device and your device.

To compare your identifiers with Alex, please select an option below.

- Not Now
- In Person
- Phone Call
Figure B.4: Phone call and QR code privacy check results

(a) Phone call: matching identifiers
(b) Phone call: non-matching identifiers
(c) QR-code: matching identifiers
(d) QR-code: non-matching identifiers

Figure B.5: Conversation log notifications

(a) Notification after a key change
(b) Notification after matching identifiers in privacy check
(c) Notification after non-matching identifiers in privacy check

Figure B.6: Privacy check indicator

(a) Default state
(b) Matching identifiers
(c) Non-matching identifiers

Figure B.7: Privacy check screen

(a) Default state
(b) Matching identifiers
(c) Non-matching identifiers

Figure B.4: Phone call and QR code privacy check results
Appendix C

Study materials for “Exploring the ‘Secure’ in Secure Email: Perceptions of Risk and Response in the Adoption of Secure Email Features”
C.1 Study coordinator instructions

Study Coordinator Guide

☐ Before the participants arrive:

1. Generate a random ordering of numbers from 1-4, and rearrange the participant’s roleplay sheets to match the ordering.

2. Identify if the participant is a tutorial user or not (round-robin).

☐ Read them the following:

We would like to invite you to take part in a study using a new email tool that we have designed. It will involve you roleplaying some standard email tasks – such as sending and receiving emails – and will end with a short verbal interview. The entire study should take approximately 30 minutes to complete, for which you will be compensated $15.

☐ Have the participants sign the consent form.

Please read the consent form I am giving you. The main points of this form are that:

1. We will have you engage in a series of email tasks using an email tool we have designed.

2. You should not use any of your personal information when performing these tasks; we will provide you with all data and forms that you will need to complete the tasks.

3. You will fill out a short questionnaire and conduct a short verbal interview after completing these tasks. Your answers will be published as
part of our research, but without any details that can be used to identify you.

4. We will record the audio of the verbal debriefing for transcription purposes, after which the audio will be deleted.

5. The study will last approximately 30 minutes, for which you will be compensated $15.

6. You may end the study at any time and we will delete all data collected at your request.

□ Tell the participant to read aloud the first page of their instructions.

□ Inform that they may use anything they’d like for name and email, and whether or not they need to check the tutorial box.

□ Once they are done, load the survey from Qualtrics:

https://byu.az1.qualtrics.com/jfe/form/SV_42P0WLkVGWCWo3zv

□ When they have completed the questionnaire, interview the participant:

- “Have you ever sent emails in the past that you would consider sensitive? Can you give me some examples?”

- “When you sent those emails, did it concern you that the information in them might get out?”

- “Is there ever information that it’d be convenient to send over email, but that you don’t because you’re concerned about email? What is it about email that makes you hesitant to rely on it in these situations?”

- “Who is it you’re concerned might gain access to your emails? For example: your email company, the government, hackers, etc.”
• “The other thing I’d like to do is to review each of the four scenarios with you. For each one, I’d like to touch on two things: 1) Whether you consider that scenario sensitive or not, and 2) whether or not you used encryption or self-destructing messages, and why or why not.”

C.2 Roleplay instructions

Participant role-play instructions

In this study, we are trying to understand how people balance the benefits that come from using improved security/privacy mechanisms against the potential inconveniences they can cause. For this reason, even though you’ll be using made-up information for the tasks in this study, we want you to consider how you would act in real life.

Our new email tool will give you options to encrypt and/or set emails to self-destruct. It is your choice whether to use these new features. In some cases you may feel it is important to protect your email with these features, and in other cases not. Any choice you make is fine; we would like you to act as you think you would in real life.

There is no right answer here. What we’re interested in is the reasoning behind your decision, whatever that may be.

Task
You will now be led through a series of four email scenarios, which you are to perform in order. All necessary data has been provided for you; use this data and not your own personal information to complete each task.

Please enter anything you’d like for the name and email fields. If prompted by the study coordinator, please check the box labeled “I’m a tutorial user.” Proceed to login.

When you are ready, go ahead and flip to the next page and begin the first task.

Task

Scenario: You and some other co-workers are planning on heading out to grab lunch. You want to know if one of your other co-workers, Jim, wants to come.

Begin a conversation with Jim by sending him an email to ask if he’s interested in coming with you.

Jim’s email address is: slimjim@2040mail.com

The study coordinator will let you know when you are ready to proceed to the next task.

Task
Scenario: It’s flu season again. You’ve just woken up and, as luck would have it, you’re pretty sure you have it now. Your head is spinning, you have a sore throat which makes it hurt to even try to talk, and your body aches all over. Since it hurts to talk, you plan to email your boss to let her know that you’re not feeling well, and that you intend to take the day off.

Let your boss know the situation by sending her an email.
Your boss’ email address is: andrea.johnson@gmail.com

The study coordinator will let you know when you are ready to proceed to the next task.

Task

Scenario: Congratulations, you’ve just found a new job, and that means it’s time for your favorite thing in the world – paperwork.

Step 1: Wait for an email from your new employer. When it arrives, respond as you would normally.

Step 2: Once you’ve completed step 1, another email from your employer should arrive next, asking for some additional information and documentation. You’ll find the documents you need in the “Downloads” folder of this laptop. (You don’t need to actually fill out the form; you can use it as-is.)
The study coordinator will let you know when you are ready to proceed to the next task.

Task

Scenario: You’ve been struggling with some stuff lately, and that’s shown in your posts on Facebook. Your friend, who’s worried, decides to reach out and see if everything’s okay.

You may use the following template if you’d like:

Hey, thanks for reaching out. The truth is that for the past ________ (day/week/month), I have been feeling ________ (anxious/lonely/hopeless/overwhelmed/stressed).

I’ve been struggling with ________ (depression/anxiety/alcohol or drug abuse/feeling worthless/thoughts of suicide)

Telling you this makes me feel ________ (nervous/anxious/embarrassed/scared), but I’m telling you because ________ (I’m worried about myself/it is impacting my school work/I don’t want to feel like this/I don’t know what to do/I trust you).

The study coordinator will let you know when you are ready to proceed to the next task.
C.3 Post-task questionnaire

1. What is your age?
   - 18-24
   - 25-34
   - 35-44
   - 45-54
   - 55+
   - I prefer not to answer

2. What is your gender?
   - Male
   - Female
   - Other
   - I prefer not to answer

3. Please specify your ethnicity.
   - White or Caucasian
   - Black or African American
   - Asian
   - Pacific Islander
   - Mixed race
   - Other
   - I prefer not to answer
4. What is the highest level of school you have completed or the highest degree you have received?

- Less than high school degree
- High school graduate (high school diploma or equivalent, including GED)
- Some college but no degree
- Associate’s degree in college (2-year)
- Bachelor’s degree in college (4-year)
- Master’s degree
- Professional degree (JD, MD)
- Doctoral degree
- I prefer not to answer

5. On a scale of 1 to 5, how would you rate your current technological expertise?

For the purposes of this survey, we’re primarily concerned with your computer and web-based skills. We’ve defined three points on the scale as follows. These tasks represent some of the things a person at each level might do.

**Beginner** (characterized as 1 and 2 on scale): Able to use a mouse and keyboard, create a simple document, send and receive e-mail, and/or access web pages.

**Intermediate** (characterized as 3 on scale): Able to format documents using styles or templates, use spreadsheets for custom calculations and charts, and/or use graphics/web publishing.

**Expert** (characterized as 4 and 5 on scale): Able to use macros in programs to speed tasks, configure operating system features, create a program using a programming language, and/or develop a database.

- (1) Beginner
6. Suppose you send an email from your Gmail account, **without** turning on any encryption or expiration features. Who do you think can read the contents of this email? (Mark all that apply.)

- You
- Your ISP (Internet Service Provider)
- Your government
- Other governments
- Google (Gmail) employees
- Hacker breaking into Gmail (not your individual account)
- Hacker breaking into your Gmail account
- The recipient
- Nobody

7. Suppose you send an email from your Gmail account, with the **optional encryption option enabled**. Who do you think can read the contents of this email? (Mark all that apply.)

- You
- Your ISP (Internet Service Provider)
- Your government
- Other governments
8. Suppose Google added a feature to your Gmail account that provided timed expiration, like what you used today. If you sent an email with the message expiration feature enabled, who do you think could read the contents of this email after the email’s expiration time has passed? (Mark all that apply.)

- You
- Your ISP (Internet Service Provider)
- Your government
- Other governments
- Google (Gmail) employees
- Hacker breaking into Gmail (not your individual account)
- Hacker breaking into your Gmail account
- The recipient
- Nobody
References


[100] Justin Wu and Daniel Zappala. To encrypt or not to encrypt: Exploring perceptions of secure email through the lens of protection motivation theory, 2019. in submission.