Understanding and Mitigating Online Scam Victimization Among At-Risk Users

Elijah Bouma-Sims

CMU-S3D-26-116

August 2026

Software and Societal Systems Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Lorrie Faith Cranor (Chair)
Nicolas Christin
Julie Downs
Michelle Mazurek (University of Maryland)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Societal Computing.

Copyright © 2026 **Elijah Bouma-Sims**

This material is based upon work supported by: the National Security Agency under Award No. H9823018D0008; the National Science Foundation under Award No. CNS-2150217 and Award No. DGE-2140739; the Carnegie Mellon CyLab Security and Privacy Institute through a Presidential Fellowship and the Secure Blockchain Initiative; the Innovators Network Foundation; Meta; and Google.

The views and conclusions contained in this document are those of the author and do not necessarily reflect the official policies or positions, either expressed or implied, of any sponsoring institution, and no official endorsement should be inferred.

Keywords: Security, Human-Computer Interaction, Scams, Fraud, Phishing, Warnings

To David, my best friend, partner, and the love of my life.

Abstract

Internet-based fraud and other deceptive schemes (collectively called "scams") are today's most pervasive digital safety threats. Although scams can victimize anyone, specific vulnerable or "at-risk" populations may face a disproportionately higher risk of harm. Prior research on scams has primarily examined these groups in isolation, and there is little consolidated understanding of why different populations are more susceptible to scams or how to design inclusive protections.

This thesis addresses this gap through three main contributions. First, it synthesizes existing knowledge by systematically reviewing 145 papers that present original findings on how internet scams affect various demographic and at-risk groups. The review analyzes the methods used in these studies, the risk factors that increase vulnerability, and the defensive measures proposed to protect these groups. It finds that factors tied to personal circumstances, such as underserved accessibility needs and resource constraints, appear to shape vulnerability to scams, though this evidence is largely qualitative. Studies of general demographic trends are often conflicting and uninformative about underlying causes. It argues for research that moves beyond demographic profiling toward a deeper understanding of contextual risk factors. These risk factors offer a lens for interpreting the empirical studies that follow.

Second, this thesis presents a series of empirical studies, particularly focused on two at-risk populations: LGBTQ+ users and minor teenagers. For LGBTQ+ users, a survey examines privacy practices among closeted individuals on online dating services, revealing how marginalization can result in vulnerability to extortion. For teenagers, two studies examine scam exposure and responses: a qualitative analysis of Reddit discussions about cryptocurrency in teenage communities, and an experimental comparison of teens' and adults' susceptibility to scams that offer free digital goods on YouTube. These studies emphasize how teenagers' resource constraints and distinct interests may increase their exposure to fraud. Complementing these group-specific studies, a large-scale thematic analysis of 1,525 posts from scam-focused Reddit communities identifies common support-seeking behaviors, scam recognition strategies, and emotional coping mechanisms.

Third, the thesis develops and evaluates a novel warning system for instant messaging scams. This system leverages generative AI to detect potential fraud and deliver context-sensitive advice to help users make safer decisions. The system design was developed through a qualitative interview study and three online experiments. These studies found that participants had a positive sentiment towards GAI explanations in warnings, and identified a design that promoted safer decisions in hypothetical scenarios. This design was evaluated in a first-of-its-kind field experiment, which found that both the GAI-generated and a baseline warning were associated with reduced clicking on simulated scam messages relative to no warning.

Ultimately, this thesis advances understanding of scam victimization and offers solutions to improve digital safety for everyone, especially underserved populations. By centering their perspectives, we can raise the baseline of protection for all users.

Acknowledgments

As John Donne wrote in *Devotions upon Emergent Occasions*, Meditation XVII (1624), "No man is an island, entire of itself." I owe many personal and intellectual debts, and while it is not possible to fully repay them, I hope these words can provide some meager compensation.

My primary and secondary schooling took place at Homewood City Schools (Homewood, AL). I received excellent schooling there from teachers such as Lisa Holland, Jane Petras, Kelly Reaves, Tim Hurry, Mark Helmers, and dozens of others. I was subsequently fortunate to receive a Park Scholarship to attend North Carolina State University, which funded a first-class education in engineering and the humanities. There, I found many important mentors, especially Dr. Brad Reaves, who taught me how to conduct security research. Dr. Ross Bassett showed me how to meld the humanities with engineering, a practice central to my current research. Dr. Yasemin Acar helped me conduct my first research in usable security and privacy, and has been a friend and mentor ever since.

As a graduate student at Carnegie Mellon University, I have benefited greatly from being advised by Dr. Lorrie Cranor. She has provided me with much invaluable guidance and feedback over the last five years. This dissertation would not exist without her. Dr. Nicolas Christin has also been a valuable mentor, collaborator, and running buddy. I have also been lucky to learn from and work closely with Dr. Lujo Bauer and Dr. Julie Downs. I also extend my thanks to Dr. Michelle Mazurek for serving as my external committee member.

In addition to those listed above, I would like to thank the other collaborators who contributed to the papers in this thesis, including Sanjnah Ananda Kumar, Hiba Hassan, Alexandra Nisenoff, Lily Klucinec, Mandy Lanyon, Enze Liu, Alexandra Xinran Li, Matt Ung, Wendy Bickersteth, Ginger Smith, and Ryan Li. I want to thank the other members of the CyLab Usable Privacy and Security (CUPS) Lab, who have provided thoughtful feedback and always been down to gossip in the cubes.

I am thankful to my family, especially my mother, Katherine Bouma, and my father, Robert Sims, who gave me so many opportunities, from piano lessons to scout camp, and set me on the path to success. My partner, David Mayes, was crazy enough to follow me to Pittsburgh and has been a constant source of support throughout graduate school. Finally, I am thankful for my dog, Bingus, who is the best research assistant any PhD student could ask for.

Contents

1 Introduction **1**

2 Background & Related Work **4**
2.1 Internet Scams 4
2.2 Online discussion of scams 6
2.3 Scams and at-risk groups 8
2.4 Human-centered anti-scam interventions 10
2.4.1 Warnings and other anti-scam design interventions 10
2.4.2 Training to reduce scam susceptibility 11

3 A Systematic Review of Internet Scam Research with At-Risk Users **13**
3.1 Methods 14
3.1.1 Paper Collection 14
3.1.2 Relevance Coding 16
3.1.3 Detailed Analysis 17
3.1.4 Limitations 19
3.2 RQ1: Study Goals and Methods 19
3.2.1 Populations Studied 20
3.2.2 Study Goals 22
3.2.3 Representations of Scam Victimization and Susceptibility 23
3.2.4 Types of Scams Studied 25
3.3 RQ2: Contextual Risk Factors and Scams 26
3.3.1 The Overarching Role of Targeting 26
3.3.2 Personal Circumstances 28
3.3.3 Societal Factors 32
3.3.4 Relationship Factors 34
3.3.5 Dispositional Traits and Behavioral Factors 35
3.4 RQ3: Protection Mechanisms for At-Risk Groups 36
3.4.1 Educational Interventions 36
3.4.2 Other User-Centric Anti-scam Interventions 38
3.5 Discussion 39

4 Exploring the Privacy Experiences of Closeted Users of Online Dating Services in the US 44
4.1 Methods . . . 45
4.1.1 Recruitment . . . 45
4.1.2 Survey . . . 46
4.1.3 Follow-up Interview . . . 48
4.1.4 Qualitative Data Analysis . . . 48
4.1.5 Correlational Analysis . . . 48
4.1.6 Inspection of Dating Services . . . 49
4.1.7 Ethics . . . 50
4.1.8 Limitations . . . 50
4.2 Results . . . 50
4.2.1 Participant Characteristics . . . 51
4.2.2 RQ1: Privacy Concerns . . . 53
4.2.3 RQ2: Privacy-protecting Behavior . . . 56
4.2.4 RQ3: Improving Privacy Experiences . . . 59
4.3 Discussion . . . 61

5 Investigating the Experiences of Teenage Cryptocurrency Users on Reddit 64
5.1 Methods . . . 65
5.1.1 Dataset . . . 65
5.1.2 Thematic Analysis . . . 67
5.1.3 Ethical Considerations . . . 68
5.1.4 Limitations . . . 68
5.2 Results . . . 68
5.2.1 Overview . . . 68
5.2.2 RQ1: How do teenage users access crypto assets? . . . 71
5.2.3 RQ2: What motivates teenagers to engage with crypto assets? . . . 73
5.2.4 RQ3: What types of harm do teenagers experience when using crypto assets? . . . 76
5.3 Discussion . . . 79

6 Investigating the Susceptibility of Teens and Adults to YouTube Giveaway Scams 82
6.1 Exploring YouTube Giveaway Scams . . . 83
6.1.1 Organic Searching on YouTube . . . 84
6.1.2 Identifying Giveaway Scams on 1% of YouTube . . . 85
6.2 Methods . . . 85
6.2.1 Experimental Design . . . 86
6.2.2 Recruitment . . . 87
6.2.3 Quantitative Data Analysis . . . 89
6.2.4 Qualitative Data Analysis . . . 90
6.2.5 Ethical considerations . . . 92
6.2.6 Limitations . . . 92
6.3 Results . . . 92

6.3.1 Participant Characteristics . . . 93
6.3.2 RQ1: Participant interactions with giveaway scams . . . 93
6.3.3 RQ2: Differences in teenager and adult behavior . . . 99
6.3.4 RQ3: Other factors impacting scam victimization . . . 100
6.4 Discussion . . . 101
6.4.1 Overview of key takeaways . . . 101
6.4.2 Recommendations to reduce victimization . . . 102

7 The Nature and Quality of Reddit Discussion about Scams 104
7.1 Methods . . . 105
7.1.1 Dataset . . . 105
7.1.2 Thematic Analysis . . . 105
7.1.3 Ethical Considerations . . . 107
7.1.4 Limitations . . . 108
7.2 Results . . . 109
7.2.1 Overview . . . 109
7.2.2 RQ1: Types of scams discussed . . . 112
7.2.3 RQ2: Type of support requested . . . 114
7.2.4 RQ3: Support type and quality . . . 119
7.3 Discussion . . . 125

8 Investigating User Reasoning with Context-Specific SMS Scam Warnings 128
8.1 Methods . . . 129
8.1.1 Recruitment . . . 129
8.1.2 Interview Procedure . . . 130
8.1.3 Message Stimuli . . . 132
8.1.4 Warning Designs . . . 133
8.1.5 Analysis . . . 134
8.2 Ethics Considerations . . . 134
8.3 Limitations . . . 136
8.4 Results . . . 136
8.4.1 Participants . . . 136
8.4.2 Reasoning about Unsolicited SMS Messages . . . 138
8.4.3 Effect of Warning Design Parameters . . . 144
8.5 Discussion . . . 147

9 A Real-World Evaluation of GAI-generated, Context-Specific Warnings 150
9.1 Pre-studies . . . 151
9.1.1 Overview . . . 151
9.1.2 Pre-study 1 . . . 152
9.1.3 Pre-study 2 . . . 154
9.1.4 Pre-study 3 . . . 154
9.2 Prototype implementation . . . 156
9.2.1 Android Frontend . . . 156

9.2.2 Server Backend . . . 158
9.2.3 Runtime Configurations . . . 159
9.3 User Study Methods . . . 160
9.3.1 Experimental Design . . . 160
9.3.2 Survey Instruments . . . 161
9.3.3 Recruitment . . . 164
9.3.4 Fraud Prevention . . . 165
9.3.5 Statistical Analysis . . . 166
9.3.6 Qualitative Analysis . . . 170
9.3.7 Ethics . . . 171
9.3.8 Limitations . . . 172
9.4 Results . . . 173
9.4.1 Participant Characteristics . . . 173
9.4.2 Behavioral Analysis of Simulated Messages ($\mathcal{H}_1$ and $\mathcal{H}_2$) . . . 174
9.4.3 Phishing Identification Task ($\mathcal{H}_3$) . . . 179
9.4.4 Warning Sentiment ($\mathcal{H}_4$) . . . 179
9.5 Discussion . . . 184
9.5.1 Key Takeaways . . . 184
9.5.2 Methodological Reflections . . . 185
9.5.3 Towards Inclusive SMS Warnings . . . 186

10 Conclusion **187**
10.1 Methodological Approaches to Understanding Scams . . . 188
10.2 Policy Implications . . . 189
10.2.1 Regulating platforms that deliver scam content . . . 189
10.2.2 Regulating platforms that process scam payments . . . 191
10.2.3 Other anti-scam policies . . . 192
10.3 Directions for Future Research . . . 192
10.3.1 Assessing anti-scam interventions . . . 193
10.3.2 Investigating scammer targeting . . . 193
10.3.3 Collecting large-scale scams data on social media . . . 195

A A Systematic Review of Internet Scam Research with At-Risk Users **196**
A.1 Codebooks . . . 196
A.1.1 Population . . . 196
A.1.2 Scam Type . . . 197
A.1.3 Risk Factors . . . 198
A.1.4 Methodological Characteristics . . . 199

B Exploring the Privacy Experiences of Closeted Users of Online Dating Services in the US **201**
B.1 Recruitment Materials . . . 201
B.2 Survey Instrument . . . 201
B.2.1 Demographics . . . 202

B.2.2 Outness Inventory Questionnaire . . . 203
B.2.3 Characterizing Use . . . 204
B.2.4 Privacy Attitudes and Behaviors . . . 205
B.2.5 Improving User Experience . . . 207
B.2.6 Conclusion . . . 207
B.3 Code Books . . . 207
B.3.1 Privacy Protecting Behaviors (Q16, Q21) . . . 208
B.3.2 Privacy Violations (Q22) . . . 209
B.3.3 Features to Improve Privacy (Q24) . . . 210
B.4 Outness Inventory Correlations . . . 212
B.5 Interview Participants . . . 213
B.6 Interview Guide . . . 213
B.6.1 Introduction . . . 213
B.6.2 Interview . . . 213
B.6.3 Conclusion . . . 214

C Investigating the Experiences of Teenage Cryptocurrency Users on Reddit 215
C.1 Code Book . . . 215
C.2 Filtering Keywords . . . 218

D Investigating the Susceptibility of Teens and Adults to YouTube Giveaway Scams 222
D.1 Artifact Information . . . 222
D.2 Survey Instrument . . . 223
D.2.1 Internet Usage Questionnaire . . . 223
D.2.2 Search result selection task . . . 225
D.2.3 Video task . . . 226
D.2.4 Post-task Questions . . . 227
D.2.5 Demographics . . . 229
D.3 Codebook . . . 232
D.3.1 Video and Website Codes . . . 232
D.3.2 Search Result Codes . . . 237
D.4 Recruitment material . . . 238
D.4.1 Parent/Child Recruitment Poster . . . 238
D.4.2 Prolific requirement advertisement . . . 239
D.5 Stimuli . . . 239
D.5.1 Spotify Stimuli . . . 240
D.5.2 Roblox Stimuli . . . 247

E The Nature and Quality of Reddit Discussion about Scams 254
E.1 Codebook . . . 254
E.1.1 Post Type . . . 254
E.1.2 Technologies & Communication Methods . . . 255
E.1.3 Type of Scam . . . 256
E.1.4 Type of Harm . . . 259

E.1.5 Type of Support Requested . . . 259
E.1.6 Remediation Steps and Other Advice . . . 260
E.1.7 Types of Personal Information . . . 262
E.1.8 Other Comment Types . . . 263

F Investigating User Reasoning with Context-Specific SMS Scam Warnings 265
F.1 Demographic Survey . . . 265
F.2 Interview Guide . . . 267
F.2.1 Introduction . . . 267
F.2.2 Background Questions . . . 267
F.2.3 Task Introduction . . . 268
F.2.4 Task . . . 268
F.2.5 Post-task questions . . . 269
F.3 Stimuli . . . 269
F.3.1 Fake Amazon Account Alert . . . 269
F.3.2 Fake Bank of America Alert . . . 270
F.3.3 Fake Delivery Alert . . . 270
F.3.4 Employment Scam . . . 270
F.3.5 Fake T-Mobile Giveaway . . . 271
F.3.6 IRS Arrest Threat . . . 271
F.3.7 FedEx Notification . . . 271
F.3.8 Healthcare Reminder . . . 271
F.3.9 Amazon OTP . . . 271
F.3.10 Overstock.com Promotion . . . 272
F.3.11 Bank of America PIN Alert . . . 272
F.3.12 Political Solicitation . . . 272
F.4 GAI Prompts . . . 272
F.5 Codebook . . . 273
F.5.1 Introductory questions . . . 273
F.5.2 Task . . . 274
F.5.3 Message Reasoning (Q7) . . . 276
F.5.4 Warning . . . 278
F.5.5 Post-task Questions . . . 282
F.5.6 Other Codes . . . 283

G A Real-World Evaluation of GAI-generated, Context-Specific Warnings 284
G.1 Main Study Materials . . . 284
G.1.1 Recruitment Materials . . . 284
G.1.2 Sample Participant Communications . . . 285
G.1.3 Screening Survey . . . 287
G.1.4 Onboarding Survey . . . 290
G.1.5 Midpoint Survey . . . 294
G.1.6 Conclusion Survey . . . 296
G.1.7 GPT 5.4 Prompts . . . 302

G.1.8 Message and Warning Stimuli . . . 304
G.1.9 Conclusion Survey Message Stimuli . . . 307
G.1.10 Codebooks . . . 308
G.2 Pre-study 1 Material . . . 313
G.2.1 Survey Instrument . . . 313
G.2.2 Message and Warning Stimuli . . . 319
G.2.3 Additional Results . . . 321
G.3 Pre-study 2 Material . . . 323
G.3.1 Survey Instrument . . . 323
G.3.2 Message Stimuli . . . 326
G.3.3 Additional Results . . . 329
G.4 Pre-study 3 Material . . . 329
G.4.1 Survey Instrument . . . 329
G.4.2 Message and Warning Stimuli . . . 335
G.4.3 Additional Results . . . 343
G.5 Pre-study Interface Designs . . . 347

Bibliography **354**

List of Figures

3.1 Overview of literature review methodology. 15
3.2 Temporal distribution of publications in the literature review sample. 19

4.1 Summary of Outness Inventory data for participants in dating privacy study . . . 51
4.2 Summary of Likert-scale questions on privacy concerns in dating privacy survey . 53
4.3 Summary of Likert-scale questions on information disclosure and privacy-protecting strategies in the dating privacy survey 57

5.1 Frequency of posts per year in teenage cryptocurrency sample 69

6.1 Example of a search result displayed in the YouTube giveaway scam experiment. 86
6.2 Participants' recommended actions in response to scam stimuli in the YouTube giveaway scam experiment 95
6.3 Participants' recommended actions in response to legitimate stimuli in the YouTube giveaway scam experiment 98

7.1 Frequency of posts per year within the sample of scam-related communities. . . . 109

8.1 Example screens displayed in the warning interview study. 131
8.2 Examples of the warnings presented in the warning interview study 135

9.1 Screenshots from the prototype used in the real-world warning experiment 157
9.2 Inbox/notification behavior by message in the real-world warning experiment . . 175
9.3 Message outcomes by warning condition in the real-world warning experiment . 177

B.1 Advertisements used to recruit participants in the dating privacy study 201

D.1 Advertisements used to recruit parents of teen participants in the YouTube giveaway giveaway scam experiment 238
D.2 Spotify legitimate 1 search result show in the YouTube giveaway giveaway scam experiment 240
D.3 Spotify legitimate 2 search result show in the YouTube giveaway giveaway scam experiment 241
D.4 Spotify legitimate 3 search result show in the YouTube giveaway giveaway scam experiment 242
D.5 Spotify scam 1 search result show in the YouTube giveaway giveaway scam experiment 243

D.6 Spotify scam 2 search result show in the YouTube giveaway giveaway scam experiment . . . 244
D.7 Spotify scam 3 search result show in the YouTube giveaway giveaway scam experiment . . . 245
D.8 Spotify clickbait search results show in the YouTube giveaway giveaway scam experiment . . . 246
D.9 Roblox legitimate 1 search result show in the YouTube giveaway giveaway scam experiment . . . 247
D.10 Roblox legitimate 2 search result show in the YouTube giveaway giveaway scam experiment . . . 248
D.11 Roblox legitimate 3 search result show in the YouTube giveaway giveaway scam experiment . . . 249
D.12 Roblox scam 1 search result show in the YouTube giveaway giveaway scam experiment . . . 250
D.13 Roblox scam 2 search result show in the YouTube giveaway giveaway scam experiment . . . 251
D.14 Roblox scam 3 search result show in the YouTube giveaway giveaway scam experiment . . . 252
D.15 Roblox clickbait search results show in the YouTube giveaway giveaway scam experiment . . . 253

G.1 Advertisement used to recruit Non-Prolific participants for real-world warning experiment . . . 285
G.2 Warning experiment pre-study generic warning . . . 348
G.3 Warning experiment pre-study spam warning . . . 348
G.4 Warning experiment pre-study long warning . . . 348
G.5 Warning experiment pre-study blocking warning . . . 349
G.6 Warning experiment pre-study accordion warning (closed) . . . 349
G.7 Warning experiment pre-study accordion warning (open) . . . 350
G.8 Warning experiment pre-study mockup texting interface shown without warning . 351
G.9 Warning experiment pre-study mockup texting interface with a spam warning . . 352
G.10 Warning experiment pre-study mockup texting interface with a blocking warning 353

List of Tables

3.1 Strings used to identify papers in the literature review sample 15
3.2 Overview of populations studied by papers in the literature review sample 21
3.3 Overview of study objectives in the literature review sample 22
3.4 Overview of methods used in the literature review sample 23

4.1 Summary of demographics of participants in the dating privacy study 49

5.1 Overview of coding results in teenage cryptocurrency study 69
5.2 Most common codes assigned to posts in the teenage cryptocurrency study 71

6.1 Independent variables investigated in the YouTube giveaway scam experiment. . 91
6.2 Overview of selected demographic and behavioral characteristics in YouTube giveaway scam experiment. 94

7.1 Overview of types of posts observed in the evaluation of Reddit discussion about scams 109
7.2 10 most common scam type codes in the evaluation of Reddit discussion about scams 113
7.3 Overview of the types of advice requested by Reddit posters in the evaluation of Reddit discussion about scams 116

8.1 Demographics of participants in warning interview study 137
8.2 Overview of action recommendations and reasoning in warning interview study . 139

9.1 Summary of demographics of participants in the real-world warning experiment . 174
9.2 Mixed-effects logistic regression predicting whether a participant opened a message in the real-world warning experiment 176
9.3 Mixed-effects logistic regression predicting whether a participant clicked a link in the real-world warning experiment 176
9.4 Mixed-effects logistic regression predicting whether a participant clicked the "Report & Block" button in the real-world warning experiment 178
9.5 Mixed-effects logistic regression predicting whether a participant noticed a good feature in the real-world warning experiment 180

B.1 Correlations between average Outness Inventory score and Likert scale questions in dating privacy study. 212

B.2 List of interview participants in dating privacy study 213

C.1 Coin keywords used to find posts in the teenage cryptocurrency study 219
C.2 Failed project keywords used to find posts in the teenage cryptocurrency study . . 220
C.3 Crypto exchange keywords used to find posts in the teenage cryptocurrency study 220
C.4 Gambling keywords used to find posts in the teenage cryptocurrency study 221
C.5 Keywords derived from general concepts that were used to find posts in the teenage cryptocurrency study 221

G.1 Warning experiment pre-study 1 nested model comparisons 321
G.2 Warning experiment pre-study 1 best fitting model for *dismiss* actions on Scam messages 322
G.4 Warning experiment pre-study 1 sentiment statistical testing results 322
G.3 Warning experiment pre-study 1 cross-tabulation of *dismiss* and *interact* outcomes by warning presence and by message 322
G.5 Warning experiment pre-study 2 message review task results 329
G.6 Warning experiment pre-study 3 nested model comparisons for scam messages . 343
G.7 Warning experiment pre-study 3 nested model comparisons for legitimate messages 343
G.8 Warning experiment pre-study 3 sentiment statistical testing results 343
G.9 Warning experiment pre-study 3 best fitting model for *interact* actions on Scam messages 344
G.10 Warning experiment pre-study 3 best fitting model for *dismiss* actions on Scam messages 345
G.11 Warning experiment pre-study 3 best fitting model for *interact* actions on legitimate messages 346
G.12 Warning experiment pre-study 3 best fitting model for *dismiss* actions on legitimate messages 347

Chapter 1

Introduction

Internet-based fraud and other deceptive schemes (collectively called "scams") are among the most common digital safety threats people face today. For example, the US Federal Trade Commission (FTC) received reports of $3.5 billion lost to imposter scams alone in 2025, with text, phone, email, social media, and other websites being the most common methods of contacting victims [507]. More broadly, Heung et al. conducted a survey on digital safety experiences with a representative sample of Americans ($n = 5,001$) in late 2023/early 2024, finding that 49.7% of participants reported having "been the target of a scam or phishing attack," the most common among the measured digital safety risks [225].

While anyone can be scammed, government reports and prior research suggest that harm from scams is not evenly distributed across the population. Most notably, adults 60 or older are the most likely to report fraud to the FTC and reported the greatest median loss of any age group [501]. People with cognitive impairments, including those with acquired brain injuries or older adults in cognitive decline, may have a diminished ability to detect deception and therefore avoid scams [155, 199, 247, 376]. Blind and low-vision users face a distinct accessibility barrier: visual security warnings [276, 562] and cues to deception, such as typo-squatting URLs [338], are often inaccessible to them. Scams also especially prey on the limited financial resources of people with low socioeconomic status [259, 524].

Although prior work provides valuable insights into the experiences of various groups with scams, there is no consolidated body of work that specifically examines why different at-risk populations are disproportionately affected by internet scams, what measures are necessary to protect vulnerable user groups, and what future research is required. These groups are heterogeneous, but examining them explicitly can surface insights that benefit both marginalized populations and the broader "average" user who faces similar difficulties but does not belong to any of the specified groups.

This thesis addresses this gap in three main ways. First, it comprehensively analyzes victimization by scams among at-risk populations. In Chapter 3, the thesis presents a systematic review of 145 papers reporting original findings on how internet scams affect different demographic groups. This review systematically analyzes the methods employed in these studies, the contextual risk factors [535] that increase vulnerability, and the defensive measures proposed to protect these groups. This review highlights that factors related to personal circumstances—such as underserved accessibility needs and lack of financial resources—appear influential in the context

of scams; however, these findings are primarily qualitative. Studies investigating general demographic trends in relation to scam vulnerability are often conflicting and uninformative about the underlying causation. The role of scammers' deliberate targeting of particular groups is also underappreciated. The review highlights several pervasive methodological concerns, particularly a lack of ecologically valid measures of online scam vulnerability. These findings reveal the need for more rigorous research on scam vulnerability that moves beyond demographic profiling to incorporate a deeper understanding of societal risk factors.

Second, this thesis presents empirical findings focused on two specific populations (LGBTQ+ people and minors), complemented by a broader analysis of how people discuss and cope with scams in online communities. These results are discussed through the analytical framework established by the literature review. Specifically, in Chapter 4, the thesis presents results from a survey of the privacy experiences of "closeted" LGBTQ+ users of online dating services who conceal their identity. Among other findings, the chapter highlights how marginalization can place users at risk for online extortion. Chapter 5 presents a study of minor teenagers' experiences with cryptocurrency based on posts from teenage communities on Reddit. This chapter offers some of the first evidence on how minors interact with crypto-assets, highlighting their reliance on adults to access mainstream investment platforms. It also provides qualitative evidence on teens being victimized by investment fraud, which is associated with the desire for easy, short-term profits. Chapter 6 also focuses on teenagers, presenting results from an online experiment comparing teens' and adults' susceptibility to scams on YouTube that purport to offer free digital goods. While not finding that teenagers were more vulnerable to the scam stimuli, teen participants were significantly more likely to report previously searching for one of the scam terms ("Free Roblox Robux"). This result exemplifies how particular scam content may target specific groups.

Complementing these demographic-specific studies, Chapter 7 presents a comprehensive thematic analysis of 1,525 posts across four Reddit communities focused on scam discussion. Online support-seeking may be particularly relevant for people of low socioeconomic status who have fewer resources to help them navigate digital safety threats [394, 524]. This analysis showed that users primarily use Reddit to seek assistance in identifying scams, discuss scammers' strategies, and cope with experiences of victimization. Underscoring the stigmatized nature of scam victimization, emotional support emerged specifically as a feature of the discussion above and beyond specific technical remediation steps.

Third, building on insights from these prior studies, the final two chapters present the design and evaluation of novel warning systems for instant messaging scams. Specifically, it utilizes generative AI to provide explanations tailored to the message content. Chapter 8 presents a formative interview study that explores user perceptions of existing warnings and of a "context-specific" design with text generated using GPT-4o. Chapter 9 further develops this design in three online experiments. A final design is tested in a first-of-its-kind real-world study in which participants were sent simulated phishing messages with and without the new warning design. This study found that participants exposed to the generative AI design found it significantly more helpful than users exposed to a generic warning design. Both designs reduced the click-through rate on phishing messages.

The thesis concludes in Chapter 10 with a discussion of the thesis's methodological approach, policy implications of the results, and directions for future research.

Thesis statement

This thesis advances the understanding of scam victimization among at-risk populations through three core contributions: a comprehensive review of existing research identifying relevant risk factors and open research challenges; a series of empirical studies investigating scam-related experiences among specific vulnerable groups and online communities; and the design and evaluation of an inclusive, generative AI-powered warning system for instant messaging scams that delivers context-sensitive guidance to support safer decision-making.

Chapter 2

Background & Related Work

This chapter first defines the term "scam" and then provides an overview of the online scam threat environment, with a special focus on instant messaging-based fraud. It then describes online discussion of scams, with a focus on Reddit. It further explores research on at-risk groups and discusses the types of scams that may target them. The chapter closes by discussing human-centered approaches to scam mitigation, including design-based interventions and education aimed at reducing scam susceptibility. [1]

2.1 Internet Scams

Scam is an informal term used in both colloquial speech and academic literature. While uses vary, this thesis adopts the definition from Button and Cross: a scam is "a deceptive scheme that seeks to trick a person(s) out of money and/or personal information...." [81, pg. 7]. Many types of scams far predate the internet; [2] However, this thesis focuses only on *internet scams*, meaning those that are initiated via the internet or related technologies (e.g., phone calls, SMS, social media, etc.). Because they manipulate people rather than exploit technical vulnerabilities, scams are commonly described as *social engineering*, or in Schneier's taxonomy, as *semantic attacks* that target how people assign meaning to content [425].

Scams are among the most common cybercrimes. For example, phishing, in which an attacker impersonates a trusted entity to obtain sensitive private information such as a user's password or tax ID number, was the single most common crime reported to the Federal Bureau of Investigation (FBI) Internet Crime Complaint Center (IC3) in 2025 [497]. Due to its significance as a threat to organizational and individual security, phishing has received the most attention from researchers [142, 349, 570].

Based on fraud reports collected by the Federal Trade Commission (FTC), email, phone calls, and text messages were the most frequently used methods for initiating scams in 2024,

[1] This chapter is adapted from the prior work underlying this thesis [57, 58, 59, 60, 61]

[2] For example, in the 1887 case of *United States v. Hoeflinger*, the eponymous Joseph Hoeflinger was indicted for mail fraud following a scheme where he sent letters to potential victims leading them to believe that they were heirs to a fortune in England and requesting a fee in exchange for his services [125, pg. 627]. As observed by Abrams [2], this ruse is still commonly used by internet scammers today.

with phone calls and text messages resulting in the largest median losses to individuals.[3] Like phishing, most scams involve impersonating trusted entities or individuals [501, 507]. To extract money or personal information, scammers typically threaten the user with harm (e.g., arrest [504], the release of explicit media [85, 366], etc.) or offer something desirable (e.g., a job offer [506], prize [503], investment opportunity [505], etc.). Scammers may craft communications to take advantage of specific persuasion techniques, like the tendency to obey authority or reciprocity [5, 176]. Alongside psychological tactics, scammers also employ technological methods, such as caller ID spoofing [48], to enhance the believability of scams. Scams can cause not only direct financial harm but also other harms resulting from the disclosure of personal information, such as account compromise [377] and identity theft [130]. Victims may also experience significant psychological harm [131, 326], especially when they have formed a romantic relationship with the scammer [544].

Instant messaging scams are of particular importance to this thesis, with Chapters 8 and 9 focusing on an intervention to better protect users from these scams. While receiving less attention than email-based fraud, several prior studies have examined how users perceive and evaluate SMS messages [250, 440, 470, 481]. Tabassum et al. conducted an interview study ($n = 29$) investigating participants' experiences with scams and their approaches to processing instant messages, showing that users rely on contextual reasoning and content-based cues, such as URL features or message formatting [470]. Katsarakes et al.'s eye-tracking study ($n = 25$) similarly found that users focus attention on the body of SMS messages rather than the sender phone number or other interface elements when making judgments [250]. Timko et al.'s survey study ($n = 187$) identified demographic correlates of scam susceptibility and found that attention to specific message features (particularly sender information and URLs) predicts correct classification [481]. Finally, Siadati et al. demonstrated the effectiveness of SMS-based attacks targeting 2FA codes and suggested text changes as a mitigation measure [440].

Several theoretical models are commonly used to explain users' vulnerability to scams. Two closely related "dual processing" models [409], namely the Elaboration Likelihood Model (ELM) [380] and the Heuristic-Systematic Model [93] (HSM), each present persuasion as occurring along two distinct cognitive routes. The peripheral (ELM) and heuristic (HSM) processing routes are simpler paths that rely on indirect cues to authenticity, such as appeals to authority or urgent language. The central (ELM) and systematic (HSM) routes are engaged when a user analyzes a proposition more carefully, using critical thinking to evaluate evidence. These models suggest that scams succeed especially by engaging the peripheral/heuristic routes, pushing victims to make quick decisions without critically examining a proposition. These models may help explain how situational variables (e.g., time constraints, emotional arousal, or cognitive fatigue) diminish a user's capacity for central processing, thereby increasing susceptibility to deceptive tactics.

Routine Activity Theory (RAT) is applied to analyze scams at a macro-sociological level [117]. RAT posits that crimes occur when a motivated offender and a suitable target meet in the absence of a capable guardian (i.e., a person or other protective mechanism). RAT and related theories suggest that differences in online routine may increase or decrease users' exposure to crime. RAT would predict that greater time spent online, particularly engaging in risky behaviors, would in-

[3] As of writing, the Federal Trade Commission has not released detailed data for 2025.

crease the likelihood of being a scam victim. While not adopting a particular theoretical lens to analyze scam vulnerability, this thesis generally focuses on how long-term macro-sociological variables affect scam-related behavior. In this sense, this thesis is aligned with RAT and related frameworks.

Before proceeding, a brief discussion of terminology is necessary. Specific types of scams can be challenging to describe due to a lack of uniform terminology in this area. Scams are commonly categorized by the distinctive features of the social engineering tactics used. For example, the term *romance scam* typically refers to schemes wherein a fraudster uses a false persona to imitate an online relationship with a victim and subsequently extract money from them [542]. Along these same lines, the term *tech support scam* refers to schemes, often (but not exclusively [3]) committed over the phone, in which fraudsters convince victims to pay for fake technical support for some non-existent issue with their internet-connected device [324, 454]. Some researchers have proposed broader categorization schemes. Most comprehensively, Beals et al. [39] developed a five-level framework for categorizing different fraudulent schemes based on the target of the fraud (i.e., organizational or individual), the victim's expected benefit/outcome (e.g., investment returns, product or service, etc.), and other attributes (e.g., the types of technologies used to facilitate the fraud, whether the victim is an older adult or not, etc.). In the absence of a generally accepted approach to terminology, this thesis generally adopts the labels for specific scams that are most informative and contextually appropriate.

2.2 Online discussion of scams

Victimization by internet scams can be difficult to characterize accurately. Some of the most comprehensive data on scam victimization comes from consumer reports to government entities like the FBI [497] or the US FTC [501, 507]. Consumer reports are not a perfect representation of real victimization patterns. Prior studies have generally found that the vast majority of scam cases go unreported [18, 231, 391]. For example, an analysis of surveys administered by the FTC in 2005, 2011, and 2017 suggests that only 4.8% of US mass-marketing scam[4] victims reported their experience to a government agency or the Better Business Bureau (BBB) [18]. Alongside other factors [253], underreporting may result from the pervasive stigma around victimization. Scam victims are often viewed as responsible for their experience and maligned as gullible [127, 131]. Survey and interview studies, where users are asked to discuss their experience with scams, can partially fill the gap. However, even in anonymous surveys or interviews, participants may still fail to report their experience when asked. For example, in a 2003 AARP phone interview study with recent adult victims of telemarketing fraud who state and federal agencies identified, 47% of known lottery scam victims and 73% of known investment scam victims reported that they had **not** been victimized by a "major Scam or swindle" in the past 3-years [1].

In this context, online discussion can provide valuable insight into the experience of scam victims who did not report their experience to government entities. This thesis includes several studies based on posts and comments on the American-based social media platform Reddit

[4]Mass-marketing scams is a term that refers to fraud perpetrated via mass-communication methods, including digital and analog means (e.g., the mail)

(Chapters 5 and 7). The platform is composed of thousands of user-created and moderated communities called "subreddits," that focus primarily on specific topics (e.g., `/r/NFL` for the American National Football League) or identity groups (e.g., `/r/TwoXChromosomes` for women). Users submit posts that link to external websites or contain original content. Users vote to determine the default order of posts, with an "upvote" generally boosting a post and a "downvote" lowering the post's visibility. Users can also comment on posts, with the order of comments determined by the same voting scheme. Reddit has many communities dedicated to receiving advice, including advice related to scams (e.g., `/r/Scams` and `/r/Sextortion`).

A growing body of research has investigated help-seeking related to scams on social media [94, 105, 310, 351]. A substantial portion of this work has focused on Reddit discussion of financially-motivated sextortion [85, 531, 539], a type of scam where victims are tricked into sharing intimate imagery and then extorted with the threat of publication. Sextortion may also refer to cases where the scammer falsely claims to have explicit media featuring the victim [366]. Wei et al. [539] used large-language models (LLMs) to identify over 100,000 posts across Reddit where users sought advice about image-based sexual abuse (IBSA), including financially motivated sextortion. They analyzed a stratified sample of 261 posts (50 of which focused on financial sextortion) to evaluate how users seek and receive help with IBSA on Reddit. Focusing on scams involving crypto assets, Childs [105] qualitatively analyzed the top 200 posts from `/r/CryptoCurrency` that mentioned the word "scam" to assess how users prevent and cope with victimization.

Most relevant to this thesis are a paper by Oak and Shafiq [351] and a study by Chanenson et al. [94], both of which use Reddit data to characterize scam-related help-seeking and both of which engage directly with the work in Chapters 5 and 7. Oak and Shafiq directly build on the methodology presented in Chapter 5 to analyze scam discussion, with their sample focusing on the largest community discussed in Chapter 7 (`/r/Scams`), supplemented by several crypto- and scambaiting-focused communities. In contrast to Chapter 7, they analyze their data using a set of broader thematic categories (e.g., grouping remediation advice into high-level themes such as technological protections, verification practices, and community vetting). Chanenson et al. follow Wei et al.'s LLM-based sampling technique and analyze original posts (excluding comments) across a broader set of subreddits, including ones not focused on scams (e.g., `/r/FacebookMarketplace`). They position their work relative to Chapter 7 and corroborate many of the post-level findings. They also apply Warford et al.'s [535] framework to identify contextual risk factors that appear within scams discussed on Reddit.

Another line of research has focused on online vigilantism, referred to as "scambaiting," wherein users pose as victims and interact with scammers to waste their time or cause harm (e.g., by deleting files). While the practice of scambaiting has existed for decades [484], in recent years it has become more prominent, with popular scambaiting creators emerging on user-generated content websites such as YouTube and Twitch [43, 410]. On Reddit, `/r/scambaiting` is dedicated to the practice, with users sharing examples of scams and their experiences interacting with scammers [156]. Most research on the practice has focused on characterizing the social dynamics of scambaiting [43, 410], without a focus on security and privacy issues. One exception is a study conducted by Edwards et al. [160], which utilized transcripts of email exchanges from scambaiting communities to measure the use of persuasion tactics in advance-fee scams and train a classifier to distinguish between scammer and scambaiter emails. In general, analysis of

scambaiting discussion has limited utility due to the artificial nature of scambaiting interactions and the potential for wholly fabricated exchanges raised by Dynel and Ross [156].

2.3 Scams and at-risk groups

Privacy and security researchers have given special attention to the needs of vulnerable or at-risk groups due to the increased risk of harm they face. Chapter 3 reviews this literature in a systematic manner; however, I provide a brief introduction here. This thesis builds on the systematization of Warford et al. [535]. They reviewed 95 papers covering 31 distinct population groups, identifying 10 contextual risk factors that augment security and privacy risks. For example, users with *access to sensitive resources*, such as healthcare workers [10, 460], may be especially targeted by attackers seeking to exploit the data or systems these individuals control. Users who are *resource or time constrained* lack the time, money, or other resources to deal with security and privacy risks effectively. Returning to the healthcare context, the high time pressure of emergency medicine can contribute to insecure practices like account sharing and password reuse [460]. In the context of scams, people with low socioeconomic status (low-SES) may be more likely to be victimized by scams purporting to offer a prize or a job opportunity due to their lack of financial resources [524].

Older adults are often considered the most vulnerable to consumer scams, as those 60 or older are the most likely demographic to report falling victim to internet fraud [497, 501]. This statistic could be affected by reporting bias, as research indicates that consumers who are more highly educated or those who experience a greater loss are more motivated to report fraud victimization [120, 253, 255, 390]. The results of past experimental studies on phishing are mixed [431, 540, 570], with some studies finding victims are more likely to be younger (e.g., [420, 438, 533]), other studies finding them more likely to be older (e.g., [30, 201, 202, 297]), and still others finding no relationship to age (e.g., [368, 419]). These studies are typically conducted with users in the workplace, so they may not generalize to the population of retired older adults. Assuming that older adults are more susceptible to scams, one partial explanation may be cognitive decline with age. Mild cognitive impairment in older adults is associated with a greater risk of prior fraud victimization [155, 247, 376]. People of any age with a cognitive deficit, such as an acquired brain injury, may also be more vulnerable to scams [199].

Other disabilities may also impact users' susceptibility to scams. For example, in their study of the security and privacy challenges faced by people with visual disabilities, Napoli et al. found that none of their blind or low-vision participants ($n = 14$) could successfully identify a spoofed website [338]. While some blind participants attempted to verify the URL of the phishing website, the spoken URL obscured the difference between the real and spoofed domain names. Similarly, Yu et al. evaluated the accessibility of email phishing warnings in Gmail with a sample of low-vision screen reader users ($n = 21$) [562]. They found that their participants often overlooked existing graphical warnings and proposed a new, more inclusive design. These challenges in avoiding phishing reflect an *underserved accessibility need* that places blind or low-vision users at greater risk [535].

This thesis includes studies focused on two at-risk populations in particular: LGBTQ+[5] users and teenagers. The following research provided the motivation to conduct these studies. LGBTQ+ people face pervasive discrimination [88, 91, 110, 157, 222] and even legal persecution [321]. As a result, some LGBTQ+ people (those who are *in the closet* or *closeted*) choose not to disclose their sexual orientation or gender identity to others [72, 211, 308, 361]. The internet has allowed LGBTQ+ people to explore their identity while remaining in the closet [145, 185, 476]; however, it also may place closeted users at greater risk of scam victimization. For example, Li et al. conducted a mixed-methods study of the effects of a 2013 Indian high court decision to reinstate a ban on sodomy that criminalized non-heterosexual sexual activities. The participants reported a variety of physical and digital victimization experiences, including financial extortion after meeting individuals online [290]. In the US context, the FTC has previously warned users about scammers using dating apps to blackmail closeted LGBTQ+ users with the threat of exposure to friends, family, or employers [280]. Within Warford et al.'s taxonomy, these risks result from *legal or political* threats and *marginalization*.

Minor teenagers (users between the ages of 13 and 17) are also a potentially vulnerable population. During adolescence, emotional and social networks in the brain mature faster than the prefrontal cognitive-control network. This makes teenagers more susceptible to heightened impulsivity, sensation seeking, and emotional reactivity. These developmental dynamics contribute to a general trend of increased risk-taking behaviors among teenagers that declines with age [278, pgs. 528 – 530]. Because risk-taking is associated with fraud susceptibility, it is hypothesized that teens may be more likely to fall victim to fraud than adults.

Only a small number of human-subjects studies have investigated scam susceptibility among minors [14, 275, 344]. For example, Nicholson et al. evaluated the ability of English children aged 12 to 17 ($n = 83$) to recognize phishing emails in an experimental setting [344]. Participants generally performed poorly, with an overall success rate of 59% correctly discriminating between legitimate and fraudulent emails, but the researchers observed no difference with age. Certain types of scams are hypothesized to primarily target minors. For example, previous research has identified a kind of scam referred to as *giveaway scams*, which purport to give valuable rewards for free in exchange for completing a task but generally involve a never-ending string of surveys or other tasks with no actual reward [28, 54, 111, 114, 257]. Some of these scams promise free access to rare or paid-for in-game materials to lure victims. It is hypothesized that these scams especially ensnare children due to their focus on games that are popular with minors [28, 54]. These scams may result in users providing personal information or downloading potentially unwanted programs without compensation [114, 257].

Some scholars have critiqued the population-centric framing of vulnerability research in computing. Tang et al. argue that this approach obscures the heterogeneity within groups, risks perpetuating stigma associated with certain identities, and masks the reality that all users—regardless of group membership—are vulnerable to digital safety risks [472]. Indeed, a 2023/2024 survey conducted by Heung et al. with a representative US-based sample ($n = 5,001$) found that 85% of respondents strongly agreed with at least one statement corresponding to a contextual risk factor, suggesting that many people feel that they experience these factors [225], not just members of specific groups focused on within the academic literature. Acknowledging these

[5]This acronym refers to lesbian, gay, bisexual, transgender, queer, and other gender/sexual minority groups

drawbacks, I argue that this analytic frame helps to surface important structural factors that contribute to internet scam victimization. These factors should not be interpreted as universal within population groups, nor are they isolated to particular groups. Throughout this thesis, I hope to highlight nuance and the effects of intersecting identities where appropriate.

2.4 Human-centered anti-scam interventions

Preventing scam communications from reaching individuals, such as through automated detection [150], is the best way to prevent victimization; however, it remains impossible to screen out all scam communications. Users, therefore, must be relied upon to identify and screen materials that evade automated detection mechanisms. I review two approaches to helping users avoid scams: 1) security warnings and other design cues that aid in scam identification, and 2) educational interventions that teach users to avoid scams.

2.4.1 Warnings and other anti-scam design interventions

Research on security warnings builds upon prior research on general warning design, including studies on risk communication from outside the computing field. The Communication-Human Information Processing (C-HIP) model, developed by Wogalter et al., is particularly influential in this context [548]. C-HIP models how a warning produces behavioral change through a series of stages, requiring the delivery of the warning from a source to a receiver via a communication channel. To prompt action from the receiver, an effective warning must capture and maintain attention, support comprehension through clear language and appropriate visuals, and ultimately motivate safe behavior.

Warnings are widely implemented to notify users of potential scams. For example, many web browsers warn users when they attempt to visit a website that has been reported for phishing, malware distribution, or other dangerous behavior [195]. Due to the potential for false positives, browsers should not entirely block users from visiting websites flagged as phishing; however, an effective warning should significantly reduce the likelihood that users will proceed to a website. Early browser phishing warnings and other security indicators had design flaws that made them ineffective. For example, Egelman et al. conducted a lab-based experiment to test the efficacy of various phishing warnings implemented in popular browsers of 2007 (Firefox 2.0 and Internet Explorer 7.0) [161]. Participants who saw an active warning, which required them to interact with the notice before proceeding, were much more likely to exit a simulated phishing website without entering information than users who saw a passive warning, which required no interaction. Moreover, they found evidence of habituation: participants often clicked through the active warning from Internet Explorer without reading it, which the researchers attributed to its visual similarity to a less severe SSL certificate warning.

Based on this and other research [63, 146, 423, 466, 550], browser-based warnings have evolved to become more effective at promoting secure behavior. Akhawe and Felt conducted a large-scale measurement of users' interactions with security warnings on Mozilla Firefox and Google Chrome in May and June 2013, finding that the majority of users heeded malware and phishing warnings in the browsers, with less than 30% of warnings being ignored [7]. In contrast

to earlier, less effective variants, modern phishing and malware warnings are active, requiring users' interaction to proceed. They are also "fail-safe," with the most prominent option being to exit without proceeding to the dangerous website. Ignoring the warning and proceeding to the website requires multiple steps, making it more difficult for users to dismiss it without thought. An experience sampling study conducted by Reeder et al. in 2015 found similarly high compliance rates for malware and phishing warnings in Chrome and Firefox [397].

A sub-area of warning design research has focused on creating explainable security warnings. Prior work suggests that contextual warnings improve users' comprehension and decision-making across contexts, including phishing emails [75, 379], phishing websites [20], Android malware [24], and phishing URLs [13]. Before LLMs, such warnings typically involved manually preparing messages or message templates or presenting important decision-making features. With the rise of LLMs, numerous researchers have applied them to produce more user-friendly warnings. Studies to date have applied LLMs to warnings on phishing URLs [389], emails [90, 143, 295], and websites [412]. Most relevant to this thesis, Wang et al. developed and evaluated *SmishX*, an LLM-based, explainable phishing detection system. The results of their online user study found that users find context-specific explanations helpful [534]. We build on these results in Chapters 8 and 9, testing the effect of LLM-generated warnings in a real-world context.

Other design interventions seek to help users detect scams by emphasizing evidence relevant to determining whether something is a scam. In practice, this type of passive intervention is easily ignored or misinterpreted [296, 298, 329, 554, 568]. For example, browsers may highlight the domain name in a URL to draw users' attention to the most salient portion for detecting phishing. Studies of URL highlighting show that it is ineffective, as many users fail to inspect the address bar in browsers, and non-experts may have difficulty parsing URLs correctly [296, 554]. Similarly, Zheng and Becker conducted an experiment ($n = 252$) to evaluate whether the presentation of email header information, including sender name and email address, helped users identify phishing [568]. They found that users generally failed to detect phishing, regardless of whether sender information was shown or hidden.

Interventions of this type that require active interaction may be more effective. Lain et al. developed and tested several "URL inspection tasks" to help users detect imposter URLs in phishing emails by forcing them to read the URL and verify that it leads where they expect [274]. For example, one task requires users to correctly type the domain name they expect to go to in a text box before proceeding to the website. In multiple role-play experiments ($n = 2,673$), participants assigned one of the URL inspection tasks upon clicking on a link were significantly less likely to fall for phishing attacks than users who performed no task, users who saw only a passive warning, and users who performed a control task that did not involve verifying the intended destination.

2.4.2 Training to reduce scam susceptibility

Much research has explored training to reduce user susceptibility to social engineering, especially email-based phishing. This research often focuses on large organizations, such as companies or universities, where regular security training is recommended as a best practice, including, but not limited to, training on phishing [235].

Researchers have tested a large variety of different educational interventions for email-based phishing [242, 311]; however, *embedded training* has emerged as a popular method to deliver targeted training for phishing. Meant to capture the teachable moment that occurs when someone falls for a phishing attack, embedded training involves sending simulated phishing emails and redirecting those who click on links in the email to some form of anti-phishing training. This approach was pioneered by Kumaraguru et al. Their initial real-world evaluation of an embedded training system called "PhishGuru" ($n = 515$) found that users who received embedded training had improved resiliency to phishing emails, even after 28 days [270, 271]. Others have since found that simulated phishing emails delivering embedded training can reduce susceptibility to email phishing [441, 467]. Other studies have questioned the practical value of embedded training, finding no or only small effects from such training [86, 227, 272, 273]. Most recently, Ho et al. conducted a simulation study at a healthcare organization with around 19,500 employees to evaluate the effectiveness of embedded training and annual security awareness training in reducing phishing susceptibility [227]. They found no effect of yearly security awareness training and only a small effect of embedded training materials provided by ProofPoint (an average 1.7% lower phish rate). They attribute the low efficacy of embedded training to the fact that only a small proportion of users (those who fall for a phish) receive it, and most users who do receive it ignore it. Users who completed an interactive embedded training session (e.g., where they were asked to label suspicious parts of an email) had a $\sim 19\%$ reduction in phishing failure rates.

Consumers may take advice from workplace training and apply it to their everyday lives, making it relevant to consumer security practices. Still, many people, especially those of lower SES, do not receive such training and must take their advice from other sources [394]. There is a vast amount of educational content related to scams that is distributed by various organizations, including government entities (e.g., [500, 503, 505, 506]) and service providers (e.g., [25, 158, 166, 374]). These generally consist of static content to raise users' awareness of particular scams and provide strategies for avoiding victimization.

There is limited research directly investigating the value of this type of educational material, but it seems unlikely to be effective. For example, Robb and Wendel tested the efficacy of three training modalities (one interactive and two static, text-based) in reducing susceptibility to Social Security Administration imposter scams. In the interactive modality, participants were shown tips about spotting imposter scams and then asked to classify six emails as real or fraudulent, receiving feedback on their performance after each email [406]. Participants were tested on correctly classifying 12 communications (eight emails, two text messages, and two letters) as legitimate or fraudulent. Only participants who received the interactive training modality showed a significant improvement in performance compared with a no-training control group, with this improvement primarily evident in email classification. A follow-up study testing a similar intervention reached comparable conclusions [139]. Jakobsson and Gerell tested the efficacy of a postcard-based informational campaign to inform Swedish adults over 70 about telephone scams [240]. The results were inconclusive, with no significant reduction in fraud reports from this demographic compared to adults who did not receive the postcard.

Chapter 3

A Systematic Review of Internet Scam Research with At-Risk Users

In this chapter, we systematize the existing literature on scam victimization among *at-risk users* through the lens of Warford et al.'s framework of contextual risk factors [535]. In particular, we seek to answer the following research questions:

1. What are the goals of scam research with at-risk groups, and what methods are used to achieve these goals?
2. What contextual risk factors contribute to increased scam victimization among at-risk groups?
3. What defensive measures and other recommendations are proposed to protect at-risk groups?

To answer these research questions, we conducted a systematic literature review of papers in computing and the social sciences on internet scams and at-risk populations. First, we used structured searches in the ACM Digital Library, IEEE Xplore, and Scopus to identify papers that contain relevant terms in their abstracts, titles, or keywords, yielding 930 bibliographic entries. Two authors then reviewed the abstracts to identify potentially relevant papers. A team of eight authors then read and analyzed this set, selecting a final corpus of 145 papers that present results about scams among at-risk users or demographic differences in scam vulnerability. This analysis was performed in two rounds, the first of which focused on papers from 2013 to 2023. The second round included papers from 2024, as well as papers identified through recommendations from domain experts.

Based on this sample, we provide three main contributions. First, we summarize the state of research on internet scams focused on at-risk populations, highlighting methodological shortcomings and knowledge gaps (RQ1). Much research has focused on documenting the experiences of older adult victims or assessing general demographic trends in scam victimization. In contrast, insights about other potentially vulnerable groups, such as immigrants and people with visual impairments, come primarily from small-scale qualitative research. More research is necessary to explore the extent to which these findings can be generalized. Additionally, some

This chapter is based on a paper that is under submission, which was co-authored with Alexandra Xinran Li, Ginger Smith, Wendy Bickersteth, Mandy Lanyon, Alexandra Nisenoff, Matt Ung, and Lorrie Faith Cranor

research in this space relies on questionable measures of scam victimization, which may limit the utility of the results.

Second, we systematize this research through the lens of Warford et al.'s taxonomy [535], thereby highlighting the structural inequalities that contribute to scam victimization (RQ2). In particular, we find that issues related to personal circumstances—including *underserved accessibility needs*, *resource and time constraints*, and *loneliness or isolation*—are the most salient in the context of scams. Societal factors, including *social norms*, *legal or political factors*, and *marginalization*, may also affect vulnerability and the ability to recover from scams. We also highlight the critical role of *targeting* by scammers, which includes both tailoring to exploit specific risk factors and using private data to identify potentially vulnerable individuals.

Third, we discuss proposed interventions to better protect at-risk groups and make suggestions for future research in this area. We find that educational interventions are the most frequently discussed and evaluated protective measures, although there is generally little research on other inclusive interventions. We call for more research exploring scam vulnerability among groups beyond older adults and encourage the use of more rigorous measures of scam vulnerability in these studies. We also recommend that researchers consider interventions beyond training that may reduce scam victimization. Finally, we urge more focus on scammers' actions to understand the nature and role of demographic targeting, which remains underexplored.

3.1 Methods

We conducted a systematic literature review of 145 papers, focusing on peer-reviewed work published in journals, conferences, or workshops between 2013 and 2024. To be included, a paper must present original results related to internet scams and their impact on at-risk populations. As our focus is on consumer fraud, we exclude papers whose target populations face risk primarily through their role within an organization (e.g., bank employees, emergency room staff). By contrast, we include marginalized independent workers (e.g., sex workers or domestic workers) as they encounter fraud as individuals navigating informal or unregulated markets. Their lack of institutional support places them closer to vulnerable consumer populations in terms of exposure profile and capacity to recover from victimization. We also include papers that examine sociodemographic characteristics (e.g., gender, age, educational attainment) in relation to scam susceptibility or impact.

In the following subsections, we describe the process of paper selection and analysis. This process occurred over two rounds, each of which followed a substantially similar procedure.

3.1.1 Paper Collection

To identify papers, we performed structured queries on the ACM Digital Library (DL) Guide to Computing Literature, IEEE Xplore, and Scopus. We selected ACM DL and IEEE Xplore for their coverage of top security and privacy venues. Scopus was included to capture interdisciplinary and social science research not indexed in CS-specific repositories. Using appropriate syntax for each database, we used the following search string to identify papers on ACM DL and IEEE Xplore: *Abstract:(<GROUP>) AND Abstract:(phishing OR scam OR fraud OR extortion*

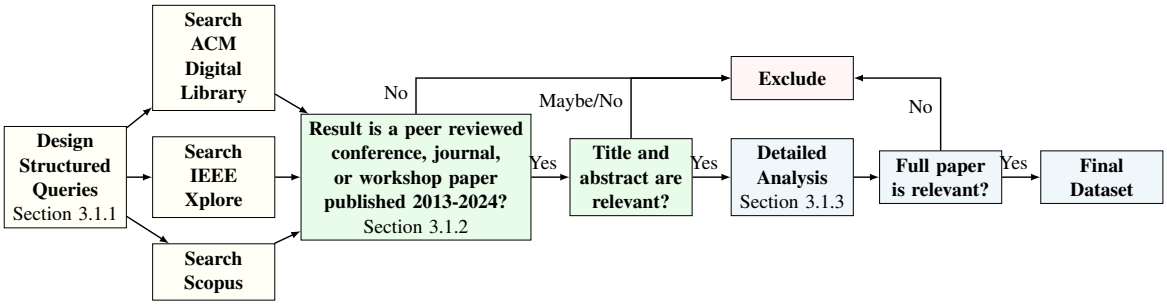


Figure 3.1: Overview of literature review methodology.

Table 3.1: Strings used to identify papers discussing certain at-risk populations. Specific phrasing was chosen to limit false positives (e.g., the term "black" was not used as its inclusion would result in too many false positives)

Group	<GROUP> String
People who are marginalized in society	*"marginalized"*
Non-western users	*"non-WEIRD" OR "non-western" OR "global south" OR africa**
LGBTQ+ People	*gay OR lesbian OR transgender OR nonbinary OR sexuality OR gender OR queer OR LGBT**
Older adults	*"older adults" OR "older adult" OR "aged population" OR "aged populations" OR elder**
People with lower socioeconomic status (low-SES)	*"Low income" OR "socioeconomic status" OR poverty*
Minors	*"young adult" OR "young adults" OR adolescen* OR teen* OR child**
Women	*women OR girls*
Racial and Ethnic Minorities	*"Hispanic" OR "Latino" OR "Latina" OR "Latinx" OR "African American" OR "afro-" OR "racial minority" OR "indigenous" OR "native american" OR "ethnicity" OR "ethnic minority"*
Victims of human trafficking	*"Trafficking" OR survivor*
Sex workers	*"sex workers"*
People with disabilities	*disability OR "visually impaired" OR "deaf" OR "hard of hearing"*
Migrants	*"immigrant" OR "emigrant" OR "migrant" OR refugee*

OR "social engineering" OR spam OR "identity theft") The placeholder <GROUP> corresponds to the strings listed in Table 3.1 that are meant to identify papers that include results related to a

particular at-risk user population. The specific populations chosen were selected based on those discussed in Warford et al. [535], with a focus on marginalized populations. Some words are not in quotation marks to allow for appropriate automatic substitutions in each database (e.g., scam $\approx$ scams).

When used on Scopus, the search string for IEEE Xplore and ACM DL produced many irrelevant results. We therefore adapted the search string to be narrower: *Abstract:(<GROUP>) AND Abstract:(phishing OR scam OR "internet fraud" OR (extortion AND ("internet" OR "online")) OR spam OR "identity theft")* This adapted search string excludes papers where *extortion* does not appear alongside *"internet"* or *"online,"* as many papers about non-internet-based extortion appeared without this restriction. It also does not include the term *"social engineering,"* as it frequently refers to societal change outside of the context of information security. Beyond narrowing the search strings, Scopus searches were configured to include only relevant academic fields ("social sciences," "psychology," "engineering," "economics, econometrics and finance," "decision science," "computer science," "business, management and accounting," and "multidisciplinary") and to exclude ineligible publication types.

Round 1 paper collection took place in August 2024 to identify papers published between 2013 and 2023. We conducted the second round of searches in August 2025 to add papers from 2024. Informed by external feedback, round 2 also incorporated suggested papers from the fields of criminology and victimology that were not surfaced in our original searches [264, 300, 301, 334, 375]. In October of 2025, we also conducted exploratory searches to identify new papers focused on groups underrepresented in round 1. For example, only one paper discussed LGBTQ+ people during round 1. Based on [474], we performed searches on terms like *"sexual minority," "bisexual," same-sex, trans* and looked at SOKs and literature reviews about at-risk users and marginalized groups [41, 316, 418]. We were unable to find additional relevant papers for our review. We performed similar procedures to attempt to identify papers focused on people with disabilities and non-Western users. Finally, two relevant papers cited in the sampled papers were added during the detailed analysis [432, 562].

3.1.2 Relevance Coding

In round 1, after screening out any results that we identified as not peer-reviewed journal or conference papers, 728 papers remained for review. Two authors performed initial coding to identify which papers appeared to meet our inclusion criteria based on each paper's title and abstract. We started by dual-coding three sets of 70 papers ($\approx 10\%$), meeting after each set was complete to discuss coding and reconcile differences. On the third set of 70 papers, the percent agreement between the coders was 85.7%. We coded the remaining 518 papers separately.

At this stage, papers were assigned to one of three categories developed in the dual coding process: "relevant," "maybe relevant," and "not relevant." "maybe relevant" included papers that did not unambiguously meet the inclusion criteria. This included:

- Papers focused on at-risk users that mention cybercrime in general, but the abstract was unclear about whether they discuss scams.
- Papers in which the abstract discusses demographic differences, but the paper is not focused on demographic comparisons.

- Papers that propose a general educational intervention to mitigate scams that is not specifically targeted at an at-risk population.
- Unless the abstract otherwise discusses another at-risk population, papers that deal with scams and non-Western users in non-Western contexts (e.g., South Africans in the South African context).
- Technical papers focused on scams (e.g., measurement of scams, AI-based classification of scams, etc.) that mention at-risk populations in their motivation, but not in other parts of the abstract.

The "not relevant" category was assigned to any paper that was not about an at-risk population and/or did not seem to contain a result about scams. Literature reviews, books, or other excluded publication types that remained in the sample at this stage were also assigned the "not relevant" code. Through this process, we identified 141 papers as relevant and 68 as maybe relevant. As the relevant category already included a substantial number of papers, we opted to review only papers coded as relevant.

Relevance coding in round 2 followed the same structure as round 1, with the same two authors performing the analysis. Due to the time gap between round 1 and round 2 coding and the smaller number of papers, we dual-coded all abstracts and discussed any disagreements to resolve them. The criteria for categorization remained mostly the same, with one exception: based on the experience of the round 1 analysis, we decided to categorize papers discussing identity theft generally as "maybe relevant." During round 1, we assumed that papers mentioning identity theft were relevant, even where they didn't mention internet scams. This led to a substantial number of irrelevant papers being read and then excluded during the detailed analysis.

For round 2, 38 papers were categorized as relevant, 28 papers were categorized as maybe relevant, and 136 papers were categorized as irrelevant. As with round 1, only papers categorized as relevant were analyzed.

3.1.3 Detailed Analysis

To analyze papers, we adopted a qualitative analysis approach based on thematic analysis [279]. This approach has been used in prior HCI literature reviews [178]. In round 1, a team of six researchers, each with at least one year of research experience in human subjects research, read the papers identified as relevant in the previous stage. For each paper, the coders wrote a short summary, described any recommendations made by the paper, and coded features relevant to our research questions. The complete codebook is located in Appendix A.1.

To answer RQ1, we coded the study's objective (i.e., descriptive, experimental, or relational) and the user study methods employed (if any). The codebook for this analysis was based upon Distler et al.'s review of methods in usable privacy and security [148]. We also coded the type(s) of scams that were the focus of the relevant results based on the Stanford Center on Longevity and FINRA Investor Education Foundation taxonomy of fraud [39]. To answer RQ2, we identified which at-risk populations and contextual risk factors were discussed in each paper. The codebook for this process was developed based on Warford et al.'s framework systematizing at-risk user research [535]. To answer RQ3, we coded whether a study presented or evaluated an educational

intervention or a prototype. We also noted any explicit recommendations made by the paper regarding research or practice.

Coding proceeded iteratively. First, all six coders read the same set of 11 papers in order to gain familiarity with the codebook and the analysis process. This initial set of papers was purposively selected to cover a range of populations and study types. The remainder of the papers were read in sets of roughly eight, each read by at least two authors. All coders met after each set to compare coding, resolve disagreements, and note emerging themes. The lead coder reviewed the summaries and coding to ensure consistency. As appropriate, we excluded some papers during this process when their contents were found not to meet the inclusion criteria. For example, some papers mentioned scams as examples of potential harm affecting a group but included no results related to scams. We found other papers to be ineligible (e.g., posters or literature reviews). Some discussed fraud broadly, with no indication that online scams were included.

Throughout the first round of coding, we made changes to the codebook to better reflect the properties of the literature. For example, after several rounds of coding, we noticed that *younger adults*, roughly 18 to 29 years old, were frequently considered at higher risk of scam victimization. Therefore, we added this population to our codebook. For this change and others made throughout the analysis process, the lead coder revisited previously reviewed papers to ensure that the code was applied appropriately across the sample.

After all papers were analyzed, the coding team met for the final time to identify common types of recommendations across papers (RQ3), synthesize recommendations for future work, and discuss other themes that emerged throughout the coding process. The results of this discussion form the basis of the recommendations presented throughout the paper.

The round 2 analysis followed a substantially similar procedure. This round was conducted by a team of four coders, three of whom had participated in the round 1 analysis. To ensure alignment between the experienced researchers and the new team member, we conducted a retraining session in which all four coders analyzed the same seven papers. Following this calibration, the remaining round 2 papers were dual-coded in pairs. The codebook refinements developed during round 1 were applied consistently across both rounds.

Ultimately, we included 103 papers analyzed in round one and 42 from round two in the final sample. Figure 3.2 shows the distribution of papers over time, reflecting a large increase in papers over the last few years. A plurality of papers were published in venues focusing on computing and information sciences, especially HCI and cybersecurity (62). A significant portion of the sample was also published in venues focused on gerontology (19) or criminology (17). The remaining papers were loosely distributed across other fields, including multidisciplinary venues, psychology, education, medicine, social work, ergonomics, law, and media studies.[1] Finally, 92 of the papers were published in journals, with the remaining 53 papers being published at conferences or workshops.

[1]Publication venues were coded by discipline by a single author using descriptions from the venue website or SCIMago. This breakdown should be read as indicative rather than precise.

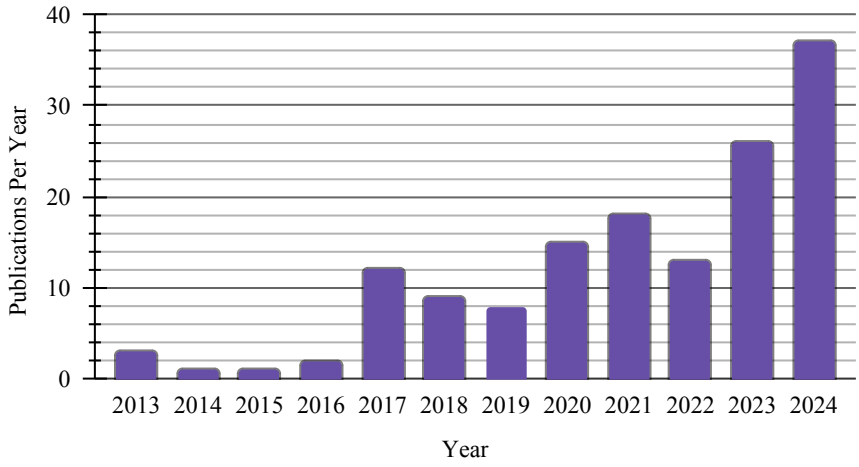


Figure 3.2: Temporal distribution of publications in the literature review sample.

3.1.4 Limitations

While our keyword-based sampling approach enabled us to identify a range of papers focused on at-risk populations and scams, it likely led us to overlook relevant papers that did not include the specified keywords or were not in the databases we searched. Keywords were selected based on populations we expected to be at risk, meaning that groups not generally considered at risk in the context of internet safety are not included. Our sample is drawn from English-language literature and primarily composed of papers focusing on the US and European context, meaning that the experiences of people in other regions of the world are underrepresented. Furthermore, while we adopt a systematic approach to identifying papers, our results consist primarily of qualitative observations rather than quantitative conclusions drawn across all studies in this area. Quantitative or semi-quantitative qualifiers (e.g., "many," "some," etc.) are made with reference to the sample of papers we viewed, and should not be interpreted as reflecting all research in this domain. Finally, we recognize that our coding results are inherently influenced by our perspectives as researchers. Other researchers may have identified different themes after reading the same papers.

3.2 RQ1: Study Goals and Methods

In this section, we provide an overview of the goals of the studies in our sample and the methods used to achieve them. The majority of studies in our sample focused on older adults or on general demographic trends in North America, Europe, or Asia. Quantitative data on non-age-related group differences were limited, with insights primarily coming from descriptive studies, raising questions about their generalizability. Most quantitative studies sought to understand how scam susceptibility varies across groups, using a variety of constructs. As a result, synthesis of study results is challenging, making it difficult or impossible to draw firm conclusions about group-

specific vulnerability. This difficulty is further compounded by the lack of uniform definitions of scam types, which makes cross-study generalizability highly problematic.

3.2.1 Populations Studied

Table 3.2 shows an overview of the papers we reviewed by population. These labels are broad and do not capture the specific details of all study samples. For example, Słupska et al.'s study of migrant domestic workers is included among the general category of "migrants" [444].

Even where the same terminology was used, definitions varied. For example, while many papers were coded as focusing on older adults, definitions of "older" varied. The majority of these studies focused on adults aged at least 60 years; some included adults aged 50-59 years (e.g., [29, 136, 202, 293]). Similar ambiguity exists with the term "young adults." For example, while most papers considered adults younger than 30, several papers included individuals in their 30s (e.g., [159, 357]).[2] These diverse approaches underscore how much these category boundaries depend on researchers' choices.

The category *demographics* refers to papers that investigated sociodemographic characteristics, such as age, gender, or socioeconomic status. Of the 62 papers that examined demographics, 31 focused solely on demographic characteristics without necessarily recruiting a particular population. In line with our search queries, the most common demographic features examined in the papers were age and gender. Demographics studies were also coded as studying specific populations when they expressed directional hypotheses (e.g., older adults would be more vulnerable than younger adults).

Some studies included participants who were not members of the population of interest. For example, several studies focused on individuals with cognitive impairment also interviewed close others responsible for care [199, 323]. Studies focused on security issues among schoolchildren recruited adult stakeholders [153, 477]. Simko et al.'s study of security and privacy issues facing refugees in the United States recruited refugee case workers as well as refugees [442]. This approach can allow researchers to gain broader insight into a population's experience and reduce the burden of research on vulnerable users.

A small number of included papers focused on scammers or artifacts generated by them [300, 313, 407, 445]. As will be discussed in Section 3.3.1, these papers offer insight into how scammers' deliberate targeting choices may help explain observed demographic differences in victimization.

The majority of the sample focused on participants from North America ($n = 53$), Europe ($n = 32$), or Asia ($n = 21$). Of the 53 papers focused on North America, 47 looked at the US and the other 6 at Canada. 7 papers involved users from countries in Sub-Saharan Africa (South Africa, Kenya, Tanzania, Somalia, and Nigeria), and 4 papers focused on MENA: Middle East and Northern Africa (Egypt, Syria, Israel, and Tunisia). Only 4 papers focused on users in Oceania (all in Australia), and only a single study reported results for users in South America. Additionally, 25 studies did not restrict their focus to a particular location or specify where their

[2]Studies with college students often included primarily young adults; however, they were classified as focusing on younger adults only when this was clearly specified as an objective.

Table 3.2: Overview of populations focused on by studies within our sample

Group	**# of Papers**	**Citations**
Demographics	62	[11, 12, 68, 73, 84, 99, 133, 137, 138, 155, 159, 171, 190, 194, 209, 212, 217, 229, 237, 241, 249, 264, 265, 267, 275, 283, 286, 292, 293, 297, 299, 330, 343, 348, 356, 357, 363, 364, 370, 373, 378, 390, 402, 403, 407, 420, 445, 462, 463, 464, 465, 473, 482, 483, 522, 541, 543, 545, 552, 557, 558, 559]
Older Adults	53	[29, 37, 38, 82, 84, 100, 112, 136, 147, 149, 155, 159, 162, 168, 174, 175, 190, 202, 212, 216, 219, 240, 241, 247, 249, 254, 284, 293, 294, 297, 300, 317, 323, 328, 346, 348, 352, 356, 357, 370, 373, 376, 398, 407, 417, 419, 435, 483, 488, 536, 545, 552, 561]
Minors	23	[27, 61, 87, 95, 106, 140, 153, 214, 217, 262, 265, 267, 312, 331, 344, 364, 383, 426, 457, 462, 477, 482, 546]
Young Adults	21	[27, 66, 68, 108, 151, 159, 190, 217, 260, 286, 297, 352, 356, 357, 381, 419, 420, 443, 465, 482, 566]
Users w/ Low-SES	11	[194, 203, 239, 259, 284, 312, 313, 334, 378, 524, 545]
Users w/ Visual Impairments	9	[52, 256, 276, 338, 432, 433, 447, 562, 566]
Users w/ Cognitive Impairments	6	[155, 199, 249, 293, 323, 376]
Migrants	5	[83, 98, 375, 442, 444]
Women	5	[153, 407, 474, 479, 545]
Marginalized Racial & Ethnic Groups	4	[108, 194, 212, 284]
Users w/ an illness	2	[293, 301]
LGBTQ+ People	1	[474]
Men	1	[359]
Other disabilities	1	[27]

Table 3.3: Overview of objectives used by papers within our sample

Study Goal	Single-Goal Papers	All Papers With This Goal
Relational	51	86
Descriptive	41	64
Experimental	8	30

participants were recruited. This result highlights the need for more research focused on the Global South, which is especially underrepresented.

3.2.2 Study Goals

Table 3.3 shows the number of studies with each type of objective within our sample. The majority of the studies we reviewed were classified as *relational*, meaning they were "designed to discover relationships among variables." [148]. Most often, these studies explored differences in measures of scam vulnerability between sociodemographic groups, most commonly age groups. Some studies examined behaviors and attitudes related to scam susceptibility, such as fear of falling victim to scams [209, 254] or awareness of anti-phishing behavior [149]. Additionally, a few relational studies primarily sought to validate scales related to scam victimization [38, 294, 473, 488].

A large number of papers in our sample presented solely *descriptive* studies, meaning they attempted to "provid[e] a snapshot or summary of participants and their opinions or behavior." [148] These were generally relatively small surveys or interview studies that provided rich insight into individuals' experiences. Many descriptive studies did not set out to investigate scams, but online scams emerged as an important safety concern for the population(s) of interest. For example, Camargo et al. conducted an interview study ($n = 12$) with Venezuelan refugees in Brazil, seeking to explore inequalities related to their use of internet-connected technologies throughout the migration process, and especially during the COVID-19 pandemic [83]. They noted that participants frequently mentioned online scams and specifically described WhatsApp-based pyramid schemes. The prevalence of descriptive studies reflects that there is a rich set of qualitative evidence, but a corresponding lack of quantitative data to demonstrate how these findings may generalize.

Relatively few studies in our sample had an *experimental* goal, meaning they "partition[ed] participants into equivalent groups and measure[d] the influence of different experimental manipulations applied to each group" [148]. Most often, experiments were designed to test the effect of educational interventions (see Section 3.4.1), but some studies explored other factors such as the effect of phishing awareness emails [68] or the effect of different types of scam stimuli [73]. As reflected in the prevalence of papers with multiple goals, many experimental studies also investigated relationships between groups. For example, Da Silva et al. tested the impact of the gender of an AI-generated trainer on the efficacy of anti-phishing training within a sample of undergraduate students [363]. They simultaneously investigated the effect of participants' background (i.e., non-technical vs. technical) and gender on phishing detection.

Table 3.4: Overview of methods used by papers within our sample

Study Method	Total	Study Method	Total
Survey	76	Other user study method	9
Interview	39	Not a user study	6
Analysis of existing datasets	29	Unclear	4
Experiment	30	Observation Study	2
Focus group, co-design, or other group activity	15	Log analysis	1

As shown in Table 3.4, *surveys* were the most commonly used method in our sample, and they were used across all three goal types. *Interviews* were the second most common method and were primarily used in a descriptive context. Some studies included individual interviews and *focus groups, co-design, or other group activities*. Studies that included both groups and individuals often used the same protocol for a single interviewee or group of participants. This was particularly common for papers that focused on families where one or more members of a household may have participated in the study [331, 524]. When focus groups were the primary method used, the reason was often to build a sense of community and make participants feel comfortable sharing their experiences, as was the case in the study by Slupska et al., which focused on privacy and security threats to migrant domestic workers [444].

Some studies—particularly relational studies—analyzed existing datasets. The most commonly used data came from several longitudinal studies at Rush University (i.e., Rush Memory and Aging Project, Minority Aging Research Study, and Religious Orders Study). These data sources were used to assess the relationships between demographics and a measure of scam susceptibility [155, 212, 241, 249, 561]. The US FTC [373, 375, 390] and other government institutions from around the world [217, 240, 254, 265, 294, 445] were the data source for several other studies. Others included data from non-profits [265, 536]. Finally, a small number of studies employed publicly available data from social networking sites [61] or dark web forums [300, 301].

Log analysis [557] and *observation* [239, 474] were the least common methods used in the sample. *Other* methods include data donations [357, 378] and simulated attacks [159, 559] that were not categorized as *experiments*. A few studies included in this review did not include any user study component [106, 140, 214], all of which discussed interventions without testing them with users.

3.2.3 Representations of Scam Victimization and Susceptibility

A key issue in experimental and relational studies assessing the likelihood of scam victimization is how to represent vulnerability. We observed three main types of approaches, each with different implications for validity: 1) measuring prior victimization, 2) assessing current susceptibility through behavioral measures, and 3) using proxy measures such as Likert scale questions.

Prior victimization by scams was most often measured via self-report questions, although some studies recruited known victims who had reported the crime (e.g., [488]). The drawbacks

to this approach are clear. Like all self-report measures, these questions rely on participants' subjective interpretation, memory, and honesty. Prior research suggests that self-report measures of scams may lead to underreporting. For example, in a 2003 AARP phone interview study of recent adult victims of telemarketing fraud identified by state and federal agencies, 47% of known lottery scam victims and 73% of known investment scam victims failed to report their victimization [1]. In our sample of papers, DeLiema et al. found that only 48% of known victims of investment fraud self-reported their victimization [137].

The wording of self-report questions may also significantly impact results, especially when asking participants to distinguish between types of scams. Given the lack of uniformity in terminology among researchers (who are presumably scam experts) discussed in the previous subsection, participants (who are less likely to be scam experts) cannot be expected to interpret broad labels consistently. Unfortunately, the vast majority of studies that used self-report measures did not clearly specify the questions they used, making it difficult to evaluate the validity of their approach.

Studies measuring prior victimization also generally lack temporal ordering between the independent and dependent variables, making it impossible to determine whether some potential risk factors preceded the scam or were consequences of the victimization itself. For example, Parti et al. found that adoption of technical protection mechanisms (e.g., antivirus software/firewalls, laptop camera covers, etc.) was actually higher in a group of past victims of online fraud than the non-victim group [370]. The lack of temporal ordering makes it unclear whether victims adopted these protective measures in response to their prior victimization or whether such behaviors were pre-existing. The only self-report study without this limitation was by Koning et al., which used longitudinal data to examine the relationship between various potential risk factors and fraud victimization in a representative Dutch sample ($n = 2,864$) [264]. While this longitudinal design is not sufficient to prove causation, it provides stronger evidence.

Behavioral measures involve participants responding to scenarios, distinguishing between benign and fraudulent stimuli, or engaging in real-world scam simulations. Only a few studies employed real-world scam simulations, which are generally considered the most ecologically valid measures of scam susceptibility. This was limited to email-based scams and was rare outside the computing domain. The frequency of stimuli and time frame of studies varied from sending the participants a single phishing email [73] to participants receiving a message every day for three to four weeks [159, 297, 356, 376]. The rarity of this approach is unsurprising, given the ethical concerns associated with simulating social engineering and the fact that certain scam types (e.g., romance scams) could not be simulated without causing significant harm.

Across all behavioral measures of susceptibility, the stimuli presented may have a large impact on the results. Many studies relied on convenience samples available to the researchers. For example, Dixon et al. stated that they selected emails for their interview study of mobile email phishing detection among young adults "based on real phishing emails received by the research team and examples observed online" [151]. Similarly, Sarno et al. presented participants with examples of authentic emails that were selected from the researchers' spam folders and internet searches [420]. While this approach may provide ecologically valid stimuli, it fails to control for the difficulty or relevance of different stimuli, hindering comparisons across studies.

Relatedly, behavioral measures cannot capture whether someone is actually likely to encounter a scam in practice. As discussed in Section 3.3.1, certain scams are often targeted at

particular groups. A group's potential vulnerability must therefore be considered in terms of the types of scams they are actually likely to encounter.

Finally, some studies employed proxy measures of scam susceptibility, including ad hoc questions and developed scales. Notably, many papers about older adults adopted James et al.'s Susceptibility to Scams Scale to evaluate older adults' vulnerability to scams [241]. This includes all papers that use data from the Rush studies discussed in the previous subsection. The scale is composed of five agree/disagree Likert scale questions, three of which focus exclusively on phone answering behavior (e.g., "I feel I have to answer the phone whenever it rings, even if I do not know who is calling."). The remaining two items focus on general awareness (e.g., "If something sounds too good to be true, it usually is"). Absent evidence that the scale is predictive of past or future victimization, its utility seems limited, especially outside discussions of phone scams.

In sum, among studies that attempted to quantitatively measure group relationships, scam susceptibility remains a poorly defined construct. These issues are not unique to research with at-risk user groups, but they undermine the ability to draw cross-study conclusions, even for groups with extensive literature.

3.2.4 Types of Scams Studied

Although we intended to use the FINRA framework [39] to provide a detailed discussion of the frequency with which different scams were investigated, in practice, the literature lacks the terminological clarity necessary to facilitate this type of analysis. Therefore, we focus on the broader thematic challenges identified across the sampled studies.

The majority of the sampled studies did not focus on scams as the primary object (i.e., descriptive studies in which scams emerged as a concern within a broader investigation), treated scams as a general phenomenon without identifying subtypes of social engineering, or focused solely on phishing. While seemingly informative, in practice, the term "phishing" was used so broadly as to encompass any form of social engineering, including message content that could fall into multiple FINRA categories. For example, multiple studies employed the Phishing Email Suspicion Test (PEST) or a shortened version (S-PEST) to measure participants' ability to distinguish between legitimate and fraudulent emails [202, 219, 376]. The emails included in the PEST cover a range of content, including spoofed account recovery emails, fake employment offers, and investment scams. While such a broad test may be instructive for assessing users' general ability to detect deception, it fails to capture how specific types of deception may exploit contextual risk factors.

When studies differentiated among types of social engineering, they used classification schemes that do not cleanly map onto FINRA categories. For example, some focused on the nature of the social engineering employed, such as a study by DeLiema et al., which classified scams into broad categories (opportunity, threat, consumer purchase, or phishing) based on the type of "offer, promise, or threat used to entrap victims" [138]. Similarly, a study by Oliveira et al. [297, 356] classified phishing emails based on which of Cialdini's six principles of persuasion they contained [113]. Both of these categorizations are informative in the context of the study, but are incompatible with a more precise description.

Finally, some studies in our sample examined broader classes of threats, with internet scams being only one type. Among older adults, scams were sometimes discussed in the context of elder financial abuse [37, 84]. Studies of "sextortion"—or extortion premised on the threat of the release of intimate imagery—generally did not distinguish between financial sextortion and sextortion committed for other purposes (e.g., obtaining child sexual abuse material) [153, 265]. Outside of computing, researchers typically discussed internet-mediated scams alongside other forms of fraud, including fraud perpetuated through analog means. For example, Fenge and Lee's interview study of older adult scam victims included those targeted by scams perpetrated both online and via the postal system [175]. These studies can provide useful insight, but the results are necessarily less granular.

Overall, the lack of uniformity in scam classification underscores the need for a rigorous, systematic framework for discussing scam types in academic literature. This is especially important in the context of the present study, given that individuals' vulnerability to different scams is likely to vary by content.

3.3 RQ2: Contextual Risk Factors and Scams

In this section, we describe the risk factors that contribute to scam victimization or augment its harms, extending Warford et al.'s framework for at-risk user research [535]. We find that these vulnerabilities are heavily exploited by fraudsters through deliberate demographic *targeting*. Unlike the remaining categories discussed in this section, targeting is not itself a risk factor, but rather a mechanism by which scammers activate and exploit risk factors. *Personal circumstances* emerge as highly salient, particularly underserved accessibility needs and resource constraints. Additionally, social isolation and loneliness (a factor which was not discussed by Warford et al. [535]) was frequently hypothesized to augment vulnerability by pushing users into online spaces and increasing susceptibility to schemes like romance scams. *Societal factors*, including unfamiliarity with institutional norms and the precarious legal positions of certain groups, both heighten risk and introduce severe stigma and reporting barriers. Furthermore, *relationship factors*, such as overreliance on third parties and ties to attackers, may amplify exposure to fraud. Finally, we discuss the role of dispositional traits and routine online behaviors, focusing on the latter as a mediator between certain contextual risk factors and scam exposure.

3.3.1 The Overarching Role of Targeting

Deliberate *targeting* of people experiencing certain risk factors emerged as a feature of many scams. This manifests in several ways. Reflecting the most diffuse form of targeting, many scams are designed to appeal to people experiencing specific risk factors. For example, as will be discussed in greater detail in subsequent subsections, many scams prey on resource constraints by purporting to offer high-paying, low-skilled jobs or other financial incentives [98, 259, 313, 357, 378]. Sellers of fake or adulterated pharmaceuticals exploit the inability of people with addiction or other illnesses to access safe forms of drugs [301]. Immigration scams target migrants' legal vulnerability by purporting to offer visas or other favorable action [375].

In addition to tailoring message content, scammers may distribute it to different demographics at varying rates. Yang et al. analyzed call records from September 2016 ($n = 698,811,827$) from a Shanghai-based telephone company with 54 million users, finding that reported fraudsters were more likely to call younger adults and more active mobile phone users [557]. At a much smaller scale, Oliveira et al. analyzed spam emails received by a sample of older and younger adults ($n = 67$), finding that older users received health-related spam significantly more frequently, while younger users were targeted with occupation-related lures [357]. Less directly, some large-scale survey studies found differences in exposure to specific scam types through self-report measures [98, 138, 264]. For example, Koning et al.'s longitudinal study of factors affecting scam susceptibility asked participants to report whether they had been victimized by specific types of fraud or if "they had experienced a fraud attempt to which they did not fall victim" [264]. They found several differences in self-reported exposure across demographic groups (e.g., men were much more likely to report exposure to multiple fraud subtypes).

The mechanisms underlying differences in exposure to scams are unclear and may not always reflect deliberate targeting. Users' behavior may lead them to encounter fraud more often, unintentionally. However, studies of fraudsters reveal that some scammers do intentionally target some demographics. Logie and Das conducted a thematic analysis of 818 posts and 1,365 comments from the Dread darknet forum [300], focusing on why and how fraudsters target older adults for victimization. They found that, across a diverse range of crime types, perpetrators actively considered victims' levels of vulnerability as motivation to target certain demographic groups. They also observed criminals seeking various forms of data (e.g., lists of prior scam victims, medical records, business customer lists) to enable attacks against older adults and other specific groups. More narrowly focused, Soares and Lazarus analyzed 50 sets of case files related to romance fraud convictions from Nigeria's Economic and Financial Crimes Commission (EFCC) [445]. They found that the (exclusively male) perpetrators showed a pronounced demographic preference in whom they targeted, with 70% of fraudsters primarily targeting female victims and 56% of the cases involving victims in the United States. The perpetrators most often identified victims through Facebook or Instagram (72% of cases), and often adopted personas designed to increase the victims' trust. For example, several perpetrators purported to hold high-status jobs, such as military officers (6 cases), engineers (1 case), or doctors (1 case). The disproportionate targeting of women is supported by a study by Whitty, which found that middle-aged, well-educated women were the most likely demographic group to be victimized by romance scams in a sample of 11,780 UK-based participants [543]. Koning et al. found no difference in reported rates of victimization by "dating scams" between men and women in a representative Dutch sample, but found men were more likely to report exposure to a dating scam [264].

Uniquely, Robinson and Edwards directly investigated demographic targeting by conducting a "randomised controlled scam-baiting experiment." They automatically communicated with scammers over email using one of four predefined victim personas. Across conversations with 296 scammers, the "kind old woman" persona was the most likely to elicit a response from scammers and had the longest average conversation length [407]. While the study is methodologically innovative, the implications of its results are unclear, as the use of highly tailored, stereotypical personas and rigid scripts appears to undermine its ecological validity, as evidenced by the frequent scammer disengagement across all personas.

The effect of targeting within experimental and relational studies was most often tested in the context of "spear phishing," meaning email phishing messages tailored to a particular individual or organization. This term had a wide range of interpretations, ranging from simply including the name of the recipient [73] to an email tailored specifically with details from the recipient's social media profile [68].

In sum, while intentional targeting is evident, its role in explaining differences in victimization or susceptibility is underexplored. In particular, the role of diffuse targeting, which focuses on broad user groups or specific risk factors, remains unclear. Moreover, although there is some descriptive evidence on the mechanisms of targeting, the extent to which scammers employ different practices remains unclear.

3.3.2 Personal Circumstances

Contextual risk factors related to personal circumstances, especially *underserved accessibility needs* and *resource and time constraints*, emerged as the most salient for online scams. *Social isolation and loneliness* also emerged as an important contextual risk factor in the scam context, which Warford et al. [535] did not identify. These factors primarily emerged through qualitative evidence, with more limited support from quantitative data. Their role in explaining demographic differences in victimization rates is unclear.

Underserved Accessibility Needs

Our analysis identified issues across three types of accessibility: cognitive, visual, and language.

Issues with *cognitive accessibility* seem to significantly impact susceptibility to scam attempts. Several relational studies with older adults examined either global cognition or specific cognitive domains (e.g., short-term memory) as predictors of susceptibility to scams. They most often found that lower cognitive function was associated with higher scam susceptibility or prior scam victimization [38, 155, 212, 241, 247, 249, 293, 376, 561], although some studies found either no effect for cognition [136] or only partial support for the role of cognitive functioning in scam susceptibility [159, 190]. Uniquely among these papers, Pehlivanoglu et al. tested the relationship between possession of a gene associated with Alzheimer's disease (APOE4) and phishing susceptibility among a sample of adults ($n = 182$). They generally found that older adults were significantly more likely to interact with at least one of the 60 simulated phishing messages over a 30-day period, with this effect especially pronounced among older adults with the APOE4 gene and lower working memory [376].

The impact of cognitive impairment on scam vulnerability also emerged in qualitative studies [199, 323, 524]. For example, Gould et al. interviewed seven internet scam victims with acquired brain injury (ABI) and their close others ($n = 6$) to gain insight into their experiences [199]. Both groups reported that impaired cognition hindered people with ABI's ability to avoid scams, making it challenging for them to understand the threat of scams or identify cues to deception. Similarly, Mentis et al. interviewed older adults with mild cognitive impairment (MCI) and their caregivers to investigate how they cope with security and privacy threats [323]. No participants reported being seriously harmed by scams, but several interviewees discussed

minor security incidents that indicated a reduced ability to detect deception. One caregiver discussed his wife repeatedly attempting to buy scam medical cures, while another participant with MCI discussed their perception that they were no longer able to detect fraudulent emails.

It is important to note that, even if cognitive impairment does make *some* older adults more vulnerable to scams, it is not clear that this translates to older adults generally being more likely to be scammed. The results of studies considering age in adult populations are mixed. While some studies found that older adults were more susceptible to scams or more likely to report prior victimization [99, 137, 201, 283, 356, 376, 402], most studies found null/mixed results with respect to age [11, 73, 138, 159, 190, 209, 217, 237, 264, 297, 299, 348, 403, 419, 420, 463, 545, 558]. Some studies reported that younger adults were more vulnerable [133, 171, 330, 370, 473, 483]. The heterogeneity of methods across these studies (discussed in Section 3.2) means that studies are not always directly comparable, and both camps include work of varying quality. It may be, as suggested by Toomey [483], that performance differences in lab contexts do not translate to real-world vulnerability due to behavioral or other contextual factors. In support of this, Koning et al. found that older adults were less likely to report exposure to various specific scam types [264]. Ultimately, despite their over-representation in government reports [496, 501], there remains insufficient evidence to conclude that older adults are more vulnerable to scams in general.

There is also a growing body of research investigating *visual accessibility* issues that may affect phishing vulnerability [276, 338, 432, 433, 447, 562]. These papers focus on the experiences of people with visual impairments (PVI), particularly those who primarily use audio interfaces to access devices. For example, Napoli et al. conducted an interview study ($n = 14$) with PVI, which included a phishing simulation task [338]. Participants were assigned to look at a website and log in/purchase if they believed it was legitimate. None of the participants was able to identify a spoofed charity website, despite having high confidence that they correctly identified the website as legitimate. The study concluded that this stemmed from a combination of issues, including the fact that many elements appear identical when read by a screen reader and that the misspelled URL for the phishing website was read the same way by a screen reader as the real website would be. Khan et al. similarly conducted a task-based interview study where PVI from Pakistan ($n = 11$) processed 10 emails (4 fraudulent and 6 benign) using a screen reader of their choice [256]. They noted the same type of screen reader issues where grammar, formatting, and spelling errors were inaudible. Their participants also exhibited behavioral patterns common among sighted users, like not verifying the destination of embedded links or checking the domain associated with email addresses.

Other studies have focused on the *visual accessibility* of security indicators, especially warnings. Lau and Peterson conducted a task-based interview study to investigate how low-vision users interact with interstitial browser security warnings for phishing, malware, and SSL errors [276]. The study surfaced issues common to warning studies with sighted users (e.g., habituation) as well as issues faced especially by screen reader users, such as an inability to use visual cues and confusion around inconsistent layouts. Yu et al. conducted a series of studies with PVI ($n = 41$) to improve email phishing warnings in the desktop version of Gmail. They found that the extant warnings were easily missed by participants due to design elements (e.g., warning position, HTML tags used, inconsistent implementation across different views, etc.) that were ill-suited to screen readers' email browsing habits [562]. Sharevski and Zeidieh conducted

a study ($n = 20$) with a broader range of email clients and warning types, including phishing and spam warnings from the desktop and mobile versions of Gmail and Outlook. They similarly found that the noticeability of email client warnings was low for screen reader users [432]. Finally, Sonowal et al. conducted a usability evaluation of anti-phishing browser extensions, noting potential accessibility issues (e.g., lack of keyboard interface, reliance on color indicators, etc.) that may make these tools less useful for screen reader users [447].

Only one study investigated the potential for audio-mediated scams targeting PVI. Sharevski et al. conducted a task-based interview study ($n = 16$) evaluating the ability of blind and low-vision people to classify 16 recordings as deepfakes or not [433]. Despite their extensive experience with computer-generated audio, participants performed only marginally better than chance. The paper argues that this raises the potential for voice phishing or other forms of AI-mediated deception, especially given PVI's reliance on speech for communication.

While these accessibility issues are important to address, it is unclear whether they make PVI more vulnerable to scams in practice. None of the studies within our sample provided broad data on PVI's real-world experience with phishing or security indicators. Moreover, visual inaccessibility may occasionally be a protective factor. For example, one participant in the study by Sharevski and Zeidieh reported detecting a phishing attempt because the email consisted solely of an image (which the participant could not view) and a malicious attachment.

Finally, *language accessibility* was also identified as an issue for people who were not fluent speakers of the dominant language in their country of residence, especially recent migrants. Simko et al. identified unfamiliarity with English as an issue affecting a variety of security and privacy tasks, including scam identification [442]. Broadhurst et al. conducted a phishing simulation study with students at the Australian National University ($n = 138$) [68]. A post-hoc analysis revealed that international students were significantly more likely to interact with phishing messages than domestic students, which the authors of the study hypothesize may relate to language barriers or differing experiences with scams in their country of origin, although further research is necessary to determine how this result may generalize and what underlying factors are actually responsible for the observed differences. The only direct evidence of language accessibility affecting scam susceptibility came from Chen et al., who found that a measure of migration stress, which included elements related to language barriers, was associated with self-reported prior scam victimization among internal migrants in China [98].

As with visual inaccessibility, the true role of language in scam vulnerability is likely complex and not unidirectional. Fedele et al. conducted an analysis of administrative data on phishing attacks targeting customers ($n = 147,751$) of an Italian bank, which serves people in the German-speaking region of South Tyrol (76% of customers used Italian while 24% used German). They found that German-speaking customers were significantly less likely to be victims than Italian-speaking customers. The authors attributed this to the fact that phishing emails impersonating the bank were often in Italian, which was incongruent with their expectation of German communications from the bank [171].

Resource and Time Constraints

Resource constraints, especially a lack of financial means, emerged as an important factor that scammers exploit. Many scams offer fake prizes or other financial opportunities (e.g., high-paid

work from home positions), which may especially target people with lower socioeconomic status (low-SES) [98, 313, 357, 378]. Multiple qualitative studies discussed low-income participants falling for these scams, even in some cases where they recognized that the offers may be "too good to be true" [83, 194, 259, 284, 524].

Due to limited financial resources, losses from scams also have a greater impact on low-income users. For example, in Lee et al.'s interview study with participants in the Money Smart for Older Adults Program, one participant reported that they had their car repossessed after falling victim to a job scam and being unable to make payments [284]. Low-income merchants interviewed by Jain et al. often depended on the previous day's revenue to purchase goods for the next day, meaning that losses of money due to theft or scams could threaten their livelihood or drive them to take out predatory loans [239].

Resource constraints also manifested as lower levels of technical expertise, making it more difficult for users to avoid scams or recover from scam victimization. Vitak et al.'s study of security and privacy challenges faced by US-based, economically disadvantaged people reported that a lack of technical literacy left participants unsure how to prevent scam victimization and sometimes even led to behavior that magnified their risk of harm (e.g., foregoing banking for prepaid debit cards) [524]. Jain et al. similarly reported that low-income merchants' lack of familiarity with digital payment apps, as well as their reliance on company agents for onboarding, made them more susceptible to in-person and digital fraud [239]. Munyendo et al. conducted interviews with customers ($n = 22$) and managers ($n = 14$) of Kenyan cybercafés to investigate their security and privacy needs [334]. They found that cybercafés play a critical role in enabling customers to access printing and government services in a country where most households lack access to computers, but that customers' limited computing experience leaves them at greater risk. Managers reported needing to warn customers about scams or other "suspicious links" and discussed the risk of fraud arising from users leaving private information on shared computers.

While these qualitative results might suggest that people of lower socioeconomic status are more vulnerable to scams, the quantitative evidence on this relationship with the sampled papers is mixed. Regarding measures of income or general financial health, some evidence suggests greater susceptibility or past victimization among those with lower incomes [38, 98, 212, 217, 241, 293, 552], while other studies find null/mixed results [136, 137, 138, 203, 264, 403, 483] or even a positive correlation [390, 402]. Results regarding educational attainment are similarly conflicting [38, 84, 98, 155, 212, 237, 241, 249, 293, 403, 483, 543, 545, 553]. As with the contradictory results regarding age, these could stem from methodological differences or various confounding factors. At a basic level, a person with more financial resources is objectively a more attractive target for scammers, regardless of whether they are more or less susceptible to fraud. More access to technology may lead to greater technical ability, but also more opportunities to encounter online scams. As suggested by multiple large-scale studies from DeLiema et al. [136, 138], it may be that both socioeconomically disadvantaged and advantaged people fall victim to scams, but the scams they fall for are different.

Only two, relatively small studies investigated the effect of *time constraints* within a specific population. Morgan et al. compared older adults' ($n = 29$) ability to detect fraudulent pop-up warnings while conducting a time-constrained memory recall task vs. in a questionnaire, finding that participants were much more likely to accept warnings during the cognitive task [328]. Sarno et al. had younger and older participants ($n = 30$) perform a phishing identification task

under time pressure and found no difference between the two groups [419]. The lack of research focused on time constraints does not suggest that they are irrelevant to scam victimization. Indeed, results from phishing studies suggest that factors like cognitive load and distraction have a large effect on the ability to avoid phishing [142, 570]. This presents an opportunity for further research into how time constraints may affect certain populations differently.

Social Isolation and Loneliness

Throughout our coding, *social isolation and loneliness* emerged as a hypothesized contextual risk factor increasing scam vulnerability, which was not included in Warford et al. [535]'s framework. Isolation and loneliness may drive users to seek online interaction, which, in turn, exposes them to scams. This was especially noted to stem from underserved accessibility needs. For example, Gould et al. found that participants with acquired brain injury (ABI) relied on online dating services to meet others, due to the loss of pre-brain-injury friendships, a lack of employment or hobbies, and difficulties forming romantic relationships [199]. When combined with a reduced ability to recognize cues to deception, this reportedly made people with ABI especially vulnerable to romance scams.

Isolation was also discussed by multiple studies with older adults. Fenge and Lee's interview study with older adult scam victims ($n = 12$) revealed that loneliness stemming from bereavement or disability reportedly led participants to be more susceptible to scams [175]. Two participants reported that a desire for social interaction drove them to respond to fraudulent communications, while another suggested that a lack of close others to confide in left them without guidance that may have protected them from harm. Similarly, a recently widowed participant in Havers et al.'s study of barriers to scam reporting ($n = 33$) described being emotionally dependent on their scammer [216].

Some quantitative studies investigating the role of loneliness, social isolation, or related variables (e.g., depressive symptoms) found that these were associated with scam susceptibility or prior victimization [98, 138, 212, 241, 293, 483, 488, 543, 545], although studies also found null results [38, 84, 264, 561] or mixed results [136, 370, 552]. In the few cases where particular types of scams were studied, loneliness seems to be more associated with particular types of scams, especially romance scams [136, 138, 543, 545].

3.3.3 Societal Factors

Societal factors were less prominent in our sample and sometimes difficult to disentangle from one another. Nevertheless, *social norms* emerged as a factor that may increase users' vulnerability and undermine recovery after scams. *Legal or political* factors and *marginalization* were also discussed. These relations primarily emerged from qualitative studies, raising questions about their broader explanatory power.

Social Norms

Unfamiliarity with *social norms*, particularly those surrounding bureaucratic processes, may make some groups more vulnerable to scams. Participants in Simko et al.'s study of refugee

security and privacy noted that recent arrivals did not know how the US government or other institutions typically operated, so phone calls about arrest warrants or overdue tax bills did not necessarily raise red flags [442]. In their longitudinal study of fraud victimization in the Netherlands, Koning et al. found that participants with a "non-Western immigrant background" were more likely to be exposed to and victimized by fraud, although they did not specifically link this to any contextual risk factor [264].

Cultural norms may also make individuals more amenable to specific forms of persuasion. Bullee et al. conducted a phishing simulation study at a Dutch University ($n = 593$), which considered age, gender, email personalization, length of employment, and power distance index (PDI) as variables affecting susceptibility [73]. PDI is a country-level measure of the extent to which less powerful members of a society accept that power is unequally held. The study found that those from countries with higher PDI were significantly more likely to fall for phishing emails, which the paper attributes to a greater tendency to obey authority in high-PDI societies. Since other confounding variables related to nationality were not investigated, it is unclear to what extent PDI explains the effect.

Social norms introduce barriers to reporting and remediating scams. For example, two studies focused on the safety experiences of teenagers found that social norms around female sexuality were a barrier to reporting victimization by sextortion. Martin et al. conducted focus groups with Tunisian teenagers ($n = 113$) in order to investigate the benefits and risks they experience online [312]. Sextortion perpetrated by strangers over social media services emerged as a common safety threat, especially targeting teenage girls. One victim described their mother blaming them for sharing photos, and other victims discussed a fear that even government authorities would view them as responsible. Dolev-Cohen et al. [153] conducted an interview study with middle and high school counselors in Israel who treated girls victimized by sextortion. They observed that some counselors, particularly in predominantly Arab schools, did not involve the parents of victims out of fear of reprisals stemming from violations of religious norms around sexuality. A study from the Czech Republic that analyzed sextortion case reports and survey data from children ($11 - 17$ years old) similarly found that victims were predominantly girls who often did not report the crime to police, although they did not connect this to social norms [265].

It is important to note that sextortion also significantly affects men and boys, although their experience is likely qualitatively different from that of women. Financial sextortion seems to especially affect men. O'Malley and Smith conducted a survey of male victims of financial sextortion who posted to online support forums ($n = 215$). They found that the majority of victims experienced suicidal ideation, reflecting the significant emotional harm of this type of fraud [359].

More broadly, scam victimization is highly stigmatized. Victims are often considered at fault and maligned as stupid or gullible [29, 199, 216]. This narrative of victimization may be internalized by victims [29, 175, 216], dissuading them from reporting the crime or seeking assistance.

Legal Factors

Scammers may take advantage of the precarious legal position of certain groups. For example, scammers may threaten migrants with deportation or offer them false promises of work permits,

naturalization, or other favorable outcomes [375]. When combined with migrants' unfamiliarity with bureaucratic norms, discussed in the previous subsection, this legal precarity can heighten vulnerability: a fraudulent demand tied to immigration status may be difficult to distinguish from legitimate government action for those unfamiliar with how such processes typically unfold.

Migrants may also not want to report scam victimization out of fear that contact with the authorities will lead to their removal [444, 524]. Quantifying this relationship, Pedroza found that the number of immigration-related fraud reports to the FTC is positively correlated with a more welcoming environment for immigrants in a particular county as measured by the presence of social safety nets, immigration attorneys, legal aid services, and better language access [375]. This analysis suggests that immigrants in environments with more community and legal support are more willing to report scams.

Marginalization

Marginalization was difficult to disentangle from other societal risk factors. For example, immigrants are often marginalized; however, determining where the effects of legal discrimination and social discrimination begin can be challenging. Likewise, the social norms related to female sexuality may be fairly interpreted as the result of marginalization.

Only one paper in our sample focused on marginalization in relation to scams. Thumboo and Mukherjee conducted a feminist analysis of romance scams based on six publicly reported case studies of incidents targeting unmarried women in India and South Africa. They argued that romance scams "exemplify the intersection of entrenched patriarchy and misogyny" and "exploit societal pressures on individuals to seek validation through romantic relationships, making [women] more vulnerable to manipulation by perpetrators who exhibit controlling behaviors" [479]. As discussed previously (Section 3.3.1), the evidence for disproportionate targeting of women by romance scams is mixed; however, almost all available evidence focuses on North American and European perspectives.

Marginalization was discussed less directly as a factor driving loneliness or isolation, which, in turn, may increase susceptibility to scams. As already discussed, participants in Gould et al.'s study of people with ABI reported loneliness and social isolation stemming from their marginalization [199]. Similarly, a participant in Tao's ethnographic exploration of online dating in the Chinese lesbian or "lala" community suggested that loneliness made queer women more susceptible to scams perpetuated in online dating spaces [474].

3.3.4 Relationship Factors

As perpetrators of many types of scams are strangers, we did not expect relationship factors (i.e., *relationship with the attacker*, *reliance on a third party*, and *access to at-risk users*) to be important in this context. Nevertheless, these were discussed as relevant factors in several cases. As with societal factors, these were primarily evidenced through small, qualitative studies.

Multiple papers highlighted users' *reliance on a third party* as a factor that may amplify risk. For example, Simko et al. suggested that, in addition to an unfamiliarity with US institutions and language barriers, refugees' reliance on community organizations and government institutions may contribute to their susceptibility to false promises of government aid [442]. Jain et al.

discussed how the low-income merchants they interviewed depended on the agents of digital payment processors to onboard them onto the applications [239]. The agents failed to adequately instruct them on using these apps, leaving the merchants unprepared to recognize or prevent in-person and digital scams. Munyendo et al. highlighted similar issues: novice cybercafe customers rely on managers for technical support, including security-related tasks such as creating and storing passwords [334]. While some managers provided advice consistent with best practices, others adopted strategies that may place users at risk, such as generating passwords from national ID numbers.

Access to other at-risk users was discussed as a factor for migrant domestic workers in the UK [444], as domestic workers often lived together and depended on one another for safety. While this conferred benefits, such as allowing workers to learn from one another, it also expanded the potential impact of digital threats.

Finally, Logie and Das highlighted how users' *relationships with attackers* can place them at risk for both traditional fraud and online scams. They observed fraudsters on Dread, a dark net forum, discussing how they leveraged their access to users they perceived as vulnerable to perpetrate fraud [300]. Examples ranged from criminals exploiting specific individuals—as in a case where a fraudster described stealing private information from an older man who came to his store seeking technical assistance—to criminals exfiltrating data from large numbers of people—as in the case of a user who sought advice on how to exploit access to medical records at a privately owned clinic. While these circumstantial, transient connections are less direct than the close personal relationships central to Warford et al.'s original systematization [535], they demonstrate that brief, transactional power imbalances can be dangerous.

3.3.5 Dispositional Traits and Behavioral Factors

Many relational studies investigated dispositional characteristics, such as personality traits, alongside demographic or contextual factors [11, 98, 99, 138, 159, 190, 209, 217, 229, 241, 247, 260, 262, 264, 330, 348, 370, 463, 464, 465, 473, 488, 522, 541, 543, 545, 552]. For example, Whitty analyzed the relationship between demographic features (age, gender, and education), personality traits (impulsivity, locus of control, and neuroticism), and self-reported scam victimization in a large sample of UK participants ($n = 11,780$) [543, 545]. Scam victimization was weakly correlated with higher impulsivity, lower external locus of control, higher neuroticism, being a man, and being more highly educated.

Dispositional traits were generally measured through self-report inventories, with only two cases where studies attempted to measure traits directly (i.e., economic games [99, 190]). This raises questions about the construct validity of these measures. Moreover, while some demographic variables may correlate with certain personality traits (e.g., adolescents tend to be more impulsive than adults [278, pp. 528 – 530]), in most cases, such traits do not seem to provide actionable insight.

Routine behavior was another frequently investigated factor in relation to scam vulnerability, and it seems to have greater explanatory power. Certain online activities almost certainly increase one's exposure to fraud, which may, in turn, lead to more incidents. Many studies demonstrated correlates between self-reported behavior and scam susceptibility [12, 98, 137, 209, 260, 264, 267, 330, 370, 402, 403, 488]. For example, DeLiema et al. investigated correlates of investment

scam victimization with a sample of known investment scam victims ($n = 214$) and general investors ($n = 813$), finding that members of the victim group were much more likely to report that they respond to digital and physical investment solicitations [137]. Beyond passive exposure, these behaviors also generate distinct digital artifacts that may facilitate the type of deliberate targeting discussed in Section 3.3.1.

Behavioral effects were often discussed through the lens of Routine Activity Theory (RAT), which was described in detail in Section 2.1. For example, Parti conducted a survey ($n = 2,589$) which measured the association between demographic variables, behavioral measures, and victimization by one of six types of online fraud in the last year [370]. The behavioral measures included values representing target suitability (e.g., time spent on computers each week, working in a computer-related occupation, general familiarity with computers, etc.) and capable guardianship (e.g., using antivirus software, asking for help when encountering scams, etc.). The study found that those who used computers more often and were more familiar with computers were more likely to have been victimized by scams, with younger adults generally being more exposed to scams. As discussed in Section 3.2.3, the study found, counterintuitively, that fraud-victim groups were more likely to adopt technical protection mechanisms. While this may seem to contradict RAT, it also may result from participants adopting protection measures in response to fraud.

Some behaviors may result from contextual risk factors that serve as mediators. For example, we might hypothesize that *loneliness and social isolation* drive some users to engage in more online activities, thereby increasing their exposure to internet scams. Likewise, a resource-constrained person may spend more time online seeking ways to make money, leading to more encounters with employment scams [259].

3.4 RQ3: Protection Mechanisms for At-Risk Groups

In this section, we describe the types of user-centric protective measures discussed in our sample. While law-enforcement efforts to stop scammers and technical tools to prevent scam messages from reaching potential victims are also important, our focus is on user-centric interventions. In general, anti-scam interventions were uncommon in our sample. The primary approach to protecting users against scams was targeted education. Other forms of intervention, such as interface design changes or safety tools, were evaluated only infrequently. All user-study evaluations of non-educational interventions within our sample focused on people with visual impairments (PVI). The evaluation of non-educational, user-centric interventions for other groups represents a significant opportunity for future work.

3.4.1 Educational Interventions

Many of the papers reviewed incorporated some form of educational intervention targeting online scams, with email phishing receiving the most attention [87, 99, 275, 292, 344, 363, 364, 383, 426, 443, 457]. A number of other studies took a broader approach, addressing scam awareness or cybersecurity education more generally [82, 112, 240, 317, 381, 435, 462].

A prominent theme within papers focused on education was the use of games. For example, Schoebel et al. presented a story-based browser game designed to teach children how to identify phishing URLs [426]. In a within-subjects experiment with 112 German children (grades 6 to 9), the game was shown to improve students' phishing detection performance, although the effect was no longer significant after 4 weeks. Chiou et al. presented, but did not evaluate, an augmented reality game meant to teach children about phishing [106]. Chung and Yeung developed and tested an anti-scam board game with a sample of older adults ($n = 109$) [112]. After experiencing the board game during a workshop, participants were better able to respond to scam scenarios, reported higher self-efficacy in scam avoidance, and had lower scores on selected items from James et al.'s Susceptibility to Scams Scale [241]. Finally, Panga et al. presented an educational Android game designed to teach teens about phishing. They performed a small-scale evaluation with secondary school students in Tanzania ($n = 30$), comparing responses to security awareness questions before and after using the app or reading notes about phishing (the control condition). Students achieved higher scores in the game condition [364].

Studies examining scams within or across groups frequently recommended targeted or tailored training, though the plausibility of these recommendations varied. For instance, proposing educational initiatives designed to address the needs of individuals with cognitive impairments (e.g., see [38, 199]) appears well justified given differences in information-processing abilities that may require tailored approaches. Similarly, anti-phishing training tailored to PVI who use screen readers (see [338]) could be useful, considering the existing focus on visual phishing identification. In contrast, proposals for gender specific training (e.g., see [285, 286, 299, 465, 522]) or training targeted at particular personality traits (e.g., see [11, 99]) lack clear articulation or justification.

Despite numerous recommendations to train older adults (e.g., [37, 174, 175, 216, 293, 328, 552, 561]), only a few papers actually evaluated anti-scam educational materials for older adults [82, 112, 240, 317, 435]. In addition to the board game presented by Chung and Yeung [112], McDonald and Mentis performed a qualitative analysis of a sample of educational materials related to security and privacy from the AARP posted online between 2004 and 2020, the majority of which focus on scams [317]. They characterize the body of articles as an "anthology of myths" about specific scams that fail to provide older adults with the skills needed to avoid harm. Sheil et al. conducted a series of workshops with Irish older adults ($n = 58$) to develop tailored cybersecurity advice, which identified scam recognition as the demographic's top concern [435]. Subsequent focus groups ($n = 31$) evaluating their resulting advice leaflets mostly highlighted participants' existing, informal coping strategies.

The most compelling studies of anti-scam education focused on real-world deployments, rather than considering one-off interventions. Button et al. investigated the distribution of scam awareness materials by a UK charity that works with older adults to improve their social connections and reduce loneliness [82]. To evaluate the campaign, the authors conducted 23 qualitative interviews (18 clients, 5 staff) and two sequential postal surveys, which yielded 1,177 baseline and 820 post-implementation responses from clients. The intervention trained volunteers through bespoke online modules to weave anti-fraud advice into social activities such as tea parties, while directly targeting clients through educational articles in a thrice-yearly magazine. However, the campaign faced severe reach limitations: only 27% of respondents recalled receiving the advice, due to low volunteer training uptake and a general reluctance among volunteers to introduce scam

topics in recreational spaces. In tension with this, survey respondents expressed a preference for receiving this type of informal advice rather than from websites or structured training.

Also focusing on a real-world implementation, Jakobsson and Gerell tested the efficacy of a postcard-based informational campaign to inform Swedish adults aged 70 and older about telephone scams [240]. The results were inconclusive, with no significant reduction in fraud reports from this demographic compared to adults who did not receive the postcard. The results of these two studies underscore the importance of considering how to deliver consumer education at scale. Even if educational material is shown to be effective in one-off evaluations, to be impactful, it must be widely distributed in a format that actively engages the target audience.

Other than older adults, the only targeted educational interventions in our sample were designed for minors [87, 106, 275, 364, 381, 383, 426, 457, 462], younger adults [66, 443], or the general population [11, 99, 292, 363]. This lack of coverage suggests room for future research evaluating tailored scam education for other groups, at least where a plausible need for specialized education exists.

3.4.2 Other User-Centric Anti-scam Interventions

Other user-centric interventions were less commonly investigated. Those that we observed primarily focused on two populations: minors and PVI. In the first category, Demirol et al. and Harfath et al. presented parental monitoring systems intended to protect children from phishing [140, 214]. These papers do not include evaluations with children; rather, they present untested systems.

Four papers presented or evaluated anti-scam interventions for PVI. Two papers focused on email phishing warnings. Yu et al. built on their results regarding the (in)accessibility of Gmail phishing warnings (see Section 3.3.2) by creating a prototype warning with an audio alert, a blocking design that overlayed the email, and the ability to use a keyboard shortcut to access the warning text. This warning was significantly more likely to be noticed by participants in a simulated email review task, suggesting that these design patterns are more accessible. Zhang et al. investigated the effect of two types of audio alerts for phishing emails. While presented as an intervention for PVI, the experiment recruited people with normal or corrected-to-normal vision ($n = 113$) who almost certainly experience such alerts differently from screen reader users. These and other methodological flaws (e.g., simulating a 40% false-positive rate) leave the practical implications of their findings unclear [566].

Two other studies focused on non-warning interventions to help detect phishing. As discussed previously, Sonowal et al. conducted an expert evaluation of anti-phishing browser extensions, focusing on the needs of screen reader users. Perhaps unsurprisingly, given that they reviewed freely available browser extensions and toolbars, none of the tools were found to be accessible (e.g., none featured audio alerts, and only two supported keyboard navigation). Based on the body of literature on visual accessibility barriers to phishing detection, Bohlender et al. developed SMILE4VIP [52], an email client add-on that adapts and combines TORPEDO [525] and SMILE [40]. While TORPEDO originally presented per-link risk assessments using visual tooltips, which are inaccessible to screen reader users, SMILE4VIP determines an email's overall risk level based on its highest-risk link and appends this assessment directly to the subject line, pairing it with auditory alerts to differentiate risk levels. Furthermore, the tool applies SMILE

modifications to hyperlinks by replacing text anchors with the domain name and top-level domain while actively resolving shortened or redirected URLs. The authors conducted an initial online study with four visually impaired volunteers, who provided qualitative feedback and suggested future features, such as per-link information options.

DiMario et al. developed the only non-educational intervention positioned to help older adults, a rubric for using ChatGPT to identify phishing emails [147]. The rubric was technically evaluated and performed poorly on novel datasets. Moreover, beyond the paper's positioning, this intervention does not appear to be specifically designed for older adults.

Many papers proposed specific design changes for certain groups, especially to address underserved accessibility needs. For example, as discussed previously in Section 3.3.2, Lau and Peterson observed that the inconsistent layout of browser security warnings, especially the lack of an option to proceed on severe warnings, made it difficult for screen reader users to navigate [276]. While these design choices are intended to make it harder for users to dismiss a warning without thinking, they caused PVI to waste time looking for buttons that weren't there. The paper suggested solutions to alert screen reader users to missing buttons, such as text indicating the missing button or a disabled button element on the page. More generally, they recommended that warning designers consider the effect that design choices intended to reduce habituation may have on people using assistive technology. Sonowal et al. recommended that anti-phishing extensions adopt keyboard shortcuts and provide auditory signals as an alternative to visual cues. Based on the strategies used by couples with mild cognitive impairment (MCI), Mentis et al. suggested implementing mechanisms to allow negotiation of access control within online services, such as requiring caregiver approval for online bank transfers [323]. The implementation and evaluation of these types of recommendations represent opportunities for future work.

Structural or procedural recommendations were also discussed, especially outside of the computing literature. For example, some papers recommended that social workers or physicians working with older adults be aware of scams, possibly screening for signs of victimization [37, 175]. Papers that focused on sextortion among teenagers emphasized the need for adult stakeholders, like school counselors or law enforcement personnel, to be aware of sextortion and be prepared to deal with it in an empathetic manner [153, 312]. Studies on migrants generally recommended improving community support structures, especially with respect to legal support [98, 216, 375, 444]. These types of proposals, which require changes to policies or procedures, are important for addressing the deeper structural issues that contribute to scam victimization or inhibit recovery from it.

3.5 Discussion

In this section, we discuss the results of our systematic literature review. We make several recommendations for future work and make connections to the most recent work in the field.

Our review revealed that research on scam victimization among at-risk populations is highly uneven. Older adults are by far the most studied group, with other potentially vulnerable populations, such as migrants or people with disabilities, receiving far less attention, and typically only through small-scale or qualitative studies. Many papers explored broad demographic variables

without examining the underlying social or structural factors that shape vulnerability. Measures of scam susceptibility varied widely in quality, with some studies using indirect measures that may not capture real-world risk. While researchers often recommended educational interventions for specific groups, few studies evaluated their effectiveness in practice. Collectively, these findings point to the need for more rigorous, context-aware research that considers how scams affect different groups.

Based on our study's results, we make several recommendations. First, **larger-scale studies with groups other than older adults are necessary to determine how well findings from smaller studies generalize**. While qualitative evidence provides valuable insights into how issues with language accessibility, lack of financial resources, or visual disability may affect vulnerability to scams, the limited scale of existing work makes it difficult to assess the extent of these problems. Scams are generally among the top digital safety concerns for people, regardless of whether they experience certain risk factors [64, 225], so the fact that they emerge as a concern in descriptive research is unsurprising and does not necessarily indicate that a particular group is more "at-risk." For example, Li et al. conducted an online survey of students ($n = 1,074$) to compare domestic (US) and international students' ability to detect scams and their past experiences with scams [289]. International students were more likely to engage with certain scenarios, such as a phone call purporting to be from U.S. Citizenship and Immigration Services; however, they were less likely to engage with a rental scam scenario. Overall, there were no significant differences in the ability of international and domestic students to detect scams, as computed using metrics from Signal Detection Theory [455]. While these results cannot speak to participants' real-world experiences with scams, they suggest that demographic differences do not automatically translate into higher overall susceptibility, underscoring the need for nuanced, large-scale studies.

Scaling up research with specific user groups presents challenges in recruiting and protecting participants. We echo Bellini et al.'s recommendation that researchers explore alternative approaches that avoid direct user interaction where feasible [41]. Other data sources, such as public social media posts, can be leveraged to explore research questions without directly engaging users or exposing them to harm. For example, Chapter 5 analyzes posts about cryptocurrency in teenage communities on Reddit, providing diverse qualitative insights, including posts describing minors experiencing losses in "pump-and-dump" schemes. A number of subsequent publications have drawn on Reddit data to study scams [94, 351], including Chapter 7. Research with social media data comes with its own ethical challenges, especially a lack of consent from "participants" [177, 178]. These must be carefully considered when conducting such research.

The recommendation for more large-scale research on groups other than older adults does not indicate that there are no opportunities for further research focused on older adults. As discussed in the following paragraphs, improving the methodological rigor of scam research involving older adults represents a significant opportunity for advancement. Still, a broader research agenda will improve our knowledge and help protect all users.

We recommend the adoption of better, more consistent measures of scam victimization and susceptibility. The use of inconsistent and low-validity measures limits comparability across studies and may obscure key differences in experience across and within groups. Real-world simulation studies are generally considered to have the highest ecological validity, but many types of scams are ethically impossible to simulate (e.g., romance scams and sextortion).

Moreover, simulations do not capture differences in victimization rates that may result from differences in activity or targeting, rather than inherent susceptibility. Longitudinal studies that measure incidents of victimization over time, possibly employing telemetry to directly measure actual behavior [184, 555], could be an effective approach. However, such studies pose several challenges, especially regarding participant retention and data protection [556].

Considering these challenges, self-report items still have an important place in scam research. When asking participants to report past victimization, research should ensure that participants are provided with a clear definition of terms such as "scam" or "victimized." The meaning of specific type labels, such as "romance scam," should be explained with examples to reduce misunderstanding. For example, Koning et al. provided definitions to participants based on the FINRA Taxonomy [39, 264]. Researchers may also consider using questions derived from long-running projects, such as the FTC Consumer Sentinel [501] or the National Cybersecurity Alliance's Cybersecurity Attitudes and Behaviors survey (2021 to 2025) [339]. While each of these options is imperfect in capturing the full diversity of social engineering, such standardized questions provide a firmer basis of comparison.

When phishing simulations or other behavioral measures of susceptibility are used, efforts should be made to standardize the stimuli. This is inherently difficult in a space where criminals' tactics are constantly evolving, but the ad hoc selection of scam stimuli confounds the study's comparative results. Researchers should purposively select stimuli that span a range of difficulty levels, for example, using the NIST Phish Scale [134]. The scale produces a categorical score of an email's difficulty based on the number and type of cues to deception in the email, as well as how well it aligns with the recipient's context. A recent real-world phishing simulation study conducted by Rozema and Davis ($n = 12,511$) suggests that the phish scale is predictive of phishing susceptibility, with emails rated as "high" difficulty resulting in significantly higher click-through-rate than "easy" emails (15% vs. 7%) [414].

The topic of messages and how this may appeal to different populations should also be considered. For example, Denisenko et al. conducted an online behavioral study ($n = 500$) to analyze how demographics and message factors (mood, topic, recency, and multimedia use) interact to affect participants' self-reported propensity to click on LLM-generated social media posts linking to news articles [141]. Since the stimuli were news articles rather than more typical scam lures, the setup may not fully capture genuine susceptibility to scams. Nevertheless, it provides a helpful example of how to explore interactions between demographics and messages.

We encourage the development and evaluation of a wider range of user-centric anti-scam interventions, in collaboration with affected communities. Our review suggests that group-specific, non-educational interventions remain almost entirely untested outside of studies with PVI, though this finding is subject to the search-strategy limitations discussed in Section 3.1.4. The focus on training, while reflecting a general trend in human-centered social engineering research [186], is overly narrow and may place an undue burden on users. Recognizing that vulnerability may stem from contextual factors or intentional targeting rather than individual failings, participatory design methods [453] may be particularly useful for identifying novel approaches to scam prevention or remediation. For example, Chew et al. conducted workshops with people with acquired brain injury (ABI), their close others, and clinicians to develop a clinical intervention (called "Smooth Sailing After Scams") to help people with ABI recover from scams [101]. The program combined cybersecurity and financial training with emotional sup-

port during 10 group sessions. A pilot program for scam survivors with ABI suggests it improves participants' emotional well-being [102, 103].

Beyond group-specific interventions, it is important to test new anti-scam technology with a range of populations. For example, Google has recently shipped an on-device model to alert users in real time if a call or text message conversation appears to be a scam [36], but the effectiveness of this consumer-facing feature has, to our knowledge, not been evaluated. Future work should test such technologies not only against an "average" user but also with specific populations for whom there is a plausible reason to expect differential performance in the underlying technology's accuracy or in how effectively users can act on its output. For example, accent or language variation may affect a model's ability to correctly detect a scam call in the first place, while cognitive load during high-pressure calls or an inability to visually perceive warnings [434] may affect whether an accurate alert actually reaches the user and is acted on.

Some significant opportunities do still remain in developing anti-scam education. As discussed in Section 3.4.1, group-specific educational interventions seem promising, especially for populations with specific accessibility needs (e.g., PVI or people with cognitive impairment). Delivering effective educational interventions at scale also remains a significant challenge. The majority of papers presenting educational interventions were one-off sessions that may be difficult to extend, whereas examples of large-scale educational programs, such as Button et al.'s study of a UK non-profit [82], revealed significant challenges. More broadly, public awareness campaigns about scams, such as Craig Newmark Philanthropies' "Take9" [123], the American Bankers Association's "Banks Never Ask That" [15], the National Cybersecurity Alliance's "Then & Now" [340], and UK Finance's "Take Five to Stop Fraud" are widespread [489], but their actual efficacy is rarely tested. van Steen et al. reviewed 17 government cybersecurity awareness campaigns from around the world [519], finding that none had been systematically evaluated. Assessment, therefore, remains confined to small, localized sessions rather than the widespread campaigns most consumers actually encounter.

We recommend a greater recognition of the role of contextual factors in scam victimization. Demographic profiling not only fails to explain the causes of apparent differences but also risks stereotyping particular groups. By shifting attention to contextual factors, computing researchers can better identify the conditions that give rise to vulnerability. For example, not all older adults experience cognitive decline, and not all migrants face language barriers. Demographic framings often imply such uniformity.

The recent large-scale survey by Heung et al. presents one possible model for this type of study. Surveying a representative sample of Americans ($n = 5,001$), they asked questions designed to assess whether participants experienced each of Warford et al.'s 10 contextual risk factors [225] and used regression analysis to examine their relationships with demographic factors and prior experiences of digital harm. 49.7% of participants stated they had "been the target of a scam or phishing attack," which was the most common of any of the measured digital safety risks. They found that having a *relationship with an attacker*, *access to a sensitive resource*, and *resource and time constraints* correlated with being targeted by a scam. The statements to assess contextual risk factors have not been validated, and the use of self-report metrics is plainly limited (e.g., men were more likely to report marginalization than women). Additionally, the statement about scam targeting is broad and may invite multiple interpretations. Still, the study represents a significant step towards operationalizing Warford et al.'s framework [225].

Focused narrowly on the effect of language, Spero et al. conducted a phishing simulation study with bilingual employees of the national government of a Melanesian country ($n = 2,000$) [451]. They sent four phishing emails, each of which participants received in English or the local national language. The local-language versions of two phishing emails (both of which included personal requests for assistance) received more clicks than their English counterparts, suggesting that local-language use was more persuasive in these contexts. The authors offer two hypotheses to explain this result: that processing text in one's native language may elicit stronger emotional reactions, or that local-language emails prime culturally specific norms related to collectivism. The authors further speculate that the lack of automated filters developed for Melanesian languages makes it more difficult to defend against these tactics. Regardless of which mechanism is at play, the results provide direct evidence of language's impact on phishing susceptibility and reveal a more nuanced dynamic than a simple linear relationship between fluency and vulnerability would suggest.

Finally, **we call for increased human-subjects research into the nature of scammer targeting**. While it is obvious that much content is tailored to exploit particular contextual vulnerabilities, the extent of active target selection remains unclear. The descriptive results of studies with scammers discussed in Section 3.3.1 suggest that at least some fraudsters preferentially target groups they believe are vulnerable. However, the exact data streams used to source these leads, as well as the balance between precise and random targeting, remain unknown. Clarifying these dynamics is essential to motivate effective, target-side interventions. Many privacy-protecting products are marketed as helping reduce contact with scammers. For example, DeleteMe claims that its service, which submits data removal requests to data brokers, helps to "Stop Spear Phishing At the Source" [135]. Yet if fraudsters do not rely on data brokers, the utility of this form of PII removal for scam prevention may be vastly overstated.

Human-subjects research with scammers is understandably rare in computing research due to logistical challenges and ethical risks. Observing dark web discussions [300] and reviewing criminal records [445, 446] have yielded valuable qualitative insights, though the extent to which these findings generalize remains unclear. A growing dimension of global cybercrime is the emergence of centralized scam compounds, where forced labor is often employed [16, 281, 532]. Because these operations operate at an industrial scale, understanding the structural practices of even a single compound could yield systemic insights that far exceed the utility of investigating individual operators in isolation.

Investigating these entities could begin with an analysis of artifacts, such as the leaked operational materials obtained by WIRED [200]. Researchers may choose to engage directly with survivors of these compounds. Because these individuals were coerced into committing fraud rather than acting as willing criminals, they may be motivated to share insights that could undermine syndicate infrastructure. However, shifting from digital artifacts to direct human subjects introduces profound ethical challenges. As documented by Zheng et al., survivors of scam-driven human trafficking endure severe physical and psychological trauma [567], presenting a high risk of re-traumatization during the research process. Ultimately, the pursuit of these insider insights must be strictly balanced against participant safety.

Chapter 4

Exploring the Privacy Experiences of Closeted Users of Online Dating Services in the US

This chapter focuses on a specific population that experiences the contextual risk factor of *marginalization*: LGBTQ+ individuals who are *in the closet*. Being in the closet or *closeted* refers to when a person internally identifies with a particular sexual orientation and/or gender identity but does not disclose it to some or all of their social relations [361, 429]. Being out of the closet is generally associated with better mental health for LGBTQ+ people [362]; however, being out also requires individuals to navigate stigma and discrimination described in Section 2.3. For people with little support, coming out may not improve their quality of life [287].

We focus particularly on online dating services, a context that requires a user to reveal potentially sensitive information, such as age, gender identity, sexual orientation, and HIV status, in a manner that is essentially public. In addition to facing the typical safety risks of online dating [115, 228], closeted users of online dating services may be targeted for violence [456] or extortion [280]. Even in the absence of a deliberate adversary, these users may also risk incidental outing. For example, in a study conducted by Geeng et al. [191], P4, a gay and trans man, described finding a coworker on Grindr—a popular online dating service predominantly for men who have sex with men—and fearing that they would reveal his identity to others in the workplace. While some research has explored the privacy risks faced by LGBTQ+ users of dating services [228, 528], no prior study has examined the unique needs of closeted individuals.

This study's scope is broader than scams; we present an online, exploratory study of the privacy-related behaviors and needs of LGBTQ+ individuals in the closet on dating services. In particular, we seek to answer the following research questions in the US context:

1. What privacy concerns do individuals in the closet have in the context of online dating services?
2. How do individuals in the closet protect their privacy in the context of online dating services?

This chapter is based on a paper originally published at the 2024 Privacy Enhancing Technologies Symposium and co-authored with Sanjnah Ananda Kumar and Lorrie Faith Cranor [57]

3. How might online dating services better protect the privacy of individuals in the closet?

To answer these research questions, we conducted an online survey with 114 participants recruited via snowball sampling, advertisements on Grindr, physical posters at Carnegie Mellon University, and digital fliers distributed to LGBTQ+-focused organizations at universities around the United States. We also conducted nine follow-up interviews with survey participants via Zoom and email exchange to provide additional context to their survey responses.

We found that closeted dating service users were most strongly concerned about institutional data-sharing practices (i.e., targeted advertising) and the risk of encountering social relations. Most relevant to this thesis, participants discussed experiencing privacy harms and other threats to their safety, including extortion on the basis of their identity. Most participants discussed some form of privacy-protective behavior, most often involving omitting certain information from their dating profiles. Finally, users proposed a range of possible features to better protect their privacy, including restrictions on downloading/screenshotting photos, improved profile visibility controls, and making some paid privacy features available for free.

4.1 Methods

We administered an online survey to 114 participants and conducted nine follow-up interviews. We also analyzed the privacy policy and privacy-related features of the five most popular online dating services used by our participants.

4.1.1 Recruitment

We recruited participants primarily through physical posters and digital advertisements in various LGBTQ+ spaces. All advertisements include a description of the study and a link to take our survey. The headline of each recruitment advertisement indicated that participants should identify as closeted. Appendix B.1 includes copies of the recruitment material.

While there is debate around the operational delineation between being in or out of the closet [173], we rely on self-identification rather than imposing a precise definition of the closet. We focus on people who consider themselves closeted, regardless of their specific identity management practices. To accommodate a spectrum of disclosure practices, we adopt a broad definition of being in the closet as referring to people who do not disclose their LGBTQ+ identity to some or all of their friends, family, and other social relations. In line with the colloquial use of the term, we expect that those identifying as closeted will rarely disclose or discuss their LGBTQ+ identity and experience the stress associated with concealment. By relying on self-identification, we risk excluding some people who meet operational definitions of being in the closet but do not identify with the label "closeted." Likewise, some participants who identify with our broad definition of the closet may not be defined as closeted under other definitions.

We placed posters on the campuses of Carnegie Mellon University in Pittsburgh, Pennsylvania, and Mountain View, California. We also distributed digital copies of the posters to LGBTQ+-focused organizations at universities around the United States. Online, we posted advertisements on LGBTQ+-related sub-forums of the social media service Reddit. We purchased advertisements on the largest LGBTQ+-focused dating service, Grindr.

To increase the diversity of the sample, we recruited a small number of participants through snowball sampling [367]: We sent the survey advertisement to several members of the LGBTQ+ community known to the authors. We asked them to share the advertisement with other eligible users of dating services who might be willing to participate. As our primary goal was to gain qualitative insights into the experience of closeted LGBTQ+ in the US, we stopped recruiting once theoretical saturation was reached (i.e., we observed no new insights from further survey responses [121, pg. 134]).

In total, we recruited 114 participants from October 10, 2022, to January 20, 2023. The majority of participants (55.5%) were recruited via Grindr.

4.1.2 Survey

The survey began with an informed consent form that disclosed the study's risks and benefits and required participants to affirm their consent to participate. In compensation for their participation, they were offered the opportunity to enter a raffle for $125 per 100 people who completed the survey.

As part of the informed consent form, participants were required to complete a series of screening questions that verified their eligibility to participate in the study. In particular, they had to be 18 years of age or older, located in the United States, fluent in English, a member of the LGBTQ+ community, to some degree "in the closet," and have used a dating app or online dating service within the last three months. We instructed participants that they are considered in the closet "if you do not reveal your LGBTQ+ identity to some or all of your friends or family."

The survey text can be found in Appendix B.2. The median completion time was 8 minutes and 56 seconds. The survey began with demographic questions (Q1 to Q6) to characterize our sample (i.e., gender, sexual orientation, age, race, etc.). In asking participants to disclose their gender identity, we used the terms "male" and "female." We recognize that "man" and "woman" are more accurate when referring to gender rather than sex [452]. We, therefore, refer to those who selected "male" and "female" as men and women, respectively. This terminology may have led to inconsistencies in recording gender, especially for trans individuals.

To quantify how approaches to identity disclosure varied between our participants, we next asked them to complete a shortened version of Mohr et al.'s Outness Inventory (OI) [327]. Our version of the OI consists of a series of Likert-scale questions that ask participants to indicate how open they are about their sexual orientation to eight different individuals and groups: their mother, father, siblings, extended family, work peers, work supervisors, new straight friends, and old straight friends. Options range from 1 ("Person definitely does not know about my sexual orientation status") to 7 ("Person definitely knows about my sexual orientation status, and it is openly talked about"), with an additional "N/A" option for questions that do not apply to an individual. For brevity, we removed items from the original OI that we felt were superfluous or irrelevant to many respondents, such as statements concerning their "religious community."[1] The OI is one of the most commonly adopted scales for measuring outness among sexual minorities (e.g., [31, 187, 287, 400, 405, 421]). It has shown high internal consistency [31, 327] and cor-

[1] In prior studies that use the OI, statements about one's religious community are often not answered (e.g., [31, 287, 327]).

relates well with other measures of outness, such as the Nebraska Outness Scale [319]. The OI has some limitations: it does not account for non-traditional family structures (i.e., those without a mother or father, those with multiple parents of the same gender, etc.), and it fails to assess how people present their gender identity. It also conflates a person's knowledge of one's sexual orientation with how often one's sexuality is discussed [173].

The next set of questions (Q8 to Q12) characterized participants' use of online dating services. We asked which dating services they use, when they first used dating services, the last time they used online dating services, how frequently they use online dating services, and what their primary purpose is for using dating services.

The next section of the survey focused on participants' privacy behavior. We first asked (Q13) participants to select what information they include in their dating profile from predefined options (e.g., name, photos, general location, etc.). About six weeks after the survey was launched, we added a question (Q14) asking whether the participant's dating service(s) required location permissions. We noticed that many participants did not indicate that they shared their location on their profile despite using services that depend on GPS information. We wanted to verify that participants understood that the dating services use location information. 80 participants saw this question.

The following two questions explicitly asked about privacy-protecting strategies. We first asked whether participants do anything to protect their privacy on dating services (Q15). If they answered yes to this question, we asked them to explain how they protect their privacy in a free-response question (Q16).

The next four questions (Q17–Q20) consisted of a series of 21 five-point Likert scale questions, requesting that participants rate their agreement or disagreement with privacy-related statements, where 1 corresponds to "Strongly disagree" and 5 corresponds to "Strongly agree." Following Lutz et al. [305], these statements were intended to gauge participants' concerns about "institutional privacy" risks (i.e., how platforms and other institutions collect and use data) and "social privacy" risks (i.e., risks from other individuals) associated with the use of online dating services. We also included statements to assess privacy-related behavior participants may engage in (e.g., "I lie about information on my dating profile(s) to protect my privacy."). These statements were developed based on the results of prior research [115, 305, 529]. Cobb et al.'s survey of the general population of dating service users is the most directly comparable prior study; we reference their results throughout to contextualize how closeted users' experiences differ [115].

After the Likert-scale questions, we asked participants again whether they did anything else to protect their privacy (Q21), as the agree/disagree statements may have prompted users to think of additional behaviors they engage in to protect their privacy. We analyzed the responses to this question separately from the first time participants were asked about privacy-protecting behavior. We refer to the responses before the Likert-scale questions as "unprimed" and those after the Likert-scale questions as "primed."

In the following question (Q22), we asked participants whether their privacy had ever been violated while using online dating services; if so, we asked them to describe what happened. About five weeks after the survey was launched, we added a follow-up question asking whether participants had been threatened on an online dating service and, if so, asked them to describe what happened. This question was marked as optional. It was added because several participants

alluded to being threatened by other users of dating services without describing the specifics of what happened. 84 participants saw this question.

The survey's final question (Q24) was an open-ended question asking participants to share any feature they would like to see improved or added to dating services to better protect their privacy. We requested that participants make specific reference to the service they use so that responses were adequately contextualized.

4.1.3 Follow-up Interview

After the survey (Q25), participants had the opportunity to volunteer to be interviewed about their responses in exchange for a $15 Amazon gift card. To protect participants' privacy, we offered them the opportunity to either participate in an audio-recorded Zoom interview or answer questions via email. Our primary goal in conducting interviews was to gain additional insight into specific experiences and feature recommendations from survey participants. They also allowed us to partially explore questions not in our survey (e.g., gender disclosure, motivation for using particular services, etc.). During the interview, we asked participants specific questions based on their survey responses. When necessary, we also asked follow-up questions to clarify their responses.

Of the 73 participants who indicated willingness to participate in a follow-up interview, we invited 21 to interview. We recruited participants who provided non-trivial survey responses. We focused on recruiting those who shared experiences of privacy violations. Potential interviewees were sent an email outlining the interview procedure and inviting them to confirm their interest in being interviewed. Nine participants ultimately took part in interviews, five via Zoom and four via email. The average Zoom interview length was 16 minutes 42 seconds.

4.1.4 Qualitative Data Analysis

We analyzed responses to open-ended questions using qualitative coding [279]. The lead coder determined codes inductively based on 20% of the responses to each question. The lead coder and another author then independently coded the entire set of responses for each question. New codes were added during this stage as needed. After all responses were analyzed, the coders reconciled any differences and finalized the codebooks for each question. The complete list of codes can be found in Appendix B.3. When appropriate, we reference interview responses to provide deeper qualitative insights. We refer to participants with the letter S followed by a number between 1 and 114. We provide specific demographic information for survey respondents who participated in interviews in Table B.2.

4.1.5 Correlational Analysis

In order to measure the extent to which participants' responses to the Likert scale questions relate to their reported level of outness, we computed Spearman's rank correlation coefficient (ρ) between the responses to each Likert scale question and the average OI score. In order to compute the average OI score, we exclude items where users selected N/A. The computation of ρ is non-parametric and does not require the data to be normally distributed. Values can range from -1 to

Table 4.1: Summary of participant demographics collected via the survey. Some demographic category names are shortened for space, but the complete text can be found in appendix B.2.

Age (Years)		Race/Ethnicity		Education	
18 to 25	49.1%	American Indian or Alaska Native	1.8%	Less than high school	1.8%
26 to 33	32.5%	Asian	14.0%	Graduated high school	20.2%
34 to 41	7.0%	Black	15.8%	Some college education	28.1%
42 to 49	5.3%	White	65.8%	Associate's degree	6.1%
50 to 57	3.5%	Self-describe	7.9%	Bachelor's degree	26.3%
Above 57	2.6%	No response	0.9%%	Degree beyond bachelor's	17.5%
No response	0%			No Response	0%
Gender		**Sexual Orientation**		**Household Income**	
Agender	0.9%	Asexual	1.8%	Less than $25,000	24.6%
Women	25.4%	Bisexual	32.5%	$25,000 to $50,000	23.7%
Genderqueer	4.4%	Homosexual	34.2%	$50,000 to $100,000	22.8%
Men	57.9%	Pansexual	12.3%	$100,000 to $200,000	14.9%
Non-binary	7.9%	Queer	12.3%	More than $200,000	5.3%
Self-describe	3.5%	Self-describe	6.1%	No Response	8.8%
No response	0%	No Response	0.9%		

1, where an absolute value above .7 indicates a strong correlation, an absolute value between .7 and .4 indicates a moderate correlation, and an absolute value between .4 and .2 indicates a weak correlation [8]. We also provide a p-value, where a significant result corresponds to $p \leq \alpha = .05$. We report the complete results of the analysis in Table B.1 in Appendix B.4.

4.1.6 Inspection of Dating Services

To better understand the existing privacy posture of dating services, we systematically analyzed the privacy policies and features of the five most popular online dating services used by our participants: Grindr, Tinder, Bumble, Hinge, and OkCupid. For each service, we first analyzed the privacy policy, recording the types of data collected and how that data is used. We then installed the Android app for each service on a Samsung Galaxy Z Fold 3, running Android version 13 and OneUI 5.1. We completed the registration process for each service using fake personal details. We then reviewed all settings in each application, recording the privacy-related features available on each service. We did not interact with any users during this process. All steps were completed on April 4-5, 2023. We reference the results of this analysis as appropriate to contextualize participant responses. For example, this analysis allowed us to determine whether certain requested features were available. It also permitted us to determine the institutional data-sharing practices of the most popular dating services. Finally, we considered these results when developing recommendations for changes.

4.1.7 Ethics

Due to the sensitive nature of this study and the risk of harm to our participants, we took ethics very seriously. Carnegie Mellon University's Institutional Review Board approved our complete study protocol. To protect participants' privacy, all participants could complete the survey anonymously. For those who wished to enter the raffle, we collected email addresses in a separate survey so that their email would not be tied to a specific survey response. Email addresses collected for compensation and facilitating interviews were deleted after the study. Audio recordings of the interview were deleted once the interviews had been transcribed. Email exchange threads were deleted once their contents were extracted. All potentially identifiable information disclosed during interviews (e.g., names, places of work, etc.) was redacted from interview transcripts and email exchange threads. Analysis was performed on password-protected devices accessible only to the research team. To protect participants' emotional well-being, all participants were informed during the consent process that they could discontinue participation at any time. Interview participants were told that they could choose not to answer any questions or stop participating at any time while still receiving compensation. No participants expressed concerns about their emotional well-being during or after the study.

4.1.8 Limitations

Due to our recruitment methods, college students, Grindr users, and men are overrepresented in our sample. In addition, participants who were willing to volunteer for our study may differ from the larger population of LGBTQ+ people who are in the closet (i.e., self-selection bias [45]). For example, they may be more conscious of privacy issues (and thus interested in discussing them with researchers) but not so private that they are unwilling to share information with researchers that they might not share with friends or family. Our broad definition of being in the closet may also have led to some people participating in the study who would not typically be defined as closeted. We could only recruit a small subset of survey participants for interviews, which may limit the validity of these results. We rely on users to self-report experiences, and the results may be influenced by social desirability bias [204] or poor recall of past events. Finally, our qualitative coding is necessarily subjective and influenced by the researchers' experience and attitudes. Another research team may have identified different codes after reviewing the data.

4.2 Results

In this section, we present the results of our study. In order to contextualize our findings, we begin by summarizing the characteristics of our participants, including their demographics, usage of dating services, and responses to the OI questions. We then discuss the results concerning each research question. Critical takeaways regarding each research question are shown in bold.

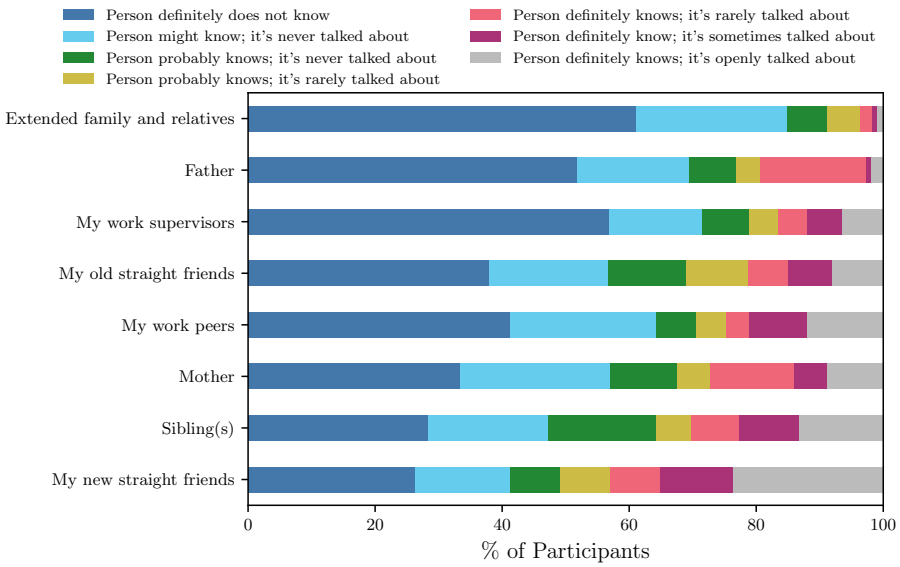


Figure 4.1: Summary of how "out" participants are to different types of social relations. The category labels are shortened to conserve space, but the complete text can be found in Appendix B.2. Categories are ordered from lowest to highest average OI value.

4.2.1 Participant Characteristics

Table 4.1 summarizes the demographic characteristics of our participants. Our sample is primarily young men. The population of US-based dating service users is also mostly young men [314]; however, the demographic imbalance is less extreme than observed in our sample. The voices of women and those older than 33 are likely underrepresented in our results.

The most common dating service used by our participants was Grindr (66.7%), followed by Tinder (57.0%), Hinge (32.5%), Bumble (27.2%), and OkCupid (7.9%). The majority (68.4%) of participants used more than a single dating service. When asked why they used particular dating services, most interviewees gave answers related to the popularity or features of the service rather than privacy. For example, S102 stated that they use Grindr and Tinder the most because "I am more likely to be able to interact with someone and have that lead to a date or a hookup." When asked whether privacy considerations affected their choice of dating service, S102 stated, "Privacy hasn't factored into my consideration of which apps to use." Similarly, when asked why they use Grindr, S66 stated, "It seems more popular in the area in which I live."

S11, a non-binary person who identifies as queer, did mention a specific privacy feature as motivation for using a dating service. When asked in their interview why they use OkCupid, they wrote, "I've heard that OkCupid is great for LGBT people. It also has a well-developed web experience so I don't have to use it on my phone: using a dating app on my phone has always been nerve wracking since a notification (e.g. 'David sent you a message!') can easily out me. I used to use Tinder and would delete the app before seeing family. OkCupid sends message notifications over email so I don't have to worry as much..." S33, a woman who identifies as pansexual, also mentioned a privacy motivation for her selection of dating service, albeit as a reason not to use a particular app. When asked in her interview why she uses Bumble and Hinge,

she mentioned having a positive experience with these services. She added that she does not use Tinder because her grandfather uses the application. She stated, "I know my grandpa only uses Tinder, so I think that gives me peace of mind because he's not on Hinge or Bumble. I don't think he really knows how to use them.... The rest of my family members, except for, like my cousins, aren't on dating apps, and I would be fine if my cousins found out."

The vast majority of our sample reported using dating services frequently and for an extended period. 73.7% of the participants reported using dating services a few times a week or daily. 84.2% of participants reported that they began using online dating services at least a year ago, with 50% of all participants using dating services for more than three years. Finally, 40.4% of participants reported that their primary purpose for using a dating service was "to have casual sex," 27.2% of participants reported that they wanted "to find a romantic partner," and 14.0% selected that they wanted "to meet new friends or acquaintances." Many users likely have multiple, overlapping purposes for engaging with dating services. For example, when asked in a follow-up interview about their reason for using dating services to find friends, S106 clarified that they look to "[meet] new friends and see if there's something more to it." They elaborated, "I wouldn't say I trust dating app[s] that much, so I wouldn't just go for a relationship... that quickly."

Figure 4.1 displays the results of the set of outness inventory questions. The participants were least likely to be open about their sexuality to their extended family/relatives, work supervisor(s), or father. Participants were most likely to be open about their sexuality to their "new straight friends." 24.6% of participants indicated that they rarely or never discussed their sexuality with any of the individuals or groups included in the OI. Only 1 participant indicated that all individuals or groups in the OI "probably" or "definitely [knew] about [their] sexual orientation."

The mean OI score was 2.74. Although the comparison is imperfect due to variations in implementation, this value is lower than that reported by other surveys with the general LGBTQ+ population in the US. For example, a 2007 study by Balsam et al. [31] reported an average OI of 4.89 with a community sample of 613 lesbian, gay, and bisexual (LGB) adults. In their 2014 online study of LGB adults, Meidlinger et al. [319] reported a mean OI score of 3.80 for 24 bisexual people, a mean OI score of 4.43 for 23 "mostly gay/lesbian" people, and a mean OI score of 5.10 for 102 gay or lesbian people. Insofar as the OI is an accurate measure of outness, this result suggests that we succeeded in recruiting a sample that is more closeted than the general LGBTQ+ population.

As previously stated, the OI does not capture how participants present their gender identity. The OI questionnaire focuses exclusively on sexual orientation. Some LGBTQ+ people may be more open about their sexual orientation than their gender identity. For example, we asked S102, a non-binary person who identifies as queer, to discuss how "out" they are about their sexual orientation and gender identity, respectively, in their interview. They wrote, "I am more 'out' about my sexual orientation... I only discuss the fact that I'm nonbinary with other nonbinary or queer people who can be more understanding, and I worry that people outside of the community will ask a lot of questions about what that identity means." In response to the same question in their interview, S1, a non-binary person who identifies as pansexual, wrote, "I wouldn't say I'm very out about either of [my identities] to anyone in my personal life... only my very best friends would know even one of [my identities], and even fewer would know both."

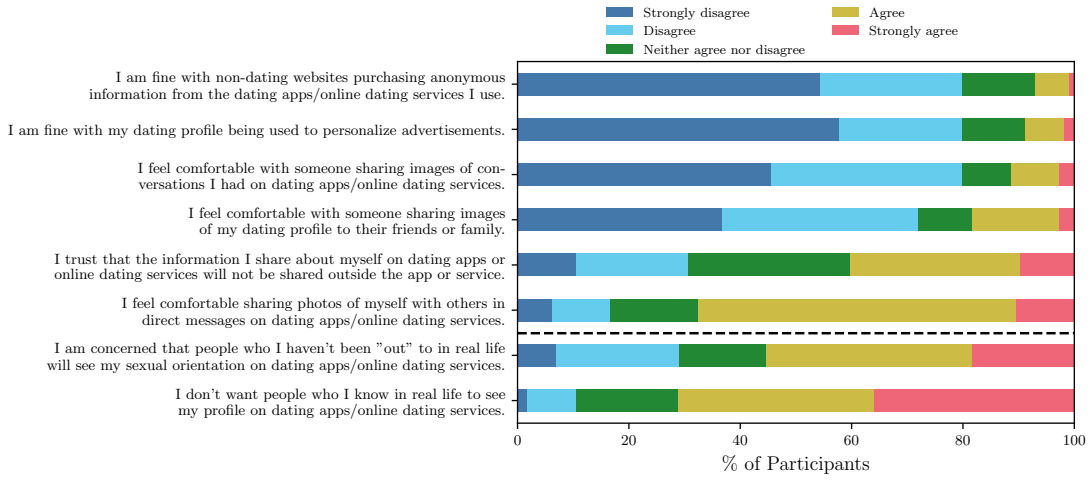


Figure 4.2: Proportion of participants selecting each response for Likert scale questions related to privacy concerns. Statements are organized from the highest to the lowest proportion of disagreement. A dashed line separates two statements with opposite polarity. The complete questions can be found in Appendix B.2.

4.2.2 RQ1: Privacy Concerns

Eight of the Likert scale questions attempt to assess the extent to which our participants have specific privacy concerns (Q17.2-4, Q18.1-5). Figure 4.2 displays the proportion of participants agreeing or disagreeing with each statement, in order from statements with the highest to lowest proportion of disagreement. Two statements have opposite polarity and are therefore separated from the other statements by a dashed line. Additionally, 35.1% of participants reported that their privacy had been violated in some way (Q22). We discuss the results of the Likert scale questionnaire below, along with relevant descriptions of privacy violations.

Study participants expressed the strongest negative sentiment towards statements describing institutional data sharing practices. An equal proportion of participants (79.8%) disagreed or strongly disagreed that they are "fine with" non-dating websites purchasing anonymous information from dating services (Q18.4) and profile data being used for personalized advertising (Q18.5). Likely prompted by these statements, several participants expressed dissatisfaction with data sharing when asked what should be changed to protect their privacy better (Q24). S91, a woman who identifies as biromantic and demisexual, wrote, "I don't know if dating apps sell that data to personalise ads, but if they do I would prefer that they dont." Similarly, S60, a man who identifies as homosexual, responded to the same question by stating that he did not want to "...have info sold to companies as a means to show personal ads." When asked if his privacy had ever been violated (Q22), S64, a man who identifies as bisexual, described that he "...used to get YouTube ads that were clearly asking to take a gay test to see how much of a gay someone could be...." While targeting based on sensitive attributes such as race or sexual orientation is generally prohibited by advertising platforms, advertisers may still target based on keywords or topics of interest that are closely related to one's sexual orientation or gender identity [537].

Participants reacted negatively to statements describing someone taking and/or sharing images of activity on dating services. 79.8% of participants disagreed or strongly disagreed that they were comfortable with someone sharing images of conversations they had on an online dating service (Q18.3). 71.9% of participants disagreed or strongly disagreed that they were comfortable with someone sharing an image of their online dating service profile (Q18.2). Notwithstanding this adverse reaction, some participants indicated that they had engaged in similar behavior. The majority of participants (54.4%) agreed or strongly agreed that they had previously taken screenshots of images or messages they had received on online dating services (Q19.4). Half of the participants agreed or strongly agreed that they had shown conversations they had on online dating services to other people (Q20.1). Almost as many (43.0%) agreed or strongly agreed that they have shown other people's dating profiles to their friends or family (Q19.5).

Most participants in Cobb et al.'s study [115] were also concerned by others taking screenshots of dating service activity. Around half of their participants also reported taking screenshots of others' dating profiles or messages. Some users may wish to share images of others' profiles for safety reasons (i.e., so the person they are meeting with can be identified if something goes wrong). Screenshots of messages may be taken for sentimental reasons [115], to publicly shame people perceived to be behaving badly (e.g., "West Elm Caleb" [450]), or for entertainment (e.g., [164, 309]). Closeted individuals may be more at risk for harm from people taking and sharing photos of activity on dating services, but their concern about screenshot sharing is not unique.

Participants expressed a strong desire to avoid people they knew in real life while using dating services. The statement "I don't want people who I know in real life to see my profile on dating apps/online dating services" had the strongest (negative) correlation with average OI ($\rho = -.449$, p-value $< .001$) with 71.1% of participants indicating that they agreed or strongly agreed with the statement. The negative correlation indicates that participants with a lower average OI score (i.e., participants who were less out) were more likely to agree with this statement. Similarly, responses to the statement "I am concerned that people who I haven't been 'out' to in real life will see my sexual orientation on dating apps/online dating services." had a weak, negative correlation with average OI score ($\rho = -.385$, p-value $< .001$). 55.3% agreed or strongly agreed with this statement. When asked if their privacy had ever been violated, a few participants (3) described incidents of "context collapse" where people they knew in real life found their dating profiles. For example, S86, a man who identifies as homosexual, shared, "A coworker approached me after finding one of my dating profiles and publicly asked questions about it." S7, a man who identifies as bisexual, wrote, "...I've had people show friends I was on apps who didn't know I was bisexual." These experiences reflect the inability of users to control who sees their dating profiles fully.

Perhaps because of the risk of inadvertent disclosure, survey participants were divided in their responses to the statement, "I trust that the information I share about myself on dating apps/online dating services will not be shared outside the app or service." 40.4% of participants agreed or strongly agreed with the statement, 30.7% of participants disagreed or strongly disagreed, and 28.9% of participants neither agreed nor disagreed. In his interview, S25 expressed trust that the LGBTQ+ community on dating services would not violate his privacy, stating, "I just like don't

expect any like queer person who is also on these apps to like, find this information, and share with, like my parents or my family, who I'm not... out to."

The high degree of concern over encountering social relations on dating services seems to distinguish closeted individuals from the general population of users. For example, while inadvertent disclosure of profile information is discussed by Cobb et al. [115] as a potential privacy risk of using dating services, their respondents expressed both positive and negative sentiments towards encountering people they knew on dating services. Their participants were most strongly concerned about dating profiles being viewed by coworkers or employers, although some expressed concern about family or friends viewing their profiles.

Most participants felt comfortable sharing photos of themselves with others; however, this was also a common vector for harm. 67.5% of participants agreed or strongly agreed with the statement, "I feel comfortable sharing photos of myself with others in direct messages on dating apps/online dating services." The most frequently reported privacy harms were related to the unauthorized saving or sharing of photos. Eight participants described how people shared their photos without permission. For example, S13, a man who identifies as bisexual, wrote that "...[someone] told me that he will share my videos and pics in websites." S39, a man who identifies as homosexual, shared that his privacy was violated when "...Someone [shared] photos and conversations to a person who then Messaged me telling me what had happened." Two other participants described people violating their privacy by downloading or taking screenshots without sharing them. S66, a man who identifies as homosexual, described how someone he met on Grindr took screenshots of photos he shared on the messaging app Snapchat. He stated, "...you get the message that a screenshot has been taken, and they don't ask. And that's just very unsettling because you don't know what they are going to do with that... picture that... you share with them."

In some cases, photos were used as a method of extortion. For example, when asked if their privacy had ever been violated, S63, a man who identifies as bisexual, wrote, "Someone used my photo and also threatened to blackmail me with an [in]appropriate video of me. I chose to ignore it and thankfully nothing happened." S74, a woman who identifies as pansexual, described a similar incident where someone tried to force them into interaction. In her survey response, she wrote, "...a couple of years ago I had someone who figured out who I really was get angry with me and they threatened to out me on social media." In her interview, she explained, "At the time I had only recently graduated from high school, and they recognized who I was by this half sliver of my face... I was like 'No... I don't really want to talk to anybody that I went to high school with' and they got angry by that, and they were like 'well, you know I know it's you and I already screenshot your picture... and if you don't hang out with me or continue to talk to me, I'm going to post your profile and this half picture of you online." Ultimately, S74 did not give in to the threat, and she faced no consequences. She shared, "I was like 'Well, I don't want you to do that, but I'm not going to hang out with you,' and they just never did anything, so I think they are really just trying to call my bluff. I suppose." Other survey participants described being threatened with outing, but did not describe a specific motive. For example, S85, a man who identifies as bisexual, shared that "Someone looked me up on Facebook and threatened to out me." In total, eight participants discussed either being extorted or threatened with outing.

Closeted users are especially vulnerable to extortion based on their LGBTQ+ identity; however, both closeted and non-closeted users alike may be threatened with extortion using intimate

photos and videos shared on dating services. This "sextortion" is relatively common both on and off dating apps, including among individuals under 18 [358, 549]. It has even led to the death of victims [32]. Notably, the extortion cases our participants described were primarily motivated by interpersonal or reputational harm (e.g., a desire to punish or coerce) rather than direct financial gain. The same leverage is also used by financially motivated scammer [280].

Participants also discussed several other types of harm previously described in the literature. Four participants described some form of harassment or stalking. For example, S111 wrote in their survey response that a "stalker tried to message me across many accounts and somehow figured out the general area i live in, continually objectifying me and pleading me for favors until i threatened to contact law enforcement." In their interview, when asked if anything happened as a result, they added, "They messaged for the first time in years a month or two ago, and I immediately blocked." Two participants described someone impersonating them. S28, a man who identifies as homosexual, wrote, "People represented my intimate pics as their own." S70, a man who identifies as bisexual, wrote, "Someone recently used my profile name and pic claiming to be me."

The non-privacy harms discussed by our participants are not unique to closeted individuals. In a qualitative study with 20 sexual minority participants living in rural areas, Lauckner et al. [277] observed similar safety harms. In their sample of mostly men, they observed "catfishing" or impersonation, harassment, and sexual coercion. Similar concerns were among those raised by participants in Cobb et al. [115], suggesting that all users of dating services face these risks.

4.2.3 RQ2: Privacy-protecting Behavior

Six of the Likert scale statements were related to information disclosure and strategies that participants may use to protect their privacy (Q17.1, Q17.6, Q19.1-3, and Q20.2). Figure 4.3 displays the proportion of participants agreeing or disagreeing with each statement, in order from statements with the highest to lowest proportion of agreement. We discuss responses to these statements below, along with other participant responses related to privacy-protecting behavior.

Most participants try to protect their privacy in some way, often by restricting the information and photos they include in their profiles. The majority (71.1%) of participants answered "Yes" when asked if they did anything to protect their privacy on online dating services (Q15), with 43.9% of participants describing more than a single strategy. Unprimed, 36 participants shared that they do not use their full name; 22 participants indicated that they avoid disclosing their exact location; 6 participants avoid disclosing their occupation or educational institution; and 14 participants specified that they generally try to restrict the amount of information they share. An overwhelming majority (84.2%) of participants agreed or strongly agreed with the statement that they limit the amount of information they include in their dating profile (Q19.1).

We asked participants to select the information they typically include in their dating service profiles (although we realize that this may vary over time and by platform). The vast majority of study participants indicated that they include their gender (85.1%), age (81.6%), sexual orientation (71.1%), physical characteristics (68.4%), general location (67.5%), or photos of self (65.8%) in their profile. Of the 80 participants who saw the question about location permis-

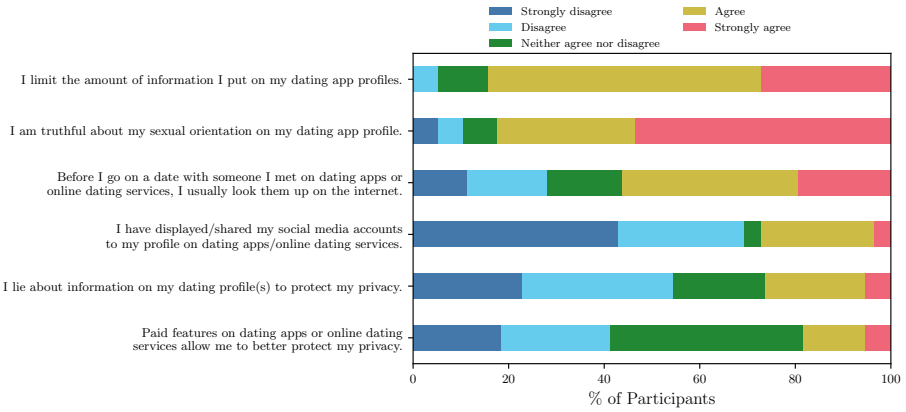


Figure 4.3: Proportion of participants selecting each response for Likert scale questions related to information disclosure and privacy-protecting behaviors. Statements are organized from the highest to the lowest proportion of agreement. The complete text can be found in Appendix B.2.

sions, 86.3% indicated that the dating service(s) they use require access to location data. Less than half of participants use their first name (38.6%) or even just their initials or a nickname (24.6%). Less than 10% of participants indicated that they included their exact location (9.65%), full name (7.02%), or other information (4.39%). Finally, about a quarter (27.2%) of study participants agreed or strongly agreed with the statement that they have displayed links to other social media services in their dating service profiles (Q19.2). Some participants (five unprimed and seven primed) described not sharing their social media profiles as a strategy to protect their privacy. While linking a social media account associated with one's real-world identity may deanonymize users, linking to other social media platforms can provide utility. For example, a user might reference a profile on a platform with additional message privacy features (e.g., Snapchat or Telegram) to make it easy to move the conversation off the dating service.

Participants also discussed restricting the photos they share in various ways to protect their privacy. Participants most often wrote that they did not include their faces in photos they shared (12 unprimed, five primed). Some participants (seven unprimed, one primed) avoid including photos in their profile altogether. Two participants (unprimed) indicated that they avoid sharing publicly available photos. Two participants (unprimed) shared that they avoid including their tattoos in their photos. Some participants described restricting photos in other specific ways (nine primed, five unprimed). For example, S5, who identifies as non-binary and queer, wrote, "I make sure that you can't tell where I am in the pictures I take, so no scenic backgrounds or anything with street names in the background." Restricting photo-sharing was also observed as a privacy-protecting strategy in Waldman's study of intimate photo-sharing practices with men who have sex with men [528, 529] and Lerner et al.'s study of the online computer security and privacy practices of transgender people [288]. Not sharing intimate images with identifying information may be an excellent way to avoid sextortion, as an attacker may not be able to prove that a photo shows the victim.

As one may expect, closeted users seem to include less identifying information in their profiles as compared to the general population of dating service users. For example, in Cobb et

al. [115]'s survey, 63.9% of participants reported including their first name on their dating profile, and only a single participant reported not including pictures of their faces. These information restriction strategies may help prevent the re-identification of users on the dating services. Not including photos in one's profile or cropping one's face out of profile pictures likely prevents identification by a passive observer. Depending on the specific combination of information that a user includes in their profile, however, it may be possible for a determined adversary to identify them. Especially in geographic areas with few users of a particular service, one's age, gender, general location, and photos may be sufficient to permit re-identification [468]. If a user provides additional information in conversation, this may help an attacker to narrow their search.

In coordination with their strategy of information restriction, many participants (11 unprimed, five primed) discussed some form of screening process that they use to determine whether to share more information with another person on a dating service. For example, S111 said that they "crop out face on my profile [and] only share it with those i become comfortable with..." When asked in their interview to explain how they chose whom to share their full photos with, they wrote, "Usually all I require is a good vibe from the other person, a conversation to tell if they're nice and not weird right off the bat or an interaction to show i can trust them to be a real person. A picture of themselves is usually mandatory." Similarly, S66 shared on the survey that "I only share pictures when trust is developed. I also work to build trust!" When asked to describe how he built trust with others, he said he has "basically multiple conversations. Really trying to get to know someone in... that virtual world."

This strategy may not be effective against an active adversary (e.g., one planning to extort a user). An attacker may be willing and able to have a normal conversation long enough to elicit photos or other identifying information. If a photo is requested of them, the attacker could use photos available online or from other users to mimic an equitable exchange.

Most participants indicated that they do not lie to protect their privacy on their dating profiles. More than half (54.4%) of study participants disagreed or strongly disagreed with the statement that they lie about information on their profile to protect their privacy, with only 26.3% of participants agreeing or strongly agreeing with the statement. Additionally, a large majority of participants (82.5%) also agreed or strongly agreed with the statement that they are honest about their sexual orientation on dating app profiles (Q17.1).

Few participants indicated using paid dating service features to protect their privacy. 41.2% of participants disagreed or strongly disagreed with the statement that paid features help them protect their privacy. 40.4% neither agreed nor disagreed. Considering only the 27.2% of participants who agreed or strongly agreed with the statement, "I have paid to access additional features on dating apps/online dating services," a plurality of 45.2% still disagree or strongly disagree with the statement that paid features help protect their privacy. Only four participants, all primed by viewing the Likert-scale statements, indicated that they use premium features to protect their privacy. For example, S64 wrote, "When I was using the paid version [of Grindr], I used to unsend pictures if I didn't trust someone or block[ed] them."

Very few participants discussed using technical means to protect their privacy. Most often, participants described using different throwaway accounts either to register for dating services or to contact others outside the service. For example, S15 wrote, "Often I will use a fake # to chat with people off the app." S85 wrote that he "...uses a basic email that doesn't link directly to [him]." A few (two unprimed, one primed) mentioned using a VPN to protect their

privacy, although they did not detail how they felt the VPN helped them protect their privacy. A VPN may be useful for preventing others on a network or the ISP from seeing their traffic. This functionality could be useful for closeted individuals who fear that accessing dating services may out them to network operators (e.g., their school, employer, family members, etc.).

4.2.4 RQ3: Improving Privacy Experiences

When asked in the survey what features they would like to protect their privacy better, users requested a wide variety of changes. We discuss the nature and frequency of suggestions made in survey free-response questions and participant interviews.

The most commonly requested type of feature (14 participants) was some form of enhanced profile visibility controls to help avoid inadvertent disclosure. Most often, this was expressed as a desire to prevent specific people from seeing their profiles. S102, who uses Tinder, Grindr, and Bumble, wanted platforms to "Give users more control over whom their profile is shown to. This would reduce the likelihood of someone they are not comfortable being out to happening upon their profile." When asked in their interview to explain how they imagine this feature working, they elaborated, "I imagine that this feature would allow you to block certain groups from seeing your profile. For example, options to 'only show my profile to: ____' would be helpful. I would love to be able to limit my profile's visibility to people within my age group, for example, or to limit my profile's visibility to other people who identify as queer on their account. It would also be nice to block your profile from other accounts that it can recognize in your contacts, so you could choose to restrict people you know from seeing your account." The idea of blocking contacts is implemented in Tinder but not the other services S102 uses. Some users requested the broader ability to hide one's profile entirely. S39, who uses Tinder, Grindr, Bumble, and OkCupid, requested the ability to "[allow] your profile to be shown or hidden." It is not entirely clear what enhancement he wants, as all four of the dating services he mentioned provide the ability to hide one's profile. He may be unaware of these settings or desire the ability to interact with others while remaining hidden.

The next most common feature request (13 participants) was a method to limit screenshotting or downloading photos to prevent their misuse. S14, who uses Tinder, Grindr, and Hinge, requested "An app that is completely bulletproof from people stealing other peoples' images." Similarly, in their survey response, S66 stated that Grindr should "[stop] the ability to screen shot... Scruff does not allow screen shots." When asked in his interview why he stopped using Scruff despite it having this feature, he stated, "...I live in a very rural area... there aren't a lot of people on [Scruff] in my area." While people may be able to circumvent restrictions, blocking screenshots would help protect users from all but the most determined attackers.

Some participants (13) expressed a desire for features that are currently only enabled by a premium subscription to be available to free users. Several users opined about the unfairness of paid privacy features. For example, S15, who uses Tinder and Grindr, wrote, "I think it's crazy that you must pay for privacy features, they should be standard." Other users named specific features. For example, S108 and S87 requested that the "incognito mode" of Grindr, Tinder, Bumble, and OkCupid be free. While the specifics vary, incognito modes allow users to hide their profiles from most users while still being able to speak with or match with new people. Other paid features that participants requested be made available for free include

the feature to unsend messages (Grindr), expiring photos (Grindr), and the ability to share private photo albums (Grindr).

Some participants (8) requested the ability to disable or limit the use of their exact location. For example, S54, who uses Grindr, Hornet, and Blued, suggested, "Maybe [they] can support the feature of disabling GPS." Of the dating services we investigated, OkCupid and Hinge allow you to use the service without providing location permissions. Allowing users to provide only a general location, as permitted by OkCupid or Hinge, could protect users from triangulation attacks. However, Tinder, Bumble, and Grindr may be unwilling to allow disabling GPS, as each service offers a paid feature where you can input a city name and view profiles at that location. Allowing users to set their location manually would make this feature less useful.

Some participants who requested the ability to limit or turn off GPS services wanted the specific ability to hide their distance from other users. For example, S111 wrote on the survey that Grindr should "add an option to not show specific distance... (EG, 1,234 feet away is way too specific)." This feature is already available on Grindr, Tinder, and Bumble, while Hinge and OkCupid only show the city or neighborhood in which a person lives. When informed that this feature does exist, S111 wrote, "Thanks! I assumed this was a premium feature, i had not been able to easily find it myself." This example reflects a trend where several users were seemingly unaware of existing privacy features.

This lack of knowledge, even among a population with high privacy concerns, may suggest that dating services should place greater emphasis on privacy education when onboarding users. Three participants explicitly requested improved privacy education. For example, S63, who uses Grindr and OkCupid, requested "More transparency in how the data is being used? Usually it's in the form of a T&C which people don't really read. It's not digestible." S86, who uses Tinder, wrote that the service could include "...maybe a quick tutorial before you add photos to remind you to be careful of adding photos with identifiable info or ones that could be easily imaged searched on google." Tinder highlights privacy and safety features by including a safety center with guides on configuring the application and staying safe in real life. During the registration process, Tinder also highlights the ability to block specific people from your contacts. In order to improve users' experience, other services could include similar privacy education measures.

Some participants (6) requested a form of user verification to reduce impersonation and fraud. S79, who uses Bumble, Hinge, HER, Zoe, and Taimi, wrote, "I think profiles should always be verified because there are a lot of scammers out there." Similarly, S49, who uses Tinder, Grindr, Bumble, and Hinge, requested "A feature to only see verified people and reduce the bots that message. Some services have this but I wish especially Grindr would add this." All of the services we investigated, except Grindr, have some form of selfie verification where a live photo or video of a user is compared to the photos in their profile. If the user matches, they receive a verified mark on their profile. S73, who uses Tinder, Bumble, and Hinge, wrote that to protect her privacy, "i verify my account and don't like anyone that isn't verified." In the interview, she further explained, "'I feel as though people who are verified are less likely to be stalkers or catfish or people that I know making fake accounts and that gives me a sense of peace."

In addition to requesting new features, some participants requested enhancements to existing features. Most frequently, participants requested improvements to how blocking other users works (5 participants). Grindr was the most frequent target for criticism as it limits the

number of people a free user can block per day. For example, S71, who uses Tinder and Grindr, wrote that Grindr should "Allow unlimited blocking." This limitation represents a clear safety risk, as users may be prevented from blocking users who are harassing or stalking them. Some users also criticized Tinder's blocking. S46, who only uses Tinder, wrote that he wants "...easier ways to report and block profiles for misconduct that remove[s] your profile from their screens." You can block users encountered in the discovery queue, but you can only "unmatch" people whom you have already liked and matched with. While this appears to be functionally equivalent to blocking, it may improve usability if a block option is included.

Many of these feature requests directly relate to the risks closeted individuals face; however, they could benefit all users. For example, given participants' adverse reactions to social interactions, it is understandable that many requested better profile visibility controls. If implemented, visibility controls could help all users avoid inadvertent disclosure, not just closeted users. Similarly, restricting the ability to download or take screenshots of photos would help all users avoid extortion and unauthorized sharing.

4.3 Discussion

In this section, we briefly summarize our results and discuss changes dating services could adopt to improve the experience of closeted users.

Respondents expressed the most significant privacy-related concerns (RQ1) about non-dating websites purchasing anonymous information from dating sites, dating profiles being used to personalize advertisements, and the risk of encountering social relations while using dating services. A sizable minority of participants described some form of privacy harm they experienced while using a dating service. The most commonly described harms involved the unauthorized saving or sharing of photos sent to other users. Some users discussed being inadvertently forced out of the closet by encountering social relations on dating services. A few users described experiencing serious safety threats such as extortion, harassment, and stalking. Most participants discussed protecting their privacy in some way (RQ2), most often through limiting the amount of information and photos they share in their profile. Finally, users described a variety of features that may help improve their privacy (RQ3). The most common feature requests included enhanced profile visibility controls and limits on downloading or screenshotting photos. Some participants also expressed dissatisfaction with privacy-enhancing features being available only to premium users.

Closeted users' desire to prevent inadvertent discovery by social relations is the privacy need that most distinguishes them from the general population. While non-closeted users may not want some people to see their dating service profiles, this need is typically less stringent. Tinder's feature that allows users to block specific contacts preemptively is the best tool we observed for helping users to prevent inadvertent outing. We believe other services should adopt it. It does have limitations, however, as the blocked contact must register with the phone number or email provided for their profile to be recognized. There is also some risk that dating services may misuse or inadvertently leak the contact data that users share. While Tinder states it only keeps contact data for contacts users choose to block, users must still share their contact list to enable the feature.

Future work should investigate the feasibility of alternate blocking schemes. For example, a service might allow users to block users from a particular location (e.g., a college campus or neighborhood) from seeing their accounts. This feature would technically enforce the practice of some participants of avoiding dating services in specific locations. Services might also add the feature suggested by S102: the ability to block other users by group identifiers, such as place of work and educational institution. An anonymous notification feature, as requested by S11, may also be helpful, as it would allow users to receive notifications from a dating service without fear that these alerts would expose them to nearby users. None of these privacy features would be foolproof individually, but, used together, they may provide "defense in depth" against inadvertent exposure.

It is also worth considering whether privacy-protective behaviors themselves could inadvertently function as a targeting signal. A profile that omits a face photo, a real name, or a precise location may indicate that a user is closeted, potentially making that profile more conspicuous to an adversary seeking a vulnerable target. We did not observe this directly in our data, but it would be worth investigating in future work whether the very strategies that closeted users adopt to protect themselves also mark them as targets.

Participants' negative reactions to institutional data practices (e.g., dating services selling anonymized data and using profile information for targeted advertising) are similar to prior results with non-closeted populations [305]. The practice which participants reacted most negatively to—sales of profile information—was not disclosed in any of the privacy policies we reviewed and thus may not be a current practice. All the services use user data for advertising, a practice that most participants also disagreed with. While eliminating targeted advertising would reduce revenue, advertising represents a small portion of the revenue for the services we investigated. On a page describing their privacy policies, the Match Group states, "less than 3% of our revenue [comes] from advertising sales." [207] In Grindr's first-quarter 2023 financial filings, indirect revenue, which primarily consists of advertising revenue, made up only 13.8% of Grindr's total revenue [206]. We could not find specific numbers for Bumble, but their first-quarter 2023 financial disclosure states that subscriptions are their primary source of revenue [74]. Future work could investigate the extent to which specific types of online advertising (e.g., contextual advertising, retargeting) may be acceptable in the context of online dating services, among both general users and populations with privacy concerns.

The request to limit other users' ability to take screenshots or download images on mobile dating apps is technically feasible and unambiguously beneficial. Grindr already limits screenshots for images shared via an album, although this feature is limited for free users.[2] Services with image sharing could allow users to opt in to allowing a conversation partner to download or screenshot a photograph. Although this would not eliminate the risk of others stealing images (e.g., by taking a photograph of an image using another device), this could reduce the risk of unauthorized pornography and extortion.

Other feature suggestions, like user verification for Grindr, offer both positives and negatives. By allowing users to create profiles with no information, Grindr is uniquely suited to facilitating private, online interactions between strangers. At the same time, the ability to remain anonymous

[2]An album on Grindr is a private collection of photos that users can share with up to 5000 other users. As of publication in 2023, free users had limited access to albums [205].

makes it easier for fraudsters, spammers, and other malicious users to create accounts and evade bans. Adding a verification system to Grindr could allow some users to interact with others on the service more safely, but it may undermine others' ability to use the service anonymously. Alternatively, services might consider verifying users by some form of ID, such as a driver's license. This feature would make it easier for dating services to identify and exclude malicious actors from using their services. Additionally, if users trust the dating service and are not required to publicly share their ID information, this would provide a path for closeted users to be verified without exposing personal information to other users. ID-based age verification is increasingly used to restrict children's access to adult content online [411].

As with our proposal for more services to adopt contact-based blocking, ID-based verification comes with the risk of companies misusing or leaking ID data. Even if dating services do not handle ID data directly (e.g., by partnering with a service like `id.me`), users may still be uncomfortable disclosing such data for online dating. This form of ID-based verification would also not work for users with a mismatch between their legal name and gender and their preferred name and gender identity. Future work could investigate the extent to which this type of user verification may be acceptable to closeted users or others with high levels of privacy concern. Services could make verification optional and provide a profile badge for verified users. They could also offer an option for those who have been verified to see and be seen by only other verified users.

It is also unclear to what extent paid privacy features could be made free. Dating services are businesses, and they must withhold some features from free users to make their paid service attractive. Some limitations on free users, like the daily block limit on Grindr, should be eliminated for safety reasons. Other features, such as disappearing messages or incognito mode, may offer privacy benefits but are not necessarily essential. Users in our sample, like S66, reported using other services that provide messaging privacy features for free (e.g., Snapchat). Enabling these features for all users may help them avoid leaving the platform to communicate with others.

Only about a quarter of users who indicated that they pay for a premium subscription also agreed that the premium features are useful for protecting their privacy. Extrapolating from this admittedly small and biased sample, it may be possible to make the premium privacy features available for free without harming subscription numbers. Some closeted users may refrain from using premium features for fear that financial transactions will be viewed by others. For example, a "GOOGLE*Grindr LLC" charge could out a person due to the service's reputation as a dating service for LGBTQ+ people. If they cannot make privacy-preserving features available for free, dating services might consider billing under alternative names less likely to attract scrutiny.

Chapter 5

Investigating the Experiences of Teenage Cryptocurrency Users on Reddit

This chapter shifts its focus to minors and their use of cryptocurrencies, non-fungible tokens, and other crypto assets [180]. Over the last decade, these have become very popular investments, especially among younger people. An online survey of 2,872 users conducted at the end of 2022 found that cryptocurrencies were the most popular type of investment among investors aged 18 to 25 [92]. Moreover, many survey respondents began investing at an extremely young age, with 25% reporting that they began investing as a minor. This raises important security and privacy concerns: Private keys are difficult for even adult users to manage safely, and any mistakes can result in immediate financial loss [268]. Most cryptocurrencies are pseudonymous, meaning that once a person's wallet address is known, many, if not all of their past transactions can be identified [320]. Most notably for this thesis, the cryptocurrency world has historically been rife with scams and other financial crimes that defraud users [33, 223, 325, 494].

Crypto asset investment also presents considerable financial risks. Many speculative crypto assets feature the characteristics of "gamblified" investments as defined by Newall et al. [342]: it is difficult for most to profit reliably [122, 165, 192, 448], it is attractive to users who are susceptible to gambling [244], and it presents the allure of out-sized profits. Even more established cryptocurrencies, like Bitcoin, are highly volatile, with the price of Bitcoin peaking around $65,000 in November 2021, before dropping to a three-year low of around $15,000 just a year later.[1]

There is good reason to hypothesize that minor teenagers (aged 13 to 17 years old) may be more vulnerable to these risks of cryptocurrency. Teenagers have developing brains, generally showing higher risk-taking and sensation-seeking behavior that declines with age [278, pp. 528 – 530]. They may also be affected more by certain motivational factors, such as peer pressure [126] or influence from social media personalities [345], some of which have been implicated in illegally promoting crypto assets [33, 223]. Teens also generally face the contextual risk factor of *resource constraint*, including both limited financial means and limited technical/financial experience. As discussed in Chapter 3, scammers often appeal to users' need for money in order

This chapter is based on a paper originally published at the 2024 Symposium on Usable Privacy and Security that was co-authored with Hiba Hassan, Alexandra Nisenoff, Lorrie Faith Cranor, and Nicolas Christin [61]

[1] Historical prices obtained from `https://coinmarketcap.com/`

to lure victims, while a lack of technical proficiency may simultaneously lower teens' ability to spot scams.

The under-regulated nature of cryptocurrencies has also raised concerns [415] that teenagers may gain access to risky assets without parental oversight. While most exchanges require users to be 18 years of age or older to create an account,[2] it is possible that teenagers could invest via unregulated exchanges with absent or lax Know Your Customer (KYC) policies.[3] Indeed, this method is advocated by online blogs [26, 282].

Despite these risks, little empirical evidence exists about how minor teenagers use crypto assets and the harms they encounter while investing. To address this gap, we analyzed posts about cryptocurrency from teenagers on Reddit, a US-based social media platform and link aggregation website. First, we used a set of 48 keywords to identify 4,979 posts likely to be about cryptocurrency in six Reddit communities ("subreddits") used by teenagers (e.g., `/r/teenagers`, `/r/ApplyingToCollege`, etc.). We then performed inductive thematic analysis on a random subset of 1,676 posts, seeking to answer the following research questions:

1. **How do teenage users gain access to the crypto asset ecosystem?**
2. **What motivates teenagers to engage with crypto assets?**
3. **What types of harm do teenagers experience when using crypto assets?**

We ultimately identified 1,409 (84.0%) posts meaningfully discussing cryptocurrency. We found little evidence of widespread use of unregulated exchanges, with minors most often reporting using accounts in their parents' names to purchase cryptocurrencies. Some also received donations from other users or mined for cryptocurrency. Teenagers appear motivated to invest for profit, but some cited entertainment, ideological motivation, or an interest in technology. Finally, we identified many of the same harms that adult users of digital assets encountered, including investment losses, fraud victimization, and loss of keys. This chapter provides the first insight into teenagers' interactions with cryptocurrency and the harms they encounter.

5.1 Methods

We conducted an inductive thematic analysis of 1,676 posts containing keywords related to cryptocurrency from a set of Reddit communities used by English-speaking teenagers.

5.1.1 Dataset

Our study is based on posts and comments on Reddit. The platform is divided into thousands of user-created communities called "subreddits," which focus on particular topics (e.g., `/r/politics` for American politics) or identity groups (e.g., `/r/gaybros` for gay men). Users can submit posts that link to external websites or present original text and multimedia content. Users can vote to affect the visibility of posts, with an "upvote" boosting a post and a "downvote" lowering the post's visibility. Users can also comment on posts, with the order of

[2]For example, at the time of this study in 2023, the user agreement of the publicly traded cryptocurrency exchange, Coinbase, states that "To be eligible to use the Coinbase Services, you must be at least 18 years old." [118]

[3]KYC regulations vary by jurisdiction but generally require financial institutions to verify a user's identity to help prevent money laundering and other criminal transactions [183].

comments determined by a similar voting scheme. Reddit is a common data source for academic research, particularly in computer science [384].

Rather than collect data directly from Reddit, we used an archived copy of the Pushshift dataset [35] with posts and comments from June 2005 to December 2022 (inclusive).[4] The dataset also includes the content of some posts that were deleted by users or removed by moderators, making it more complete than the data available directly from Reddit. Pushshift is widely used for academic research on Reddit [35, 384].

Anyone 13 or older can register for Reddit, and an estimated 14% of American teenagers have used the platform [19]. For our content analysis, we focus on six communities that we assume are predominantly used by teenagers: `/r/teenagers`, `/r/ApplyingToCollege`, `/r/SAT`, `/r/ACT`, `/r/HighSchool`, and `/r/PSAT` (henceforth referred to collectively as the "teenage subreddits"). `/r/teenagers` describes itself as "the biggest community forum run by teenagers for teenagers," `/r/ApplyingToCollege` has over 1 million users and focuses on topics related to college admissions, with an emphasis on students enrolling directly out of high school. Similarly, `/r/SAT`, `/r/ACT`, and `/r/PSAT` are subreddits focused on the discussion of college admissions exams that are most frequently taken by teenagers. Finally, `/r/HighSchool` is a subreddit that includes a wide range of content related to the secondary school experience. Previous studies [47, 104, 459, 565] have considered users' participation in Reddit communities like `/r/teenagers` as an indication of a user being underage.

To identify posts that discussed cryptocurrency, we used a keyword-based sampling technique. Previous studies of Reddit have also used keyword-based sampling to study cryptocurrency [245, 365] as well as other topics [182, 551]. We first selected a heterogeneous set of keywords based on popular cryptocurrencies (e.g., Bitcoin, Ethereum, Dogecoin, etc.), cryptocurrency exchanges (e.g., Binance, FTX, etc.), interest-yielding services (e.g., Nexo, Blockfi, etc.), cryptocurrency gambling services (stake.com, Cloudbet, etc.), and concepts related to cryptocurrency (e.g., blockchain, decentralized, etc.). Our goal in selecting keywords was to identify a diverse set of posts about cryptocurrency with a low false positive rate. The complete list of keywords and their frequencies can be found in Appendix C.2.

We then collected all posts that contained one or more of these keywords in either the title or the body of the post. We initially collected 6,408 posts; however, as discussed in the next subsection, several keywords were removed during codebook development. The final dataset contained 4,979 posts (0.06% of the total posts on the teen subreddits). For the thematic analysis, we selected a random sample of 1,676 (33.7%) to review.[5] We also collected all the comments ($n = 3{,}738$) associated with the posts in the sample. Most posts (80.0% of the sample) included only a single unique keyword. The most common keyword was some variation of "NFT," with 680 posts (40.6% of the sample) including at least one instance of this word.

[4]In response to changes in Reddit's policies, the Pushshift API is no longer publicly available to all users. We maintained an archive of the data available prior to this change and used it in this study.

[5]This sample size was selected based on the sample necessary to achieve a 99% confidence level with a 2.5% margin of error.

5.1.2 Thematic Analysis

Thematic analysis was performed by two authors in multiple stages. To generate an initial codebook, the lead coder reviewed a subset of 100 posts from the 6,408 post dataset and performed inductive thematic analysis. The coder viewed the post title, post body, and all the comments associated with the post. If a post contained a link, the coder also considered this content if it was available.

Codes were selected to identify content relevant to our research questions and provide valuable categories for subsequent analysis. For example, we identified discussions of different types of behavior (trading, mining, using crypto for payment, etc.). We also identified different types of harm (e.g., monetary losses, crimes, and gambling). We also selected codes that characterized the kind of post (e.g., discussion of crypto-related news, joke/meme, question about crypto assets, etc.). Most posts had more than a single code assigned.

Some posts were recorded in the Pushshift dataset with the body text replaced by "[removed]" or "[deleted]." This indicates that, before Pushshift could collect the post, it was removed by moderators or deleted by the user who posted it. Often, it was still possible to assign a theme to these posts, although this analysis is necessarily less reliable. For example, a removed post titled "Artist Seed - The NFT project that births a metahuman through cryptoart" evidently promotes the Artist Seed project. Removed or deleted posts that could not be assigned a theme were coded as "Removed/Deleted (Ambiguous Content)." For example, a removed post with the title "Dogecoin" was given this code.

To help contextualize each post, we also viewed user and post flairs. Flairs are tags that can be added to a post or a user profile to indicate something about them. On `/r/teenagers`, post flairs are required to indicate the type of content (e.g., "Meme," "Art," "Discussion," etc.). User flairs on `/r/teenagers` are used to indicate a user's age. Flairs are unique to each subreddit. For example, on `/r/ApplyingToCollege`, user flairs are used to indicate a user's year in school (e.g., HS Freshman, College Senior, Graduate Student, etc.). When quoting users, we list their flair if available in Pushshift. All deleted posts and many removed posts had the author field replaced with the text "[deleted]" and the user flair removed. We are, therefore, unable to report age information for users with deleted posts.

During codebook development, some keywords were eliminated. For example, we realized that the keyword "ether" mostly resulted in false positives, as all posts returned by this keyword in the subsample were people misspelling the word "either." The keywords eliminated during this stage are listed in Appendix C.2 with the note that they were removed.

After developing the initial codebook, the lead coder and another author collaboratively coded the remainder of the posts. The process was broken into blocks of 10% to 20% of the sample. For each block, the coders independently reviewed the sample and assigned codes based on the definitions in the codebook. The coders then met to compare their analyses, discuss differences, and select a consensus code for each post. Additional codes were added as needed throughout this process. During coding, the coders also recorded memorable or archetypal posts to serve as examples for the paper.

After all the posts were assigned an initial set of codes, the coders reviewed every post a second time to gain additional insights into the nuance of particular codes and refine the initial categorization. For example, for posts that were categorized as "Working on a crypto project,"

the coders enumerated the specific types of projects that users described. During this secondary analysis, the coders changed the codes assigned originally to some posts as needed. The final codebook with examples for each code can be found in Appendix C.1.

5.1.3 Ethical Considerations

Our research is not human subjects research, as we relied on publicly available data.[6] We recognize that social media users may not expect their posts to be reviewed and analyzed as part of research [177]. Moreover, teenagers, as members of a vulnerable population, deserve special consideration to ensure their protection. We do not report the names of users in the dataset. We will not publicly post our curated dataset, although interested researchers may contact the corresponding author to request an archived copy.

5.1.4 Limitations

The generalizability of our results is necessarily limited by the structure and demographics of Reddit. Due to the Reddit voting scheme, posts and comments that are more broadly appreciated are more likely to be seen and interacted with. Reddit users are more likely male, with 20% of teen boys in the United States using Reddit as compared to only 10% of teen girls [19]. Moreover, we cannot verify the claims made by users, and some posts and comments may contain fabrications or exaggerations. The subreddits we investigated are not age-restricted, and some users are likely to be adults. Finally, we focused exclusively on English posts and users who do not speak English are, therefore, not represented.

5.2 Results

In this section, we describe the results from our qualitative analysis. We begin by giving an overview of the sample before discussing results relevant to each of our research questions.

5.2.1 Overview

Figure 5.1 shows the frequency of posts about cryptocurrency in our sample. To account for the varying number of posts per month, we normalized the number of posts each month by dividing by the total number of posts in the teen subreddits that month. The highest posting volume is associated with an increased discussion surrounding the Reddit NFTs, which were given away to many users in the second half of 2022 [306]. Only 267 (16.0%) of the posts in the random sample did not contain relevant content about cryptocurrency. Most of these posts (181) were false positives, which did not contain any mention of cryptocurrency. The rest of these posts (86) contained some mention of cryptocurrency, but in such a manner that they provided no useful information. For example, one user made a post in `/r/SAT` where they stated that they "want to apply to Computer engineering (preferring specialization in blockchain tech., cloud computing,

[6]Our study was submitted to our Institutional Review Board(s), which confirmed that human subjects research review was not necessary.

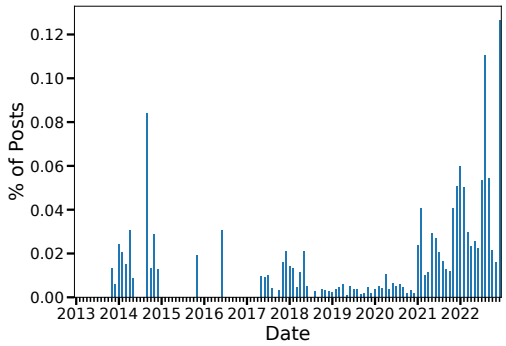


Figure 5.1: Frequency of posts in the random sample that contained relevant content each month, normalized by the number of posts in the teen subreddits per month. There were 8,803,440 posts on the teenage subreddits from January 1, 2013 to December 31, 2022. No posts in our sample were posted before 2013.

Table 5.1: An overview of the results of coding the sample. The percentage for "# in Sample" refers to the proportion of the entire sample. The percentage for all other columns refers to the proportion of posts from each subreddit.

Subreddit	**# in Sample** (%)	**# Irrelevant (%)**	**# Promotional (%)**	**# Removed/Deleted (%)**
`/r/teenagers`	1486 (88.7%)	204 (13.7%)	241 (16.2%)	497 (33.4%)
`/r/ApplyingToCollege`	109 (6.5%)	52 (47.8%)	10 (9.2%)	15 (13.8%)
`/r/Sat`	47 (2.8%)	2 (4.3%)	43 (91.5%)	42 (89.4%)
`/r/ACT`	15 (0.9%)	6 (40.0%)	7 (46.7%)	7 (46.7%)
`/r/APStudents`	11 (0.7%)	0 (0.0%)	3 (27.3%)	2 (18.2%)
`/r/highschool`	8 (0.5%)	3 (37.5%)	1 (12.5%)	2 (25.0%)
`/r/psat`	0 (0.0%)			
Total	1676 (100.0%)	267 (16.0%)	305 (18.2%)	565 (33.7%)

data tech.)" While this indicates that they have some interest in crypto-related technology, it does not indicate anything about whether or not they currently use cryptocurrencies.

1,027 of the 1,409 relevant posts in our sample (72.9%) were made by users without user flair or with a flair that did not indicate their age (e.g., "2 MILLION ATTENDEE"). The most common age flairs identified the author as under 18 (385 or 81.7% of posts with age flair). Only 78 of the relevant posts (17.4% of posts with age flair) indicated that a user was 18 or 19.[7] The remaining 4 relevant posts (0.8% of posts with age flair) were made by users with the flair "OLD," indicating that the user was 20 years old or older. These age distributions should be viewed skeptically, as they are sparse and user-provided. Users may forget to change their flair after a birthday or purposefully lie about their age. Still, this result supports our assumption that most of the discussion in our sample comes from users under the age of 18.

Table 5.1 provides an overview of the posts in our sample, broken down by subreddit. The overwhelming majority of posts in the sample were posted on `/r/teenagers` (88.7%). This

[7]For this analysis, we assume that users with flair "HS Senior" are 18. We assume that users with flair indicating another class year (E.g., "HS Junior" are under 18.

result is in line with the distribution of posts in the PushShift dataset, as /r/teenagers is by far the largest subreddit we examined. /r/PSAT contained no posts containing any of the cryptocurrency-related keywords. Across the entire sample, 33.7% of the posts were removed by moderators or were deleted by users prior to collection by PushShift. 67 of these posts (11.9%) had no code assigned (i.e., the probable content was ambiguous). The remaining 498 posts (88.1%) were assigned one or more codes, albeit at a lower confidence level than others. When reporting the frequency of codes in Appendix C.1, we specify the number that had removed or deleted content.

Table 5.2 shows the frequency of codes in our dataset. The most common post type in the sample were posts that promoted a specific project or service (305 instances or 18.2%). These posts seemed fairly ineffective: 228 (74.8%) were removed or deleted, and most (200 or 66.6%) received no comments. Many of these posts advertised the same projects repeatedly, often using identical or near-identical text. The most commonly promoted projects were Pi Network (8.9% of promotional posts), Ethereum Name Service (7.8% of promotional posts), and Dogecoin (7.2% of promotional posts). Another common feature of these posts was some sort of giveaway or sign-up bonus. For example, many posts promoting the Ethereum Name Service stated "Ethereum Name Service ($ENS) is Airdropping Tokens worth up to 5000$ for the first 1000 People To Claim it." Airdrops are free distributions of digital assets given away to promote a project [4]. When promotional posts were not removed right away, users often reacted negatively. For example, a post promoting Pi Network in /r/teenagers in August 2019 received negative replies from 5 different users, many calling it a scam (e.g., "That's 100% scam. Watch out").

As one might expect from a social media website, jokes and meme posts were also common. 254 posts (15.2%) were coded as being a joke or meme, mostly because they were explicitly labeled as such (i.e., using post flair). An additional 52 posts (3.1%) were tagged as probable sarcasm. We used this code when a post was not labeled as a joke but seemed so absurd as to be unbelievable or was otherwise clearly intended to be a joke. For example, one deleted post from 2019 to /r/teenagers stated, "$20 for a pic of my belly PM me accepting Bitcoin only (pic unrelated)." This post is unlikely to represent a genuine attempt at selling photos for Bitcoin. There was a great heterogeneity among the different types of jokes. However, a common theme seemed to be criticizing the intangibility of cryptocurrencies in general and NFTs in particular. For example, many users made jokes about the ability to "screenshot' or "right-click and save" NFTs. When we coded a post with "joke or meme" or "sarcasm," we refrained from applying any other codes to that post since we assumed the post does not reflect actual behavior.

Many of the relevant, non-joke posts were uninformative for our research questions. For example, 69 posts (4.1%) discussed news related to crypto assets, such as the failure of the exchange FTX. While these posts demonstrate that users were interested in discussing crypto, they lacked information about users' experience with crypto assets.

Posts discussed a range of interactions with crypto assets, including obtaining/holding crypto assets (226 posts or 13.5%), mining cryptocurrencies (59 posts or 3.5%), selling or otherwise transferring crypto assets to others (30 posts or 1.8%), short term trading (25 posts or 1.5%), using cryptocurrencies for payment (25 posts or 1.5%), and gambling with crypto assets (13 posts or 0.8%). It is important to note that not all of these posts described actual, current behavior. For example, 10.2% (23) of the posts tagged as "discussion of obtaining/holding crypto" discussed a

user's desire to obtain a crypto asset and 5.8% (13) conveyed a question about obtaining crypto assets.

Table 5.2: Most common codes assigned to posts in the dataset. Complete definitions and examples can be found in appendix C.1

Code	**Frequency (%)**
Explicit promotion of project	305 (18.2%)
Joke or meme	254 (15.2%)
Discussion of obtaining/holding crypto	226 (13.5%)
Irrelevant	181 (10.8%)
Reddit NFT	156 (9.3%)
Giveaway	145 (8.7%)
Criticism of crypto	110 (6.6%)
Subjective question about crypto	105 (6.3%)
Irrelevant, Crypto mentioned in passing	86 (5.1%)
Discussion of news in crypto	69 (4.1%)

5.2.2 RQ1: How do teenage users access crypto assets?

The most common way for teenagers in our sample to gain access to cryptocurrency services was apparently by creating accounts in the name of their parents or another trusted adult. While most posts about obtaining crypto assets did not explicitly address how the user accessed the crypto ecosystem, we observed users discussing this when they asked others how they could access crypto assets. For example, in November 2021, one user posted a thread to `/r/teenagers` titled "Are you able to get into the stock market and cryptocurrency as a teenager." The body of the post stated "I wanna buy one of those funny looking monkey pictures but I'm clueless on how it works." One 17-year old user replied "you can have your parents set up your account." A different 17-year old user stated "I trade crypto under my moms name but to my knowledge you have to be 18 at least where I live." In response to a deleted post titled "Had to sell a lot of my crypto," made to `/r/teenagers` in November 2021, one user wrote "Dude I am sorry... can you tell me why you had to sell it, how you got in, and where do you sell your crypto (I've been tring get in for a while but i dont know how)." A 15-year-old user replied, writing, "I'm not OP but for crypto I would use Binance. It's an online exchange with pretty low fees. You'd need your parents' permission though, since it requires you to be over 18 and provide a drivers license..." Similarly, in reply to a post promoting Bitcoin in February 2021, one user commented "I wanna use Blockfi and earn interest but I'm not 18 :(((((((." A 15-year user replied with advice, stating "If you have a parent, you could do it with their account." Purchasing cryptocurrency through a parent's account is likely the safest way for teenagers to access the crypto markets, assuming they supervise their child's investments. This dependence, discussed further in Section 5.3, may also introduce risks.

Users also discussed mining cryptocurrency to use it for speculation. For example, one 17-year-old user made a post to `/r/teenagers` in March 2021 titled "Decided to start mining

for eth and made some btc out of it. Time to hold and watch it raise!" The image attached to the post showed $100 worth of Bitcoin in a wallet. Mining refers to the process of generating cryptocurrency by running a computer that validates transactions on a blockchain. Once mined, cryptocurrency could be transferred to an exchange and traded for other cryptocurrencies. This is presumably what the user meant when they said they were "mining for eth and made some btc." If an underage user were to use a decentralized exchange or other service that does not deal in US dollars or other fiat currencies, they could avoid age restrictions and KYC regulations. Another user more explicitly described this process in a post to `/r/ApplyingToCollege` in July 2017: "Recently I started to get engaged with the cryptocurrency market. Ive been building multiple mining rigs and buying/sellimg the currencies in an attempt to make money. Would this count as an extracurricular?" Throughout our analysis, we encountered many posts similar to this, where users described some kind of involvement in crypto to see if it was worth discussing in their college applications.

Mining was also recommended to some users who inquired about obtaining cryptocurrency as a minor. For example, a 17-year-old user posted a thread to `/r/teenagers` titled "Is there anyway to buy crypto without being 18." The body of the thread stated "Since i cant participate in stonks without being 18 crypto i legally can, is there any service that doesnt ask for age verification?" Along with several replies recommending that the user get help from his parents or another adult, one 16-year-old user replied that they could "GPU mine for a while... and trade with that."

Some users also gained cryptocurrency via gifts from other users on Reddit. For example, in May 2021, a 15-year-old user posted to `/r/teenagers` "thank you to whoever tipped me dogecoin like a few months back." The body of the post added "i managed to reinvest in crypto and now have £290 from like £40" We also found a few examples of threads where users gave away small amounts of cryptocurrency to users who commented in `/r/teenagers`. In January, 2014 a 16-year-old made a post titled "[Other]I'm back with another bitcoin giveaway!" The poster used the `/u/changetip`[8] bot to gift users fractional amounts of bitcoin. The post received over 900 comments. Similarly, in September 2014, a user made a post to `/r/teenagers` with the title "Free Bitcoin - Just Comment." The thread also received hundreds of comments.

This kind of post was rare. Most of the threads tagged with the "giveaway" code were promotional posts advertising some sort of airdrop or sign-up bonus given out by a service. All of the peer-to-peer giveaways occurred in 2014. Users directly donating to others could be motivated by a sense of philanthropy; however, donating small amounts of cryptocurrency directly to other users is also a good way to introduce new users to the ecosystem. Indeed, the poster in the January 2014 giveaway explicitly stated this, writing, "...I am doing this giveaway to get my peers involved in bitcoin. I believe in bitcoin as a currency, and not just an investment, and I think it could benefit other teens..." As Bitcoin became more mainstream, it became less necessary to introduce users through free giveaways.

Some teenagers created their own crypto projects. For example, one user made a post to `/r/ApplyingToCollege` in August 2018 where they asked if they should update colleges

[8]This bot is now defunct. While operating, it allowed users to gift cryptocurrency using commands in Reddit comments [226]

about a cryptocurrency project they launched: "So I just recently finished making my own crypit currency with a relatively high market cap of 100k ish. I didn't mention this on my app at all (besides interest in crypto) because I didn't think I would get it done in time, but I did. Would it be worth emailing my colleges about this new stage in my life..." The open-source nature of major cryptocurrencies makes it relatively trivial to fork an existing project, tweak some properties, and give the currency a new name. Dogecoin, for example, began as a fork of an existing project (LuckyCoin) [109].

We found little evidence for teenagers using unregulated exchanges to purchase cryptocurrency. We only found one comment or post where a user stated that they used an unregulated exchange to avoid "know-your-customer" (KYC) practices. In reply to a comment expressing frustration about being unable to purchase cryptocurrency, one user in mid 2021 stated "Lol I'm a minor and I bought Bitcoin using a non kyc site, and a vpn..." Using a VPN could allow users to avoid KYC requirements and other regulations that are only enforced for users who appear to be in a particular jurisdiction. Some exchanges, such as Binance International, were complicit in helping US-based users avoid geo-blocking by recommending the use of a VPN [303]. We did find another comment where a user recommended Binance International in a discussion where a different user shared that they were interested in getting started in cryptocurrency: "yeah go for it man, hmu if U need help, im using binance international." This user did not indicate that they were under 18 nor did they state that they were using Binance International to avoid regulations.

There may be more users who used unregulated exchanges but did not discuss it for various reasons (e.g., they may not want to admit to trying to circumvent exchange rules). Our keyword-based sampling approach may have also prevented us from finding more examples. Still, finding only a single example of the use of an unregulated crypto exchange suggests that the practice is less common than the other ways minors may obtain crypto assets.

5.2.3 RQ2: What motivates teenagers to engage with crypto assets?

A desire to realize relatively large, short-term profits seemed to be the most common motive for teens acquiring crypto assets. The most valuable insight into users' motives came from 23 of the posts coded with "discussion of obtaining/holding crypto" where users discussed their intention or desire to purchase crypto assets. The potential for large, relatively short-term gains was commonly brought up. For example, one 14-year-old user made a post to `/r/teenagers` in March 2021, stating, "Hey guys so my parents are giving me 100 dollars to invest in whatever I want... I think a good crypto to buy is Bitcoin... I've been tracking it... it is fairly stable and usually doesn't go down a lot and since by 2030 my investment of 100 will grow ten fold if I buy it now." A different 14-year-old user made a post to `/r/teenagers` in June 2021, writing "I'm becoming obsessed. I wanna transfer some of my personal money into Bitcoin so I can have enough for a car when I'm 17 in 3 years. Anyone else involved at all in crypto?" In reply, several users recommended against trading crypto. However, a 15-year-old user shared an anecdote that might reinforce the sense of easy profits: "I put $500 on doge when it was 0.04 and when it hit 0.72 it became $5,000."

Users actively engaged in crypto asset investment discussed how profitable it seemed to be compared to other ways to make money. For example, in August 2021, a user made a deleted post

to /r/teenagers titled "Lmao. Trading stocks seems like a scam when I see shit like this. With Crypto trading I've seen 50-100% profits in literally minutes. Still learning but starting with small amounts. Like 18/20 of the trades I made so far were at least 50% profit." The post linked to a picture of text explaining that day trading stocks are typically unprofitable. Similarly, in November 2021, a user made a post to /r/teenagers titled "If Shiba Inu lists on Robinhood I'll have enough money to move out of my parents house." The body of the post stated "And yes that's my only hope... because dog crypto coins have made me more profit in a week than minimum wage pays in 8 months."

Short-term profit potential was also frequently emphasized in posts promoting crypto asset investing. For example, one user wrote a post to /r/teenagers in January 2021, stating "My advice to everyone here is to invest, whether it be in cryptocurrency or the stock market... Cryptocurrency (like bitcoin) is more wild and fluctuating than the stock market, so it's riskier... If you do it right, a measly $1K can pay your entire college fund in a few years." Similarly, an 18-year-old user wrote in a post to /r/teenagers in May 2021, "...I have reason to believe that Safemoon is going places... this crypto could rocket up, and if you invest now you could make a good profit if it does get anywhere..." This example is particularly disturbing, as the company and founders of SafeMoon were subsequently charged with defrauding investors [513].

Interestingly, several users described being pushed by their parents to develop NFTs based on a desire for out-sized profits. For example, a 14-year-old user made a post to /r/teenagers in July 2022 titled "Any Crypto Teens here to answer this?" The body of the post shared "I do art as a hobby, and my dad's been telling me to turn them into NFTS. I've refused twice, but on the third time, he told me that I'm missing out on the chance to make him and I millionaires. Do I go for it or not?" Similarly, a 14-year-old user made a post to /r/teenagers titled "So uh." in January 2022. The body of the post stated "My parents are asking me to make NFTs and sell them because 'you have some knowledge about it, why not go and make some money like him...'" The post linked to an article about a computer science student who made millions of dollars from selling an NFT collection [132].

Users and their parents' desire to profit off crypto assets is generally misguided. While some popular NFT projects are very profitable, most make little to no money [355]. The story is similar in cryptocurrency markets. Most retail investors sold at a loss following the collapse in cryptocurrency prices in 2022 [122, 165], and many fall prey to institutional investors [448]. Short-term profits may trick users into feeling that they are skillful or have discovered a successful strategy, however, the ambiguity of the crypto asset market makes it difficult for traders to consistently profit over time [192]. Crypto asset trading may appeal particularly to teens due to their limited wealth and low incomes. A modest gain from buying and holding traditional equities may seem trivial to a person who only has a few hundred dollars to invest.

Some users discussed crypto investing as a form of obsessive entertainment akin to gambling. For example, one user made a post to /r/teenagers in November 2021 titled "I invested like $100 into crypto and [can't] stop looking at the charts and stuff." In response to a commenter who stated that "The market is too volatile imo," the original poster wrote "That makes it fun." A 17-year-old user replied to the thread, sharing the same experience: "I got interested in it a year ago and I couldn't stop looking at charts even during online classes for months." Another user shared that they went on a "...wild rally of not sleeping and keeping my eye on the

charts." Increased checking of markets has previously been associated with investment in crypto assets [210].

User justifications for purchasing high-risk assets also mentioned the "fun" of the activity alongside the potential for a high reward. For example, in July 2021, a user made a post to `/r/teenagers` titled "i invested $1000 into a dog meme crypto haha." In the body, they explain "...I saw a lot of videos about it on tiktok and want to make some money so decided to stick $1,000 into a cryptocurrency meme coin called hoge lol... whether I become a millionare or make nothing it would be a fun experiment." In the comments, they stated they were under 18. Similarly, in response to a post promoting a project called PhunWallet in January 2021, one 17-year-old user wrote, "im interested, just for fun yk... I'm more into well established cryptos like btc,eth,sol,doge,aval,xrp,xlm,etc etc" The original poster replied, stating, "Yeah this is a start up, costs nothing to get into this... maybe in 5 years it could be worth something."

The comparison with gambling is not just inferred. As Johnson et al. [245] observed, users in our sample directly compared the experience of crypto trading to gambling. For example, an 18-year-old user wrote in a post to `/r/ApplyingToCollege` in "...I got into investing around that time as well, and I flipped $10 to around $2.7k after multiple trades. It was honestly just gambling..." Similarly, a 17-year-old user responding to a critique of the riskiness of crypto assets wrote "u see risk = profits... cryptos are like gambling rn also its the best way to pay and receive money without paying extra because of tax... anyways if it gets me money i'mma do it."

The design of real-time trading services may contribute to the entertainment appeal of crypto asset investing. The continuous feedback and speed of transactions facilitated by these services can make them more engaging. Additionally, so-called "gamification" techniques (e.g., investing leaderboards, rewards points, etc.) are commonly used in trading services [388]. These have been shown to increase the frequency of trading, potentially leading to harm [49].

We also observed motives for engaging with crypto assets besides profit. A few users shared ideological justifications for using cryptocurrency, especially in posts promoting cryptocurrency use in general. 35 (2.1%) posts were coded as "General promotion of crypto." For example, one user made a post to `/r/teenagers` in February 2022 titled "Dudes fr never gave a single fuck about the environment until they want to use that as an excuse to hate crypto." The body of the post read "It's so ignorant too, banks use more energy and cause a lot more pollution than crypto. We're trying to phase out evil and corrupt banks, we can't do that without using any fucking power..." This kind of justification for the use of cryptocurrency has its roots in the first posts from the anonymous creator of Bitcoin who expressed a deep distrust of the conventional financial system [336].

A few posts, particularly in `/r/ApplyingToCollege`, discussed crypto assets in relation to a more general interest in technology. For example, one user wrote in a post in July 2018, "I kind of have a passion for home automation/raspberry pi and arduino projects. Using these mini computers (running python and c++) I've made my own automatic door lock, a weather sensor and display, ... and a bot that manages and displays crypto prices." Another user made a post in October 2022, describing how they were the co-founder of their school's "Intel-Tech Club" in 11th grade. They explained that, as a leader in the club, they "...hosted Fintech, Business sessions; conducted bimonthly contests; hosted seminars on AI, Blockchain, ML, AR/VR, etc." Ultimately, this kind of purely technical interest in crypto assets was rarely discussed.

The Reddit collectible avatar NFTs introduced many users to crypto assets. The Reddit NFT was by far the most discussed crypto asset in our sample. 156 posts (9.3%) were coded as being about the Reddit NFT. Moreover, 99 (43.8%) of the posts coded as "discussion of obtaining/holding crypto" were about the Reddit NFT. Bitcoin (26 posts or 11.5%) and Dogecoin (17 posts or 7.5%) were the next most popular crypto assets that users discussed holding.

Reddit NFTs are hosted on the Polygon blockchain, and allow holders to uniquely customize their Reddit avatar. The NFT was given out for free to some on Reddit [306] in late 2022. In the wake of this giveaway, many users discussed receiving the avatar on `/r/teenagers`, with dozens of posts featuring images of the avatars titled things like "I got one of those free NFT avatars." and "I got random free NFT avatar while surfing the home page :/ and it is weird af." Some users reacted negatively, particularly to the idea of the avatars being NFTs. For example, one user made a post in August 2022, stating "I just got a fucking reddit nft avatar and I feel like a shitty crypto bro."

Unlike other crypto assets, users interested in purchasing a Reddit NFT discussed their aesthetic value. For example, a 15-year-old user wrote a post titled "am i the only one who's tempted to buy a reddit nft." By way of explanation, they commented "GUYS THEY LOOK COOL I DONT HAVE REDDIT PREMIUM I CANT MAKE MY [AVATAR] COOL LIKE URS." Similarly, a 13-year-old user made a deleted post to `/r/teenagers` in September 2022, writing "I want Reddit nft." In response to another user they explained "I want the reddit avatar items that come with the nft." These posts suggest that the Reddit NFT may be understood more like a cosmetic item in a video game rather than a financial asset.

5.2.4 RQ3: What types of harm do teenagers experience when using crypto assets?

Users discussed a variety of harms, including fraud victimization, wallet loss, and financial losses from poor investments. 27 posts (1.6%) were coded as including "crime," including crimes perpetrated by users and crimes victimizing users. Some crimes were only tangentially related to cryptocurrency. For example, several users discussed accounts being compromised to post spam related to crypto assets. This type of anecdote appears in our dataset because of the inclusion of crypto-keywords, however, this example of abuse is not related to the use of crypto assets.

The most common type of crime that victimized users was extortion facilitated by cryptocurrency. For example, one teen made a post titled "What do I do?" to `/r/teenagers` in 2020 asking for advice on a "sextortion" message he received: "I just got an email about something very weird. Someone emailed me some with the name of my password... it said that if I did [not] pay this certain address $2000 dollars in bitcoin it would send 3 random people in my contact a video of me wanking. Idk if this is real or just a scam to get money out of me. I am only 14 and don't know what to do. Please help me..." Thankfully, this user's post received several comments reassuring them that the extortion attempt could be safely ignored. For example, one 14-year-old user wrote "Lol it's a scam. Everyone and their mother has gotten that e-mail. You should change your password tho because it was leaked..." Bitcoin and other cryptocurrencies are often used to facilitate extortion and other scams [494], as they are pseudonymous and lack payment-reversal mechanisms such as credit card chargebacks [336]. This kind of support is

explored in greater depth in Chapter 7, which examines scam-related help-seeking across several Reddit communities.

Technical aspects of cryptocurrency were rarely discussed; however, we observed issues with key management. The giveaway posts (see subSection 5.2.2) featured the most discussion of wallets and key management, as users needed to be on-boarded into the ecosystem to receive their gift. Most of this discussion was vague and non-specific, providing little insight into users' key management practices.

Key loss was mentioned in a small number of posts and comments in the dataset. For example, a 15-year-old wrote in a post to `/r/teenagers` in January 2021 "I had $0.60 or whatever in ETH a long time ago, now it's worth $18 and I can't find the private key :(But i did find the key to an account with $11 of bitcoin so that's cool." This comment is typical of what teenage users described when discussing key loss: small amounts of forgotten money that became valuable in retrospect. More often than dedicated posts about key loss, users mentioned losing wallets in unrelated contexts. For example, a 16-year-old wrote in a giveaway thread, "About a year or so ago, I decided to try bitcoin mining, but my computer was pretty lame... I mined like one whole dollar! I wish I still had that hard drive :'(" This type of user error leading to financial harm is similar to that discussed by Krombholz et al. [268].

Users also shared experiences with investment and trading losses. More posts were coded as discussing profit (51 or 3.0%) than loss (17 or 1.0%). This likely does not reflect the overall frequency of investment losses relative to profits, as users may be too ashamed to discuss their losses. A few users shared extreme examples of loss. For example, a post titled "Screenshot of my crypto portfolio in October vs what it is now" in `/r/teenagers` in July of 2022 showed images of $18,000 in value dropping to about $100. Only the value of the assets was shown in the screenshots, so it is difficult to tell what type of investments this user made or if they engaged in active trading. A 17-year-old user shared an experience of making a huge gain on paper, before losing most of it. The post was made to `/r/teenagers` in April 2022 with the title "I made $160K on crypto and my social anxiety dissapeared. Then I lost it and It's back but worse lmao." They explained in the body of the post that "...i made a high risk bet that paid off... I was worth 160k for like a couple weeks and i felt like the fucking man, my social anxiety literally crumbled i was so confident..." In the comments, they explained that they started with a few hundred dollars and provided a screenshot of their portfolio, which consisted of a handful of obscure cryptocurrencies.

Most losses from investment were much smaller than shown in these examples. For example, on `/r/teenagers` in December 2021, one 16-year-old user wrote a post titled "I bought yesterday shib for 50$ and it dipped in the night." The body of the post added "Pain. And btc also dipped from 170 I had mined to 140. Ehh just remember HODL."[9] In response, a 14-year-old user shared "my portfolio literally went from $9.5k to $7.5k this week lol." These losses are the inevitable flip side of engaging in high-risk investing for easy, short-term profits, particularly in hard-to-predict markets like those for crypto assets. While some users will be lucky enough to pick the right coin and sell it at the right time, others will experience losses. While many of

[9]The word "HODL" (a joking misspelling of the word hold) alongside the phrase "diamond hands" are commonly used in crypto communities to encourage users to hold onto assets, regardless of price activity [193].

the losses we observed are small in absolute terms, they may represent a large proportion of a teenager's total wealth.

We found several examples of teenagers losing money to pump-and-dump schemes. A pump-and-dump is where an investor or group of investors drives up the price of an asset through false or misleading statements, hype, or other manipulative tactics (the "pump"). Once the price has been inflated to a certain level, the perpetrators sell their holdings (the "dump"), causing the price to plummet and leaving many investors with losses. Pump and dump schemes are illegal in most jurisdictions. In a thread on `/r/teenagers` in April 2014, a 16-year-old user commented. "Want to know the struggle? I lost 47k on paper with cryptocurrencies. Yes, USD." In response to a reply, they elaborated "So, when I was dumb and gullible in the world of crypto earlier this year, I would try to ride these pump and dumps that people on twitter/IRC would do..." They explained the concept of a pump and dump before concluding with the statement "I was a bag holder in a twitter pump and dump of BlackCoin since early February. I had 180k of them. BlackCoin reached between 0.0008 and 0.009 at one point and currently hovers around 0.0005. If I had sold at 0.0008, I would have had 144 btc, a 7200% ROI. Actually looking at it now, that'd be about 72k USD. Instead I sold at 0.00008, taking a near 60% loss." Their comment suggests that they participated in multiple pump-and-dump schemes before the loss they discussed.

Another thread posted to `/r/teenagers` in February 2021 also describes a pump-and-dump, although the user seems to understand less about what occurred. In the body of the post, they explain that they participated in a pump and dump: "Today a thing called pump(a lot of people buy a coin at the same time so the price go up and the you sell it have like 200% of profit) was made and i said ok this Is my moment... When the name of the coin was released I run to my investing app and spend 20€ in that coin... I wait to the price to raised and then I sell it. I don't no how but I finished losing money." This user seems to have participated in a pump and dump group similar to those described in Nizzoli et al. [347] but without the awareness that they risked losing their investment. These examples highlight the risk of uneducated and inexperienced teenagers participating in under-regulated markets.

One user claimed to have taken a leadership role in market manipulation. Using a one-time-use or "throwaway" account, they posted a thread to `/r/ApplyingToCollege` titled "How should I describe a crypto extracurricular?" requesting advice on how to discuss running a pump-and-dump group on college applications. The body of the post stated "I run a discord server... that has a lot of members and can move prices on obscure crypto (think <15 million market cap). I started with a small sum of ∼$40,000 from previous investments and small gifts, and have made nearly 100 times profit. I spend nearly all of my spare time doing this, so my grade in my foreign language dropped (from A to B) and I don't have any other ECs to put on my application other than really trivial things." Teenagers have previously been identified as perpetrators in various types of cybercrimes, including high-profile security breaches (e.g., "MafiaBoy" [224]). Leading a crypto pump-and-dump requires no significant resources or special skill, so while this description is concerning, it is unsurprising that a teen might engage in this kind of unethical and likely illegal behavior. The poster, however, insists their actions were legal, stating, "I cleared it with my parents and other family friends who are securities lawyers..."

Teenage users also described participating in cryptocurrency gambling. 13 posts (0.8%) were coded as containing content about gambling, with 4 of these posts promoting a service. These are distinct from posts where users described their trading behavior as analogous to

gambling (see Section 5.2.3), as they involve users discussing unregulated, online crypto casinos. These include centralized services and decentralized gambling applications hosted on a blockchain [71, 322]. In the US and Europe, the legal minimum gambling ages range from 18 to 25 [44], so minor teenagers are not legally permitted to gamble.

The most concerning example was a user who discussed what they describe as an addiction to both trading and gambling in a post to `/r/ApplyingToCollege` in November 2022. The post asked about how they should discuss cryptocurrency in their college applications. In the relevant part, they stated "...I did crypto/nfts/skins reselling during Quarantine. Most of my revenue came from altcoins mooning, such as Ada and Solana... Probably made over $250k+ in revenue (did this probably like 5-8 hours a day on average, I was fucking addicted)... Only came out with like $3k in profit... I was a fucking gambling addict. I was addicted to Roobet, Bustabit, and Stake.com... I deeply regret doing crypto, it was such a mentally draining and useless thing..." Some users replied with similar experiences. For example, one high school senior wrote "i also had an addiction, crypto makes it way too easy to lose it all lol. glad to hear you're better..." Another user wrote "for reference, i made (roughly) 2 million on the crypto spike during 2020-2021 i am also a stake addict $\tilde{\$}$450k gambled i did also start a sports betting server and get involved with certain things i would rather not discuss, but overall i actually ended up profiting from gambling..."

Most examples of users discussing gambling were less extreme. For example, an 18-year-old user made a post in `/r/teenagers`, stating "Hey guys, I got into Dogecoin a while back and I've been able to gamble my way up for about $5 from mining, to $30... I've since lost all of that and now I'm sitting at about $0.30 in doge, and I'm sad about it. I... will try to make the money back from the same gambling process with tips made on reddit..." It is startling that this user lost almost all their cryptocurrency gambling, but still felt that it was a reasonable way to make money. Incidents like this suggest that increased regulation in this space may be needed to protect teenagers and adults.

5.3 Discussion

This section discusses the implications of our results, particularly in contrast to adult populations. We also discuss ways to better protect users from the risks of crypto assets and the potential for future work investigating teen crypto users.

Underage teenagers on Reddit seem to most often gain access to crypto assets with the help of parents or other trusted adults rather than international services or decentralized exchanges, as was feared by Ryan [415] (RQ1). Teens' motivations for using crypto assets focus on profit, similar to adults (RQ2), as long-term investment, trading, or other forms of financial speculation are the most commonly reported uses of crypto assets in adult populations [458]. Parental pressure was a unique motivational factor we noted in several posts; however, this seems rooted in parents' sense that crypto could be highly profitable. Engaging in speculative investing and trading with crypto assets during a formative period may teach bad habits that could harm teens future financial success. We also identified many of the same harms experienced by adults [268], including financial loss due to poor key management, speculative investing, and even fraud (RQ3).

Underage teens' method of accessing crypto assets seems to be the primary factor distinguishing them from adult users. While using an exchange account in the name of a parent may allow for some level of parental supervision, exchanges such as Coinbase and Binance do not support traditional "custodial accounts" [475] that are explicitly designed to enable parental oversight of finances. Implementing account types specifically designed for teens, perhaps with limited access to riskier assets, could promote safer crypto investment behavior.

At the same time, not all parents are equally wise or benevolent. Minors' dependence on their parents, termed *reliance on a third party* within Warford et al.'s framework, can place them at greater risk, such as in cases of parental identity theft [341]. A parent who is a poor financial decision-maker, or who exerts control over an account nominally opened for their child's benefit, may introduce risk rather than mitigate it. Although this was not observed in our sample, whenever a minor's assets are held in an account controlled by an adult, that adult has the practical ability to use, redirect, or withhold those assets.

This dependence is not inherent to crypto-assets, however. Some teens in our sample avoided reliance on a parent's account altogether by acquiring cryptocurrency through other means, such as mining, peer gifts, or converting proceeds from online sales, none of which require an adult intermediary. A follow-up survey study by Klucinec et al. both confirmed the common practice of adults buying cryptocurrency for minors and highlighted how crypto assets fill gaps in existing payment systems by allowing teens to send/receive money online without the ability to create accounts with banks or money transmitters (e.g., PayPal) [261].

Increased regulation of crypto assets may help protect both underage teenagers and adults in crypto markets. Some types of crypto assets may be regulated under existing laws governing securities and commodities. For example, in 2023, the US Securities and Exchange Commission (SEC) charged multiple crypto companies for operating unregistered securities exchanges, including Kraken [512] and Coinbase [510]. Forcing crypto assets to register as securities could increase transparency and allow greater market surveillance to prevent fraud. It would also promote legal clarity by unambiguously defining market manipulation, including pump-and-dump schemes, as illegal. However, treating crypto assets like traditional financial assets is far from a one-size-fits-all approach. As illustrated by the Reddit NFT example, not all crypto assets are purchased for profit, and some may be more akin to collectibles, which are generally not treated as securities or commodities. Moreover, some crypto assets purchased for their profit potential do not fit the classic definition of securities [221].

Rather than attempt to regulate unbacked crypto assets as financial services, some argue that crypto trading should be treated like an illicit activity. For example, a 2022 report from the Treasury Committee of the UK Parliament [119] recommended that crypto assets be regulated as gambling, in part, to prevent a potential "halo" effect that may result from them being considered regulated investments. Presumably, a gambling-based regulatory framework for crypto assets would make it illegal to allow minors to engage in crypto asset trading. It is unclear how effective this type of regulation could be, considering that teens already manage to avoid age restrictions on gambling.

Restrictions on gamification and other practices meant to increase investor activity may help to prevent teens from engaging in harmful trading practices. In July 2023, the US SEC announced proposed rules that require online brokers to consider conflicts of interest related to attempts to increase user activity [509, 511]. Design elements that are adopted to increase trad-

ing activity based on data analytics could be covered by this rule. It is unclear whether this rule would apply to design elements that were adopted purely to increase entertainment. Regardless, real-time trading services should adopt positive design principles, such as nudges to diversify investments [97], to help users make better financial decisions.

Future work should explore teenage experiences with crypto assets more directly using human-subjects research. While we have confirmed that teens participate in the crypto markets, quantitative studies could determine the frequency of underage market participation and the various harms we identified. Qualitative studies could be employed to probe teens' motivations for investing in crypto and more deeply explore individual experiences. Such work is essential to further protect vulnerable populations like teenagers.

Chapter 6

Investigating the Susceptibility of Teens and Adults to YouTube Giveaway Scams

While Chapter 5 examined teenagers' behavior through observational evidence, this chapter presents a direct study of teens' engagement with "giveaway" scams. As introduced in Chapter 2, in these scams, fraudsters create websites that purport to offer free goods or services (e.g., iPads, gift cards, mobile game currency, access to movies, etc.) and are often promoted through social media advertisements [114], YouTube videos [54, 111], hijacked websites [78], and other means [28]. On such websites, victims are directed to complete surveys or other tasks, such as downloading mobile apps, but ultimately receive no compensation. Although this scam often does not cause direct financial harm, victims waste their time and may end up giving their sensitive personal information to fraudsters. Moreover, the frequent use of mobile game currency as a lure may make teenagers (minors aged 13 to 17) particularly susceptible to these schemes [28, 54]. Teens' resource constraints, discussed in Chapter 5, may also increase their motivation to seek out free goods and services, even at some risk.

Much prior research has characterized various variations of giveaway scams [114, 257], which are often distributed on social media platforms such as YouTube [28, 54, 111, 518]. These types of scams may entrap a significant number of victims. For example, Badawi et al. estimated that a subset of giveaway scams called the "Game Hack Scam," received 60 million impressions from 2014 to 2018 [28]. Additionally, cryptocurrency giveaway scams perpetuated on YouTube have been found to cause millions of dollars of loss [518].

Despite this evidence of harm, no human subject research has characterized how or why participants fall for these schemes. Moreover, there is a lack of research on social engineering involving media other than static text [76]. To help address these gaps, we conducted an online, scenario-based experiment to evaluate the reactions of both adults ($n = 205$) and teens ($n = 85$) to giveaway scams advertised via YouTube videos. Participants were asked to advise a hypothetical friend who was searching for either free in-game currency for Roblox, called Robux, or access to Spotify Premium. The experiment was designed to simulate the path a victim would take, with participants first viewing search results, then a YouTube video, and finally a video

This chapter is based on a paper originally published at the Network and Distributed System Security (NDSS) Symposium 2025 that was co-authored with Lily Klucinec, Mandy Lanyon, Julie Downs, and Lorrie Faith Cranor [58]

of the websites linked from the YouTube video. At each step, participants indicated what their friend should do and why they would recommend this action. Participants were assigned to view one scam and one video presenting a legitimate free offer (e.g., a free trial) randomly selected from three scams and three legitimate offers for each topic. Stimuli used in the experiment were derived from actual YouTube search results found during a qualitative investigation of giveaway scams.

Through our experiment, we sought to explore three main research questions.

- **RQ1:** How do users reason about YouTube giveaway scams, and what actions do they recommend when encountering them?
- **RQ2:** Do teens and adults interact differently with YouTube giveaway scams? If so, do these differences in behavior result in teens being more vulnerable?
- **RQ3:** Do other demographic and behavioral factors (e.g., income, experience with scam victimization, etc.) impact users' interactions with giveaway scams?

We found that most participants would not fall for these scams, with less than 10% indicating they believed the scam stimuli were legitimate. Users generally exited the less convincing scams immediately, whereas the more convincing ones prompted them to do more research into the offer (RQ1). Teens were not significantly better or worse at identifying giveaway scams, nor were they more likely to indicate that their friend should follow the scam's instructions. We did note some significant differences in behavior, with teens more likely to recommend that their friends report the scam videos that they saw. Teens were also significantly more likely to have previously searched for free Roblox Robux, the keyword we used to find scam videos on YouTube (RQ2). This offers direct behavioral evidence for the content-based targeting discussed in Chapter 3, showing how scammers' content can be effectively, if not necessarily deliberately, targeted at populations. None of the behavioral or demographic factors we investigated significantly influenced users' behavior regarding scams, although women expressed greater uncertainty and were therefore slightly less likely than men to accurately identify scams (RQ3). Ultimately, this chapter provides insight into an understudied type of social engineering and presents one of the first studies comparing adults' and teens' interactions with internet fraud.

6.1 Exploring YouTube Giveaway Scams

In this study, we examine teens' and adults' reactions to giveaway scams and test whether they have different abilities to distinguish scam websites from legitimate ones using an ecologically valid set of stimuli. To this end, we first investigate and describe the landscape of YouTube giveaway scams and then use these findings to identify stimuli for our human-subjects experiment. The basis for this investigation is two-fold: 1) conducting organic searches on YouTube and 2) identifying giveaway scams in a sample of 1% of YouTube collected for a previous unrelated research study in 2021 [6].

6.1.1 Organic Searching on YouTube

We conducted three rounds of organic searches on YouTube to identify giveaway scams that appeared to target teenagers. Organic searches were used to best simulate how teenagers interact with YouTube. Prior work has also used organic searching to find giveaway scams [28, 54]. Our multi-stage method of "progressive focusing" was inspired by Wei et al.'s [538] study of surveillance described in TikTok posts. The first round of searching was designed to determine if search results differed based on demographic features, and the second round of searching narrowed down the relevant search terms. Finally, the third round of searching used the narrowed list of terms to find videos to include in our survey.

To perform the searches in the first round, we created six YouTube accounts to cover all combinations of age (13 or 17) and gender options (male, female, or prefer not to say). These ages were selected to compare the results associated with each end of the age range. Each account was created and used within a separate emulated Android environment, using the Genymotion Android emulator. We used the following search terms and analyzed the first 10 results: "make money online for teenagers," "free robux," "free walmart gift cards," and "free ipad." These specific search terms were selected based on the expectation that they would reveal giveaway scams [54] and their relevance to teenagers [404]. We did not find enough qualitative differences in the search results between the accounts and decided to only use one account for the sake of time moving forward.

For the second round of searching, we created a new account with gender listed as prefer not to say and age as 15 years old. These demographics were selected to fall in the middle of our age range and be neutral in terms of gender. We collected the first 10 results from each of the following search terms: "easy online jobs," "free gift cards," "free netflix," "free roblox robux," "free spotify premium," and "make money online for teenagers." Changes to search terms were informed by the results of the first round of searching (e.g., finding no scams from the terms "make money online for teenagers" or "free ipad"). We analyzed the search results for both legitimate opportunities for financial gain, often through programs such as Microsoft Rewards, or for giveaway scams [28, 54, 114, 257]. Of the six terms, only "free spotify premium" and "free roblox robux" produced the variety of results desired, and the other terms were dropped.

For our final round of searching, we searched for videos matching our desired criteria to include as stimuli in our study. We defined three major categories of videos related to monetary gain without payment on YouTube a priori: legitimate, clickbait, and scams. We defined legitimate videos as real and largely safe opportunities for people to receive a monetary benefit, such as earning points through Microsoft Rewards or signing up for a free trial of a streaming platform. We defined clickbait videos as videos having titles unrelated to their actual content, mainly to trick people into viewing them without delivering what they claimed, or including non-harmful but ineffectual "methods" of receiving a benefit, such as editing the HTML of a website. Finally, we define scam videos as those that encourage viewers to pursue unsafe methods of obtaining rewards that will never be received, such as "human verification" to unlock a reward, which leads to an endless loop of tasks to complete, such as filling out surveys with personal information or downloading apps. This criterion was adapted from Bouma-Sims et al. [54], which broadly defined a scam video as one "that attempts to attract viewers through misrepresentation, including

fraudulently offering tangible, intangible, or financial awards." Two researchers independently classified each video, discussing to resolve disagreements.

We conducted searches over a two-week period on the same account used for the second round of searching until we found three results in each aforementioned category for both the "free spotify premium" and "free roblox robux" search terms. All of the legitimate Spotify videos involved free trials through Spotify directly or brand partnerships between Spotify and other trusted companies like PayPal. Similarly, all of the legitimate Roblox results involved earning points through Microsoft Rewards to trade in for Robux. Most of the scams across both search terms involved websites posing as coupon hosting services, with the exception being a website with many "Download" buttons that lead to an "Unlock Link" page once clicked.

6.1.2 Identifying Giveaway Scams on 1% of YouTube

To quantify the number of giveaways scams that are similar to those presented in our survey, we analyzed data from a previous, unrelated study, collected using "random prefix sampling" representing approximately one percent of all videos on YouTube between August 2020 and March 2021, for a total dataset of "about 10 million videos that have at least 800 views" at the time of sampling [6]. We analyzed the data using regex commands on JSON files in the dataset to pull out all videos with a variation of the words "robux" or "spotify" in the title, as these were the two search terms with the most relevant results in our organic searches. We found many results for the "spotify" search, but none were the giveaway scam format we desired, which is likely due to the high usage of "spotify" in videos posted by music artists. We found 451 relevant results for the "robux" search, and proceeded to analyze each result.

We next attempted to view each result on YouTube and assigned one of five codes based on each video's availability on YouTube as of April 2024: "available," "not available," "private," "unlisted," and "not English." Only the videos that were coded as "unlisted" or "available" were investigated further, leaving a total of 161 Robux-related videos to be analyzed out of the initial 451 (approximately 36% of the relevant results). We analyzed each of the remaining result videos, and those matching the giveaway scam format were flagged as scams. Of the 161 available videos, we determined 17 of them to be scams, which is approximately 11% of the available video results from the "robux" search. These results indicate that there would be approximately 1,199 Robux giveaway scams on YouTube between August 2020 and March 2021 based on this random sample of approximately 1.4% of videos on the platform [6]. We anticipate that the actual number is larger based on our organic searches, as many of the effective scams were likely taken down over the past four years.

6.2 Methods

We describe our online behavioral experiment designed to evaluate adult and teenage participants' ability to recognize YouTube giveaway scams and to assess how they reason about and interact with them. We first discuss our experimental design before discussing recruitment, data analysis, ethical considerations, and limitations of our work. All materials necessary to replicate

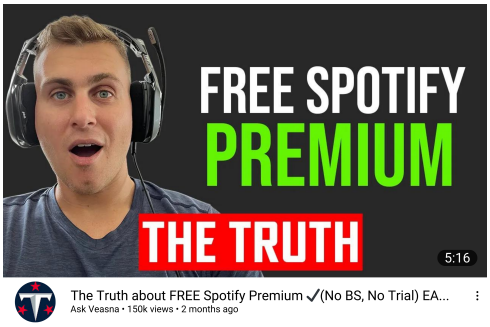


Figure 6.1: Example of a search result displayed to participants in the Spotify condition (Spotify Scam 2).

our experiment (e.g., survey instruments, recruitment materials, stimuli, etc.) are available in Appendix D or the associated artifact (see Appendix D.1).

6.2.1 Experimental Design

To evaluate our hypotheses, we employed a scenario-based design intended to mimic the process of being victimized by YouTube scams. In the experiment, participants were asked to imagine they were advising a friend searching on YouTube for ways to earn free Roblox Robux or gain free access to Spotify Premium. We decided to have participants advise a friend so stimuli could be presented randomly, without suggesting that the participant had done something they did not want to do. Previous studies of security and privacy behavior have also asked participants to imagine they are advising a friend [62, 252]. Participants were randomly assigned to one of these two topics and told the normal cost of Spotify Premium or Roblox Robux accordingly. Participants then completed several tasks.

Search result selection task

Participants first viewed nine YouTube search results in a random order (see Figure 6.1). Among these search results were three legitimate videos, three clickbait videos, and three scam videos (see Section 6.1). Participants were asked to select the videos they would advise their friend to click on or indicate if they would not select any of the videos. Participants were asked to justify their response by selecting which of the six features of the search result they considered (i.e., number of views, date of publication, the title of the video, thumbnail image, name of the channel, profile picture of the channel, or the length of the video). They were also asked to explain how they selected search results in an open-ended question.

Video review task

Participants were next shown two YouTube videos, one presenting a scam and the other presenting a legitimate free offer. These videos were selected at random from the three scam videos and three legitimate videos for each topic, regardless of which search results the participant indicated that they would select in the previous task. They were asked to watch each video and advise

their friend on what to do next. Participants could opt to visit the website, exit the video, report it to YouTube, search online to learn more, look at the comments, leave a comment, or indicate that they would do something else. They were also asked to justify their answer and identify which features of the video influenced their decision. Immediately after answering questions about each YouTube video, participants were told that their friend had gone to the website linked in the video. They were then shown another video of someone scrolling through the website. As with the YouTube video, they were asked to advise their friend on what to do next and justify their answer. The legitimate stimuli and scam stimuli were shown in a random order to control for potential order effects. Participants needed to watch at least half of each video before they could advance in the survey.

Scam rating task

After collecting survey responses from 30 teen participants and 10 adult participants, we added a question to the end of the survey directly asking whether participants thought each of the videos they saw was a scam or not. Participants were shown the search result for each of the two videos they saw and were asked to rate on a five-point Likert scale, ranging from "definitely legitimate" to "definitely fraudulent." 55 (64.7%) of the teen participants and 195 (95.1%) of the adult participants answered these questions.

Demographic and behavior questions

Before and after the experimental tasks, participants were asked questions to contextualize their responses and measure potential predictors of fraud victimization. At the beginning of the survey, participants were asked questions about their internet usage habits developed based on the Common Sense Media census [404] (e.g., what social media services they use, how much time they spend doing digital entertainment activities, etc.). After the experimental tasks, participants were asked about their prior usage of Spotify/Roblox and whether they had ever looked for free Roblox Robux/free Spotify Premium before. Participants were then asked how often they displayed behaviors that we hypothesized may be associated with scam victimization (e.g., prior victimization by similar online scams, cryptocurrency usage, etc.). To encourage truthful answers, we included distractor questions about benign behaviors (e.g., using coupons, receiving cash back from a credit card, etc.). In the adult version of the experiment, the survey concluded with a series of demographic questions. These questions were also asked of the parents of teenage participants during the consent process (see Section 6.2.2).

6.2.2 Recruitment

We recruited two samples of participants: a sample of minors aged 13 to 17 (n = 85) and an age-balanced sample of adult crowdworkers (n = 205).

Recruiting Teen Participants

Teen participants were recruited through a digital flier distributed to their parents/guardians. The flier informed parents about the study and included a link to an informed consent form. We used

non-specific language to describe the study tasks to avoid priming participants about the study (i.e., "[we are] investigating how teens interact with social media posts on YouTube"). The flier was distributed in various ways over time, as explained below. More details on the recruitment procedure can be found in a companion publication [56].

Once a parent/guardian gave consent for their child to participate in the study, they were redirected to a survey asking questions about their child's demographics (e.g., state of residence, age, household income, race, etc.). Parents were then given a link to an assent form for their child to complete. The child assent form explained the study procedure and the potential risks of participation at an 8th-grade reading level (based on Flesch reading ease). If teen participants affirmed that they were eligible for the study (i.e., they were fluent in English, located in the US, used YouTube, and were between 13 and 17 years old) and that they assented to participate in the study, they were then redirected to the main survey instrument. After completing the survey, they were redirected to a final survey where they could enter their email address to be compensated with a $5 Amazon gift card.

The recruitment flier was initially distributed via snowball sampling by sending the flier to parents known to the authors who have children in the appropriate age range for the study. These parents were also asked to share the flier with other families that have eligible children who may be willing to participate. We recruited 30 teenage participants using this approach from November 11, 2023, to December 7, 2023.

On December 13, 2023, we began distributing fliers via Peachjar,[1] a K-12 school communication platform that allows third parties to pay to distribute digital fliers to families with children enrolled in particular schools. Our goal in using Peachjar was to obtain a larger and more diverse sample, so we submitted fliers to school districts around the US. We initially selected districts based on the median income where the school district was located, as reported in the 2022 American Community Survey (ACS) [490]. We selected the largest high school in school districts with at least one high school evenly from three income groups: less than $42,606, $42,606 to $74,580, and greater than $74,580. $42,606 was the threshold for a three-person household to receive reduced-price lunches starting July 2022 [493]. $74,580 is the national median household income reported in the 2022 ACS [490]. We used this method to select 25 schools to distribute our flier to from December 13, 2023, to January 8, 2024.

On February 26, 2024, we submitted our flier to 50 more schools. We selected school districts by sampling those with the most high schools. We expected these to have the largest student populations and would result in more participants. We also translated our materials for parents (i.e., flier, consent form, and parental survey) into Spanish to send fliers to schools that required Spanish and English fliers. Teenage participants were still required to be fluent in English to take the survey. Five additional fliers were submitted to schools on April 30, bringing the total number of schools to 90.

We initially did not implement measures to prevent spam, as we expected only parents in our selected school districts to see the fliers. We received many responses that appeared to be spam (i.e., duplicate responses, meaningless answers to open-ended questions, failed Captcha, etc.) between January 2-4, 2024. Once we noted this issue, we paused data collection and retracted the 17 Peachjar fliers posted up to this point. Out of caution, we removed all survey responses

[1] https://peachjar.com/

(122) received during this period, including a small number that wrote plausible responses to open-ended questions.

We relaunched our survey on January 8, 2024, with new anti-spam checks. We generated posters with a unique URL for each school district and required parents to select which school district their child attended from a drop-down list to progress through the survey. Parents who selected a school district or state different from the one corresponding to their flier could not proceed with the study. After implementing this procedure, most attempted spammers were screened out. Moreover, using a custom URL for each flier allowed us to quickly block spam from a particular flier without shutting down the entire survey. We had to remove a flier on March 8, 2024, due to one or more spammers discovering the correct answers to the security questions. 125 responses from this flier on March 7-8, 2024, were removed.

Recruiting Adult Participants

Adult participants were recruited using the online crowdworking platform Prolific[2]. We recruited adult participants in 10 gender-balanced age buckets (18-22, 23-27, 28-32, 33-37, 38-42, 43-47, 48-52, 53-57, 58-62, and 63 or older). We recruited 19 to 24 participants from each group, comprising 205 participants. Pre-experiment power analysis was performed using GPower 3.1 [169] to determine how many adult and teen participants we needed. This calculation used a repeated measures ANOVA model, assuming an effect size of 0.25, power = 0.95, $\alpha = 0.05$, 2 measures, 12 groups (i.e., the total number of legitimate and scam stimuli), and a 0.5 correlation between measures. The result of this calculation was a total sample size of 312.

As with the flier for parents of teens, the Prolific recruitment advertisement described the study broadly so as not to prime participants to expect scam videos. Adult participants were recruited over an extended period. An initial 10 participants were recruited on November 20, 2023. The remaining adult participants were recruited incrementally, one age bucket at a time, as responses were received from teen participants. The final adult participants were recruited on April 10, 2023.

6.2.3 Quantitative Data Analysis

We conducted statistical analysis to determine whether participant responses varied significantly ($p < \alpha = 0.05$) during the search result selection task, video selection task, or scam rating task. Participants' responses during the search result selection task (Q9) and video review task (Q12 and Q15) were multiple select questions. Each choice was thus treated as a separate binary outcome variable (e.g., whether the user recommended exiting the YouTube video or not, whether the user recommended reporting the video or not, etc.). The responses to the scam rating task were interpreted as a binary variable corresponding to whether the participant correctly identified a video as a scam or legitimate.

The independent variables we investigated are listed in Table 6.1. Fisher's exact test [181] was used to determine whether the distribution of dependent variables varied significantly with respect to categorical and binary independent variables. For ordinal independent variables, we

[2]`https://www.prolific.com/`

used the Cochran-Armitage test, which determines whether there is a significant linear relationship between an ordinal variable and a binary response variable [23]. To quantify the size of statistically significant differences, we computed Cramer's V [124] for categorical or binary dependent variables and Freeman's θ [188] for ordinal dependent variables. Both values range from 0 to 1, with 0 indicating no relationship between two variables and 1 indicating a perfect relationship. All planned comparisons involving the same dependent variable were corrected using the Benjamini and Hochberg procedure [42, 238]

For the search result task, we conducted additional statistical testing to determine whether the selection rate varied significantly between search results. For this purpose, we used Cochran's Q-test, which is suitable for determining whether a binary response variable differs between conditions for within-subjects data [116]. Serlin et al.'s maximum-corrected measure of effect size, η^2_Q, was used to quantify the size of differences for these tests [430]. η^2_Q ranges from 0 to 1, with 1 indicating a perfect relationship and 0 indicating no relationship.

After completing the above analysis, we conducted some additional unplanned comparisons to explore potential confounds. Most notably, we performed statistical testing to compare the responses to behavioral and demographic questions between men and women. We also tested whether certain relationships remained significant when controlling for other variables.

Due to space constraints, we are unable to report all results. We focus on reporting significant results related to scam stimuli. Anonymized data from our experiment, the code used to run statistical tests, and the complete testing results are available in the associated artifact (see Appendix D.1).

6.2.4 Qualitative Data Analysis

Open-ended questions were analyzed via inductive coding performed by two authors. The lead coder reviewed 100 randomly selected responses to develop an initial codebook for each set of questions. The lead coder and another author then independently coded samples of 10% to 20% of the responses, meeting regularly to discuss differences in coding and decide on a consensus code for each set of responses. When new codes were added to the codebook, the lead coder re-reviewed previously coded responses, adding the new codes where necessary.

The coders started by analyzing the responses to the two questions where participants were asked to explain their response to the YouTube video. The lead coder realized during codebook development that the responses to these questions were often interdependent, so these two questions were analyzed together. Similarly, the responses to the two questions in which participants were asked to explain their response to the video of the website linked by the YouTube video were also analyzed together. The codebook developed while analyzing the video explanations served as the basis for the website explanations, as similar reasoning was displayed across these questions. A unique codebook was developed to analyze user explanations for YouTube search result selections.

The codebooks for each set of questions with definitions and example quotes can be found in the Appendix D.3. Each codebook is divided into two categories: reasoning codes, which describe participants' decision-making process, and feature codes, which describe the specific elements of the stimuli mentioned by a participant. The majority of participant responses were assigned more than a single code.

Table 6.1: Potential independent variables.

Variable	**Type**	**Note**
Adult	Binary	True if age ≥ 18
Condition	Categorical	Video shown
Time on digital entertainment (daily)	Ordinal	Response to Q1
Time watching online videos (weekly)	Ordinal	Response to Q4
Time playing mobile games (weekly)	Ordinal	Response to Q5
Time playing computer games (weekly)	Ordinal	Response to Q6
Time using social media (weekly)	Ordinal	Response to Q7
Time using other websites (weekly)	Ordinal	Response to Q8
Previous use of Roblox or Spotify	Ordinal	Response to Q18
Experience purchasing Robux or Spotify Premium	Binary	Response to Q19
Experience receiving Robux or Spotify Premium for free	Binary	Response to Q20
Experience with shopping scams	Ordinal	Response to Q27
Experience doing online tasks for money	Ordinal	Response to Q29
Experience purchasing cryptocurrency or NFTs	Ordinal	Response to Q31
Experience doing online tasks for money, but not receiving compensation	Ordinal	Response to Q33
Household income	Ordinal	Response to Q34
Gender	Categorical	Response to Q35
Type of community	Categorical	Response to Q39

6.2.5 Ethical considerations

Our study protocol was approved by our Institutional Review Board (IRB). To prevent participants from visiting scam websites, links, and other methods of accessing websites (e.g., images of Google search results) were redacted from all videos, including legitimate videos. We also included a debrief at the end of the survey that identified which videos the user saw were scams and gave advice on avoiding victimization. Our primary ethical concern in recruiting teen participants was ensuring we received consent from their parents/guardians. If we recruited teenagers directly, it was likely that many would pretend to be their parents when asked to obtain parental consent. In consultation with our IRB, we opted to recruit teens through their parents, as this would ensure that consent was received before the teen even saw the study.

6.2.6 Limitations

Participants encountering YouTube videos promoting giveaway scams in their normal lives may behave differently than they did in our study. Participants also may have had different perceptions of scams and legitimate websites if they had been able to interact with them directly rather than viewing a video of the website. While redaction of URLs was necessary to protect participants, this reduced participants' ability to detect scams accurately. Much of our quantitative analysis is based on self-reported experiences and behavior, which may be influenced by social desirability bias [204] or other forms of self-report bias. Our qualitative analysis is necessarily subjective, and a different research team may have identified different themes in the same data. Due to the difficulty in recruiting teen participants, our experiment may be slightly underpowered (see Section 6.2.2 for more details on the power analysis).

While we tried to recruit a diverse sample, our sample is not representative, which may limit the generalizability of our results. In both the teen and adult samples, there is a notable lack of participants identifying as Hispanic or Latino (see Section 6.3.1). The adult sample recruited from Prolific may have a higher level of technical expertise than the general population. Tang et al. [471] found that samples from Prolific are comparable to a probabilistic sample of the US population when asking about security and privacy perceptions and experiences, but not security and privacy knowledge. Additionally, teens from low-income households seem to be slightly underrepresented in our sample (see Section 6.3.1). While we did not find a relationship between income and scam susceptibility, teens from low-income households may experience different financial pressures, which might lead them to behave differently. Ultimately, future work should seek to explicitly study the behavior of marginalized groups when encountering internet fraud.

6.3 Results

In this section, we review the results of our behavioral experiment. We first discuss participants' demographics and behavior, highlighting the differences between adult and teen participants. We then discuss the results of each of our research questions.

6.3.1 Participant Characteristics

Table 6.2 contains an overview of participants' demographic characteristics, as well as the distribution of the most important behavioral characteristics. Participants were from 41 US states and Washington, D.C. The most common states where participants resided were California (12.4%), Florida (9.3%), and Texas (6.9%). Teens came from households with slightly higher incomes than adult participants (43.3% of teens vs. 26.6% of adults who reported income live in households with income of $100,000 or more; $p = 0.048$, $\theta = 0.178$). Parents were less likely to report household income on behalf of their teens than Prolific crowd workers (21.2% of parents vs. 2.9% of the adults recruited from Prolific did not provide a household income). Prolific crowd workers may be more comfortable sharing income information due to their greater past experience with online surveys.

Unsurprisingly, the reported behavior of adult and teen participants varied in many ways. Most notably, adult participants reported much more previous experience doing online tasks for money (82.9% of adults vs 7.1% of teens reported doing online tasks five times or more; $p < 0.001$, $\theta = 0.824$). They also reported more experience doing online tasks without being paid (67.8% of adults vs 31.8% of teens reported doing tasks for money without being paid at least once; $p < 0.001$, $\theta = 0.386$). This almost certainly results from our adult sample being recruited from Prolific. Adults also reported more previous victimization by online shopping scams (58.0% of adults vs 37.6% of teens reported purchasing something without receiving the item or a refund at least once; $p < 0.003$, $\theta = 0.222$). Adults reported more experience purchasing crypto assets, such as cryptocurrencies or NFTs (34.8% of adults vs 17.6% of teens reported purchasing crypto assets at least once $p < 0.001; \theta = 0.203$).

Teen participants reported more previous usage of both Spotify and Roblox. Teens had much more experience with Roblox (77.3% of teens vs. 27.1% of adults in the Roblox condition had played Roblox at least once; $p < 0.001$, $\theta = 0.549$), but were only slightly more likely to have used Spotify at least once (100% of teens vs. 89.0% of adults in the Spotify condition had used Spotify at least once $p < 0.005$, $\theta = 0.090$). Adult participants were more likely to be frequent users of Spotify. Teen participants were also more likely to have purchased Roblox Robux (38.6% of teens vs. 13.5% of adults in the Roblox condition had purchased Robux; $p < 0.002$, Cramer's V $= 0.265$). Teens report more weekly usage of mobile games (91.8% of teens vs. 72.2% of adults reported playing mobile games for some time each week; $p = 0.034$ $\theta = 0.240$).

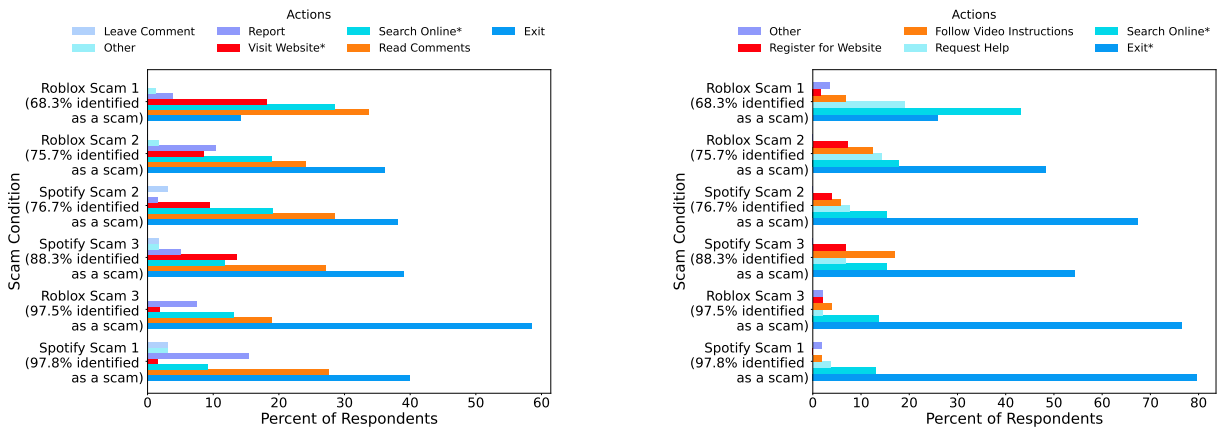
6.3.2 RQ1: Participant interactions with giveaway scams

Most participants (84.4%) correctly identified scam stimuli as definitely or probably fraudulent, with 9.2% indicating that the video was definitely or probably legitimate and 6.4% selecting the "I'm not sure" option. Participants' success varied significantly based on which video they saw ($p < 0.004$; Cramer's V $= 0.311$). 97.8% of participants successfully identified the most easily spotted scam (Spotify Scam 1), while only 68.3% successfully identified the hardest-to-recognize scam (Roblox Scam 1).

Figure 6.2 shows what actions participants recommended based on the scam video they saw and the corresponding video of someone visiting the scam website, alongside the percentage of

Table 6.2: Overview of selected demographic and behavioral characteristics. Questions about Roblox and Spotify were only asked of participants in the appropriate conditions.

	Adults		**Teens**			**Adults**		**Teens**	
	n	%	*n*	%		*n*	%	*n*	%
Gender					*Household Income*				
Female	98	47.8	37	43.5	Less than $20,000	20	9.8	3	3.5
Male	102	49.8	46	54.1	$20,000–$39,999	38	18.5	10	11.8
Non-binary	4	2.0	1	1.2	$40,000–$59,999	36	17.6	16	18.8
Prefer not to say	1	0.5	1	1.2	$60,000–$79,999	33	16.1	4	4.7
Age					$80,000–$99,999	19	9.3	5	5.9
13	—	—	9	10.6	$100,000–$149,999	32	15.6	14	16.4
14	—	—	16	18.8	More than $150,000	21	10.2	15	17.6
15	—	—	27	31.8	Prefer not to say	6	2.9	18	21.2
16	—	—	17	20.0					
17	—	—	16	18.8	*Roblox use* ($n = 140$)				
18–24	31	15.1	—	—	Never	70	72.9	10	22.7
25–34	42	20.5	—	—	1 or 2 times	15	15.6	13	29.5
35–44	45	22.0	—	—	3 to 5 times	4	4.2	5	11.4
45–54	33	16.0	—	—	More than 5 times	7	7.3	16	36.4
55+	54	26.3	—	—					
Race					*Spotify use* ($n = 150$)				
Am. Indian/Alaska Native	2	1.0	2	2.3	Never	12	11.0	0	0.0
Asian	20	9.8	7	8.2	1 or 2 times	22	20.2	14	34.1
Black/African American	38	18.5	14	16.5	3 to 5 times	6	5.5	8	19.5
Hispanic/Latinx	13	6.3	5	5.8	More than 5 times	69	63.3	19	46.3
Native Hawaiian/Pac. Isl.	0	0.0	1	1.2					
White	148	72.2	62	72.9	*Weekly video viewing*				
Self-describe	3	1.4	0	0.0	None	0	0.0	1	1.2
Prefer not to say	1	0.5	5	5.9	Less than 1 hour	10	4.9	0	0.0
Community Type					1 to 5 hours	70	34.1	21	24.7
Rural	38	18.5	6	7.1	5 to 10 hours	53	25.9	24	28.2
Suburban	105	51.2	44	51.8	10 to 15 hours	25	12.2	19	22.4
Urban	62	30.2	30	35.2	15 to 20 hours	19	9.3	9	10.6
No response	0	0.0	5	5.8	More than 20 hours	28	13.7	11	12.9
Total	**205**	**100**	**85**	**100**					



(a) Scam Video

(b) Video of Someone Visiting Scam Website

Figure 6.2: Participants' recommended actions in response to the scam stimuli. Participants could select multiple actions. An asterisk in the legend indicates that the rate of recommending this action differed between conditions. The scam conditions are ordered along the y-axis from most to least convincing based on the scam rating task ($n = 260$). Those who did not identify a video as a scam were either unsure or thought the video was legitimate. We use red and orange to indicate the most dangerous actions.

participants who recognized the stimuli as a scam. For the easiest-to-recognize scams, participants were less likely to recommend that their friend visit the website ($p < 0.005$, Cramer's V $= 0.289$) or search online to learn more about what was shown in the video ($p < 0.005$, Cramer's V $= 0.301$). Participants viewing the videos of the hardest-to-recognize scam websites were less likely to recommend that their friend exit the website ($p < 0.003$, Cramer's V $= 0.360$), but were not significantly more or less likely to indicate that their friend should register for the scam website or follow the instructions from the scammer. Rather, users were more likely to want to search online after viewing the website ($p < 0.003$, Cramer's V $= 0.326$). This suggests that the most obvious scams are easily dismissed by users, while more convincing scams lead users to want to do research to learn more. Across all conditions, only 5.5% of participants recommended that their friend register for the website or follow the instructions from the scam without doing additional research or asking for help.

Many users who suggested visiting the website after viewing the scam video also recommended other actions. For example, 37.1% of participants who recommended that their friend visit the website also suggested that their friend should search online to learn more. In their open-ended responses, many of these users expressed a desire to learn more before proceeding. For example, one adult participant who suggested that their friend should visit the website wrote, "I would tell them to spend some time checking the safety of the website before doing any downloads to make sure it is a safe site, in case this is a website or company that is not widely known." Searching online about these scams may prevent victimization, as victims report scams

to crowd-sourced review websites like Trustpilot.[3] For example, one of the scam websites we showed participants has 29 negative reviews on TrustPilot as of July 10, 2024.[4]

More so than searching online, users viewed looking at the comments as a good source of information on other users' experiences. 60% of users who recommended that their friend visit the website linked in the video also recommended that they should look at the comments. One adult participant explained, "I would check comments to see if it worked for anyone else and if so then visit the website on the video." Similarly, another adult participant wrote, "Comments will determine action to take." While potentially intuitive, relying on comments for information is a bad idea, as the video poster can control what comments are visible or even post fake testimonials to convince others of the scam. Many scam videos featured testimonials in the comments, claiming that the website worked. A few participants recognized the potential unreliability of YouTube comments. For example, one adult participant wrote, "...I would recommend they checked out a reliable online source about the website (NOT the comments which can be fake)."

For participants who recommended exiting the video or website, the necessity of doing tasks in order to get a reward was the most common feature brought up by participants, with users typically finding these suspicious or not worth the effort. 41.9% of users who recommended that their friend exit the video after viewing the scam video, and 31.4% of those who recommended that their friend exit the website after viewing the video of someone visiting the website pointed to the presence of tasks. For example, one teenage participant wrote, after viewing the second Spotify scam, "...having to complete untrustworthy offers seems suspicious." Upon viewing the web video, they added, "Completing online offers most likely never gives you what you want."

Users who mentioned the tasks in their responses did not necessarily realize that the videos they saw were scams. 16.9% of the participants who recommended exiting the video and 8.4% of participants who recommended exiting the website were coded as thinking that following the provided instructions would not be worth the effort. For example, one adult participant who saw the second Roblox scam wrote, "...it is too much hassle for what little bit you get. I'm sure they're getting plenty from you for signing up. Not worth it."

Users' reasoning about tasks went beyond just the fear that they would not be compensated or feeling that the compensation was not worth the effort. Some participants (13.2% who recommended exiting the video and 11.5% of those who recommended exiting the website) objected to the need to provide personal information to complete tasks. For example, one adult participant who saw the first Spotify scam video wrote, "If you have to enter or complete other things that you have to input you[r] information I don[']t trust [it]." Some participants pointed to specific security or privacy risks associated with performing the tasks. For example, one teen participant who saw the first Roblox scam video wrote, "Because it says 'Download apps' it's obviously a scam. Everyone with slight common sense should know that downloading random apps and opening links can give you a virus." 6.6% of those who recommended exiting the video and 11% of those who recommended exiting the website were coded as mentioning a security or privacy risk.

Participants often based their decisions on less direct indicators of deception. For example, the second most common feature that participants pointed to was the presenter of the scam video.

[3] https://www.trustpilot.com
[4] https://www.trustpilot.com/review/boujeecoupons.com

22.8% of those who recommended that their friend exit the video mentioned the presenter. The first Spotify scam video, which was one of the easiest to recognize as a scam, is presented by a man in a mask with a concealed voice. Many users found this suspicious; 46.2% of participants who saw this video mentioned the presenter in their justification for their recommended action. One adult participant wrote, "I bet the reason he is dressed like that is... because what he is doing is illegal...." Videos with computer-generated voices rather than real human voices were also viewed as suspicious. One adult participant who viewed the third Spotify scam wrote, "This looks like a scam." They pointed specifically to the fact that it featured an "AI voice over instead of a human" as unconvincing.

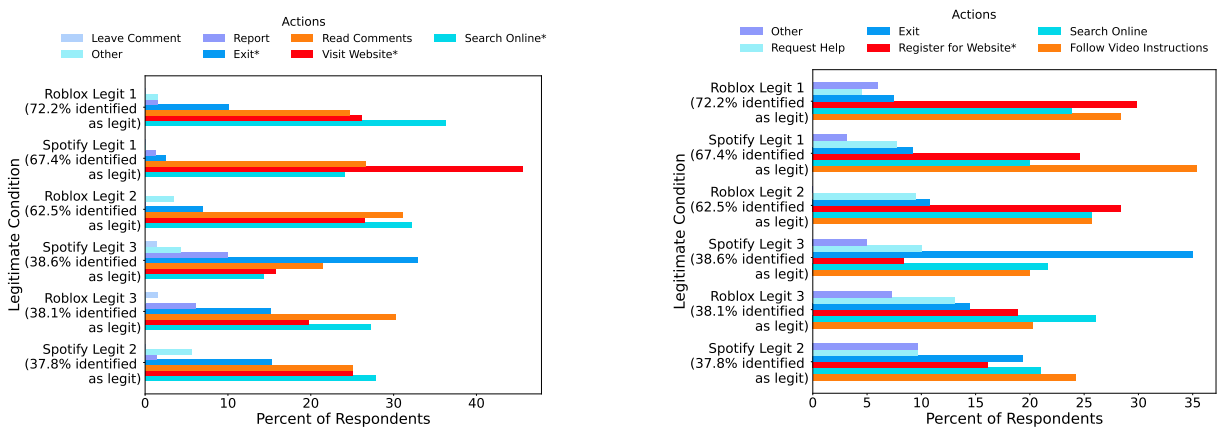
The aesthetic quality of the website was another important influence on users' decision-making. 30.4% of those who said that their friend should exit the website pointed to some aesthetic feature. Roblox Scam 3, which was one of the easiest to recognize as a scam, was hosted on a website that was filled with ads that were made to look like download links. 60.4% of participants who saw the video of this website mentioned its poor quality. For example, one adult participant explained their decision to exit the website by stating, "The website is very messy and littered with cheap, scammy ads." A focus on aesthetic elements can be useful in identifying the most obvious scams, but may not protect users from more polished schemes.

People who fell for the scam (i.e., those who stated that their friend should follow the instructions/register for the scam website) often did not provide specific reasoning beyond the fact that they believed the video. For example, one teen participant who recommenced that their friend follow the instructions in the Roblox Scam 3 wrote, "...if they want their Robux they should do that." Some participants emphasized that simple instructions were provided. For example, one adult participant who recommended that their friend follow the instructions in Spotify Scam 3 explained their answer by stating, "The steps are easy to follow." Similarly, a participant who recommended that their friend follow the instructions from the Roblox Scam 2 explained their answer by pointing to the fact that it gave "simple instructions, [and] offer[ed] help."

Users' reactions to the legitimate videos may suggest that they were more paranoid than they would be during normal browsing. Only 52% said the legitimate video they saw was probably or definitely legitimate, 12% were unsure, and 36% said it was probably or definitely fraudulent. As with the scam stimuli, participants' ratings varied significantly based on the video they saw ($p = 0.013$, Cramer's V $= 0.295$). 72.2% participants who saw the most convincing legitimate video (the first Roblox video) identified it as legitimate, while only 37.8% of those who saw the least convincing legitimate video (the second Spotify video) were convinced. Lower accuracy in identifying legitimate stimuli as compared to fraudulent stimuli has been noted in other studies, such as Lastdrager et al.'s [275] study of phishing training for children.

Figure 6.3 shows what actions participants recommended based on the legitimate stimuli they saw. There was less difference between the conditions based on the legitimate web stimuli, with only the rate of participants suggesting that their friend should register for the website varying significantly based on the condition ($p < 0.005$, Cramer's V = 0.277). This result is unsurprising, as all participants who saw the Roblox stimuli saw the same web video (i.e., the Microsoft rewards website). The first and second Spotify web stimuli showed the same Spotify free trial website.

Participants' reasoning about the legitimate stimuli was less related to security fears. For example, the most common feature pointed to by people who recommended that their friend exit



(a) Legitimate Video

(b) Video of Someone Visiting Legit. Website

Figure 6.3: Participants' recommended actions in response to the legitimate stimuli. An asterisk in the legend indicates that the rate of recommending this action differed between conditions. The legitimate conditions are ordered along the y-axis from most to least convincing based on the scam rating task ($n = 260$). We use the same colors for actions as we did in Figure 6.2, but here all actions are safe.

the video was that the video presented a trial offer and would not be free once the trial period ended. 18.6% of participants who recommended exiting the video mentioned the trial. For example, an adult participant who recommended that his friend exit the third legitimate Spotify video wrote, "...the video is somewhat misleading and is just showing him how to sign up for a free trial, rather than showing ways to get Spotify premium for free."

For those who believed the legitimate stimuli were presenting a scam, the inability to see URLs was often a factor. Users (reasonably) feared they could be directed to a phishing scheme. For example, a teen participant who saw the second legitimate Spotify video wrote, "...I would avoid going to a direct link provided in a video like this as there are a number of ways it could be malware or a phishing scam." Presumably, if the participants could see that the link led to the actual Spotify website, they would not have thought that they were being led to a scam.

Users generally showed a bias towards legitimate stimuli during the search result selection task. For the Roblox stimuli ($p < 0.001$, $\eta^2_Q = 0.104$), the most frequently selected results were the first and second legitimate videos, which both were selected by 37.4% of participants. The least selected thumbnail was the first scam thumbnail, which was only selected by 7.2%. 64.0% of participants selected at least one legitimate search result, while only 38.1% of participants selected at least one scam search result. For the Spotify stimuli ($p < 0.001$, $\eta^2_Q = 0.0814$), the most selected thumbnail was the first legitimate search result, which was picked by 36.7% of participants. The least selected thumbnail was the third clickbait thumbnail, which was picked by only 6.7% of participants. 57.3% of participants selected at least one legitimate search result, while only 41.3% of participants selected at least one scam search result.

It is unclear why participants gravitated toward selecting legitimate search results. When asked which features of the thumbnail they considered, participants most often indicated the video's title (64.5%), the thumbnail image (61.7%), and the number of views (30%). The open-

ended questions provided little additional information on users' selections, with participants often just reiterating which parts of the image they looked at. The most common form of reasoning was vague, with users expressing that they found some thumbnails more reasonable than others. For example, one participant who selected the first legitimate Spotify search result and the second Spotify scam search result wrote, "I wanted information that didn't look shady or underhanded." For the Roblox search results, some participants seemed to recognize the creators of the legitimate videos. For example, one participant who only selected the first legitimate thumbnail wrote, "The guy in the video is popular in gaming." The legitimate stimuli generally had a greater number of views as compared to the scam stimuli, which may have guided participants' selections.

6.3.3 RQ2: Differences in teenager and adult behavior

We noted several significant differences between adult and teenage participants. However, none suggested that teens were more susceptible to the scams.

Teen participants were not significantly better or worse at identifying scams or legitimate stimuli. They also did not behave differently during the search results selection task. During the video review task, they were significantly more likely to indicate that their friend should report scam videos than adults (21.2% of teen participants vs. 4.4% of adult participants; $p < 0.001$, Cramer's V $= 0.250$). Teen participants who recommended reporting the video seemed to have a desire to protect others. For example, one teen participant who indicated that their friend should report the second Roblox scam video explained that they wanted "to lessen the percentage of people to see and fall for it." Similarly, a teen participant who recommended that their friend report the first Spotify scam video wrote, "They should do this to help others from making bad decisions...."

Very few participants addressed why they did not report a scam video, so it is difficult to draw conclusions about their motivation for not reporting. One possible explanation for the difference in reporting between teens and adults is that adult participants may be more jaded about the result of reporting videos. An adult participant who saw the second Spotify scam wrote, "...I wouldn't bother reporting it because I feel actions like that are fighting an uphill battle that have little or no effect."

Teen participants were also significantly more likely to recommend that their friend ask for help in response to the legitimate website ($p = 0.025$, Cramer's V $= 0.192$). In reasoning why they think their friend should ask for help, teens often pointed to the fact that personal information or a credit card was required. For example, one teen participant who saw the second legitimate Roblox video justified asking for help by saying, "make sure you dont give any information without an adult." Similarly, a teen participant who saw the third legitimate Spotify video wrote, "They may not have a card and [the website] might need to be fact checked."

Concerningly, teen participants were significantly more likely to report previous experience searching for one of the phrases we used to find scams, "free Roblox robux," or something similar ($p < 0.001$, Cramer's V $= 0.385$). 38.6% of teens vs. 6.3% of adults reported previously searching for the term. This is at least partially, if not wholly, explained by teens having more exposure to Roblox than adults (see Section 6.3.1). Searching for free Robux was more strongly associated with the number of times a person played Roblox (8.3% of participants who played

Roblox two times or fewer vs 43.8% of participants who played Roblox three or more times reported searching for free Robux at least once; $p < 0.001$, $\theta = 0.545$) and whether or not participants had purchased Robux before (43.3% of those who had purchased Robux vs. 9.1% who had not purchased Robux had previously searched for free Robux; $p < 0.001$, Cramer's V $= 0.356$). There was no significant relationship between any tested independent variable and past searching for "free Spotify premium" or something similar. 20.7% of participants who saw Spotify stimuli reported previously searching for free Spotify Premium.

As a result of searching for these terms, some participants shared that they saw similar results to those in the survey. For example, one adult participant shared that searching for free Spotify Premium "...showed me a bunch of websites that contained verification steps which never worked in giving me access." Similarly, a teen participant wrote that they found "Another... scammy website that make[s] you do surveys or download apps" after searching for free Roblox Robux. Some participants, especially in the Spotify condition, found genuine offers. For example, one adult participant shared, "I used the 3 month free trial from Paypal." Overall, there were 33 participants who reported negative experiences, 10 neutral or unclear experiences, and 8 positive experiences using similar search terms.

In addition to describing past experiences when explicitly asked, some participants' explanations for recommended actions reflected some level of prior experience with these scams. For example, one adult who recommended exiting out of the third Roblox scam video said the video was "Definitely one of those generators where you do tasks, like watching ads or filling out a form, to get your Robux that never works." Participants' prior experience also influenced their response to the legitimate videos. One teen user who recommended that their friend visit Microsoft Rewards after viewing the third legitimate Roblox video explained, "...i have actually done this. they give free points towards free robux." Participants' experience could also be misleading. Some participants believed the scam websites were real crowdworking websites. One adult participant who recommended that their friend follow all of the instructions on the second Roblox scam website justified the answer by writing, "Obviously I also use survey sites for extra income." Overall, 14.1% of participants referenced their prior knowledge in some way in their responses.

6.3.4 RQ3: Other factors impacting scam victimization

Women were less likely to correctly identify scam videos (92.8% of men vs. 75.8% of women correctly identified scam videos; $p < 0.004$, Cramer's V $= 0.223$), but were not significantly better or worse at correctly identifying legitimate videos. They also did not behave significantly differently in any of the tasks. This could partially result from women having lower confidence in their answers, as almost half (44.8%) of the women who failed to identify the scam video selected "I'm not sure." Women were also more likely than men to select "I'm not sure" when rating the legitimate stimuli; however, this difference was not significant. Prior studies have observed that men will show higher confidence in responses than women [422].

Participants who reported more daily time on digital entertainment activities were more likely to correctly identify scams (90.0% of participants reporting more than 4 hours vs. 77.3% of people reporting less than 4 hours correctly identified scams; $p < 0.004$, $\theta = 0.322$), but were not significantly better or worse at identifying legitimate stimuli. This result may be confounded

by the fact that women reported significantly less daily time on digital entertainment activities (56.3% of women vs 38.5% of men reported four hours of digital entertainment per day or less; $p = 0.016$, $\theta = 0.200$). This effect remains weakly significant when looking at women alone ($p_{uncorrected} < 0.006$, $\theta = 0.181$), but not men alone.

We found some statistically significant relationships between the tested independent variables and the probability of selecting particular search results, but there were no factors that were consistently associated with users picking search results. The number of times the participant had previously played Roblox was most frequently associated with search result selection. Users who had played Roblox more frequently were significantly more likely to indicate that they would not recommend that their friend pick any of the search results (17% of those who played Roblox five times or less vs. 45.5% of those who played Roblox more than five times selected no thumbnails; $p < 0.05$, $\theta = 0.294$). These participants may have a greater understanding of the risk of this search term, due to their greater experience searching for "free Roblox Robux" or something similar.

We noted several relationships that are likely spurious. Participants who reported more experience doing online tasks for money were less likely to indicate that they would report the scam video ($p = 0.015$, $\theta = 0.338$) or the legitimate video ($p = 0.022$, $\theta = 0.448$). They were also less likely to report searching for "free Roblox Robux" or something similar ($p < 0.001$, $\theta = 0.520$). As discussed in Section 6.3.1, teens were much less likely to have prior experience with online tasks. When looking at teenagers or adults alone, these relationships are no longer significant in either sample. Participants who reported more weekly time playing video games on a computer or game console were more likely to correctly identify scam stimuli ($p = 0.022$, $\theta = 0.275$). Men reported significantly more weekly time spent playing video games on a computer or console than women ($p < 0.001$, $\theta = 0.419$), and the rate of successfully identifying scams does not vary with time spent playing console or computer video games when looking at men or women alone.

6.4 Discussion

In this section, we summarize the key results from our experiment and discuss how they compare with other studies on different types of internet fraud. We then provide recommendations for how to reduce giveaway scams on YouTube and other similar platforms.

6.4.1 Overview of key takeaways

We found that most participants were able to recognize that the YouTube giveaway scams were fraudulent, with only 9.2% of participants indicating that they believed the scam they saw was legitimate (RQ1). While most participants came to this conclusion based on the necessity to do tasks, others used potentially unreliable heuristics, such as checking YouTube comments or focusing on aesthetic elements of the scam website. Prior studies on phishing have also observed that users may place too much weight on the perceived aesthetic quality of fraud messages and the presence of elements like company logos [51, 368, 369]. This type of reasoning is inherently

subjective and may lead users to trust polished scams and distrust poorly crafted but genuine communications.

We also found no significant difference in scam victimization between adults and teens. Teens were more interested in both of the topics we investigated and more likely to have previously searched for "free Roblox Robux," (RQ2). These results suggest that, while minors are not necessarily worse at scam identification than adults, their greater interest in the topic may lead them to encounter and be victimized by giveaway scams more often. This is consistent with the notion of content-based targeting, described in Chapter 3. We noted no significant differences in behavior due to other demographic or experiential factors, although women were slightly less likely to correctly identify scams (RQ3). As discussed in Section 6.3.4, this seems more attributable to lower expressed confidence than a lower ability to identify scams among women.

As the requirement to complete tasks was the most salient aspect of the scam stimuli, these results are likely most applicable to other giveaway scams that involve tasks.

6.4.2 Recommendations to reduce victimization

Where possible, it is important to prevent users from encountering fraud in the first place. AI methods are often employed to detect specific types of fraud, such as phishing messages [89], in order to block them before they reach a user. YouTube already employs machine-learning-based screening methods to identify and block content that violates its community guidelines [560]. It may be possible to tune these tools to better identify giveaway scam content; however, fraudsters will continually adapt their scams to avoid moderation. One simple approach to prevent users from encountering giveaway scams on YouTube and other video-sharing platforms would be to intentionally block search terms associated with victimization. Some platforms have already employed this approach. Users on TikTok, for example, are blocked from searching for "free Roblox Robux" or "free Robux." Inputting either term on the platform returns an empty results page that states, "This phrase may be associated with behavior or content that violates our guidelines." This approach is heavy-handed, potentially making some legitimate content harder for users to find. Moreover, it is impossible to anticipate all the terms scammers may use to defraud users. Still, this may be effective for the most commonly abused search terms.

Education may be effective at reducing fraud victimization, especially when embedded within fraud messages [437]. While videos that are clearly identified as scams should be removed, YouTube and other video-sharing platforms could add warnings about scams to videos that appear to offer ways to earn free goods. YouTube already uses a similar approach to counter misinformation. YouTube automatically adds links to Wikipedia articles to contextualize videos that speak about topics that are prone to misinformation (e.g., vaccines, the moon landing, etc.) [233]. Future work should evaluate the effectiveness of these interventions in reducing giveaway scam victimization.

The difference in propensity to report between adults and teens may suggest an opportunity to promote reporting through platform design. Increasing reporting of giveaway scams could allow YouTube to find and eliminate those that are not automatically screened more quickly. For example, YouTube could add "gamified" [144] elements to reward successful reporting, such as giving users a profile badge for submitting a certain number of valid reports. Any feature intended to promote reporting must be tested to ensure it does not increase false reports.

Beyond YouTube, many giveaway scams rely on cost-per-action (CPA) advertising networks that pay scammers for each completed task. While CPA can be an ethical way to monetize content, many networks turn a blind eye to fraud or even actively aid scammers. For example, in 2023, WIRED reported that the network CPABuild provided templates for scams [78]. Similarly, the blog of OGAds[5], another CPA network frequently used by scammers, features advice on how to avoid YouTube take-downs [354]. Eliminating these services is challenging, as they are generally run anonymously and can easily change names. CPABuild reincorporated as AdBlueMedia[6] following the scrutiny they received from WIRED's reporting [78]. Still, pressure should be placed on Cloudflare, Amazon Web Services, and other critical infrastructure providers to deplatform these services in order to eliminate scammers' ability to monetize giveaway scams.

[5] https://ogads.com
[6] https://adbluemedia.com/

Chapter 7

The Nature and Quality of Reddit Discussion about Scams

While the previous three chapters have focused more narrowly on specific populations, this chapter looks generally at online help-seeking for scams. As discussed in Chapter 3 and observed directly in Chapter 5, online support for digital safety issues is widespread [85, 94, 105, 310, 351, 531, 539], and may fill a gap for people without traditional avenues for support, such as those experiencing the risk factor of *resource constraints* or *social isolation and loneliness*. Moreover, observing Reddit provides greater insight into a phenomenon that is often underreported.

Building on the methods employed in Chapter 5, we conducted a multi-stage thematic analysis of a sample of 1,525 posts and 1,883 associated comments from four communities centered on scam discussion on Reddit, with the majority of posts coming from `/r/Scams`. We sought to answer the following research questions:

1. What types of scams are discussed on Reddit?
2. What types of scam-related support do users seek on Reddit?
3. What security advice or other help do victims of online scams receive on Reddit, and what is the quality of this advice?

We found that Reddit discussions focus on internet-mediated scams that constitute crimes. Discussion on Reddit focused on identifying scams, discussing the strategies scammers employ, and seeking advice on dealing with victimization. In addition to seeking general remediation advice, users commonly sought reassurance that they were safe from harm and other forms of emotional support. We noted variation across communities, with posters particularly using `/r/Sextortion` to give and receive emotional support.

We conclude that Reddit serves as a valuable resource that helps users avoid scams and cope with victimization. Still, the prevalence of scam-related discussions highlights the difficulty that users have in avoiding scams online. We encourage the development of automated tools to help users reason about scams, to discuss potential future research based on Reddit data, and to suggest increased law enforcement outreach to Reddit's scam-related communities.

This chapter is based on a paper originally published at the 2025 ACM SIGSAC Conference on Computer and Communications Security and was co-authored with Mandy Lanyon and Lorrie Faith Cranor [59]

7.1 Methods

We conducted a multi-stage thematic analysis of 1,525 posts from a set of Reddit communities dedicated to discussing scams.

7.1.1 Dataset

We did not collect data directly using Reddit's API. Rather, we used an archive of data collected via PushShift [35] from June 2005 through December 2022 and ArcticShift[1] from January to December 2023 (inclusive). This dataset includes the content of some posts deleted by users or removed by moderators, so it is more complete than the dataset available directly from Reddit. ArtcicShift is a successor to the widely used PushShift project, which is no longer generally available due to changes in Reddit policy. Other recent research on Reddit has also used data from ArcticShift (e.g., see Shen et al. [436]).

For our content analysis, we selected four communities dedicated to discussing scams: `/r/Scams`, `/r/Sextortion`, `/r/scammers`, and `/r/phishing`. `/r/Scams` has nearly a million members as of February 2025 and is by far the largest community focused on scams on Reddit. `/r/scammers` is a much smaller community that also focuses on the general discussion of scams. `/r/Sextortion` and `/r/phishing` focus on specific types of scams, as indicated by their names. These communities were selected based on the authors' prior knowledge of Reddit and by entering the terms "scam," "fraud," and "phishing" in Reddit's community search.[2] We chose not to include large communities that seemed to focus on online vigilantism, referred to as "scambaiting," in which users pose as victims and interact with scammers to waste their time or cause other harm [156, 484] (e.g., `/r/ScamBait` and `/r/ScamNumbers`) due to the artificial nature of scambaiting interactions and the potential for wholly fabricated exchanges raised by Dynel and Ross [156].

We collected 278,231 posts and 2,302,800 associated comments from these communities made between June 14, 2008 (the date `/r/phishing` was created), and December 31, 2023. The content of 84,091 of these posts (30.2%) was deleted by the original poster or removed by moderators prior to being archived by PushShift or ArcticShift. Of the remaining 194,140 posts, we randomly sampled 1,525 (0.8%) with 15,240 associated comments. This sample size was chosen to achieve a 95% confidence level with a 2.5% margin of error. The final sample was composed of $1,311$ posts from `/r/Scams` (86.0%), 144 posts from `/r/Sextortion` (9.4%), 42 posts from `/r/scammers` (2.8%), and 28 posts from `/r/phishing` (1.8%).

7.1.2 Thematic Analysis

To investigate our research questions, two authors performed thematic analysis in multiple stages. They developed codebooks that can be found in the Appendix E.1

[1] `https://github.com/ArthurHeitmann/arctic_shift`
[2] `https://www.reddit.com/search/?type=communities`

Stage 1: Initial Categorization

The first stage of coding was intended to characterize the type of content posted to the four selected scam-discussion communities to facilitate further analysis. To develop an initial codebook, the lead coder reviewed a subset of 100 posts from the sample and performed thematic analysis. The coder viewed the post title and post body. For non-spam posts, if a post contained a link to content that was not clearly irrelevant, the coder also considered this content. At this stage, the coder did not systematically analyze the comments, although we often read comments in part to aid in categorizing the post (e.g., when the poster provided additional information in the comments).

We primarily categorized posts based on their apparent purpose (e.g., *scam identification*, *request for support*, *report of victimization*, etc.). Many posts linked to images or videos to show the scam that was being discussed. If these were no longer available, we coded the post as "media missing" and did not review it during subsequent analysis (255 posts, 16.7% of the sample). For posts that discussed a scam, we characterized the scam based on the initial method used to communicate with the potential victim (e.g., social media, instant messaging, etc.) and other technologies used in the scam (e.g., crypto assets, gift cards, etc.). We also assigned a general "type" code that was meant to capture the distinctive characteristics of the scams (e.g., *relationship scam* for scams that involve a scammer pretending to be a friend or romantic partner, *ecommerce seller* for scams perpetrated by a purported online seller of goods/services, etc.). The specific codes we chose were informed by the language used by Reddit users and the terminology used in prior research. Finally, for posts where users reported victimization by scams, we assigned a code for any harm experienced due to the scam (e.g., financial loss, exposure of personal information, etc.).

After developing the initial codebook, the lead coder and another author collaboratively coded the full sample of posts. We dual-coded 400 posts, with the coders independently coding blocks of 100, comparing the results, and meeting to resolve any disagreements. During this phase, we added new codes as appropriate. On the fourth set of 100, the coders achieved Krippendorff's α [266, 315] of 0.951 for post type, 0.895 for initial communication method, 0.905 for technologies used, and 0.966 for type of harm experienced by a victim. Considering the time-consuming nature of dual coding the entire sample, we divided the remaining posts equally between the coders and coded them separately. The coders met regularly throughout the separate coding process to discuss difficulties and emerging themes.

Stage 2: Analysis of Advice

The second coding stage was intended to analyze the experience of scam victims requesting support on Reddit and the type of advice they received. As such, we focused on the 262 posts (17.2%) that were coded as *requests for support* in the first stage.[3] For these posts, we identified the type of support that the victim requested, what remediation steps they had already taken (if any), what private information they had disclosed to the scammer (if any), and the magnitude of financial harm they had experienced (if any). At this stage, the comments on posts were also

[3]Coding was initially performed with both posts coded as *report of victimization* and *request for support*. After codebook development, we decided to code only support requests.

reviewed systematically to identify the support given to users. We read up to 15 comments per post, reading each thread from the earliest to the latest. We chose this ordering, as we expect that victims are more likely to see and rely upon comments posted sooner rather than later. The default comment ordering shows the most-upvoted comments first, so this ordering does not reflect how other users viewing the post would experience the comments section. Where appropriate, we revisited the initial categorization of post type and scam type at this stage.

Coding followed a procedure similar to the first stage. The lead coder reviewed 50 posts and associated comments to develop an initial codebook. The two coders then dual-coded two sets of 50 posts and associated comments, meeting after each round of coding to compare results and resolve disagreements. On the second round of dual coding, the coders achieved Krippendorff's α of 0.822 for the type of support that the victim requested, 0.829 for the remediation steps the victim had already taken, 0.823 for the type of private information disclosed, and 0.933 for the comment codes. Considering the substantial agreement, the remaining posts and associated comments were divided evenly and coded separately.

Stage 3: Codebook refinement

After all posts and comments were assigned initial codes, the coders met to review and refine the initial codebook. In particular, the coders sought to group codes into larger categories and generate new codes that better captured the diversity of the data. Most notably, we grouped scam types into four nonexclusive categories, partially inspired by the categorization used in Deliema et al.'s investigation of the factors contributing to scam victimization [138]: impostor schemes involving impersonating a trusted entity (e.g., a business, the government, family/friends, etc.), fraudulent goods and services schemes premised on the sale of goods/services (e.g., non-delivery, non-payment, fake or defective products, etc.); opportunity schemes where the victim is attracted by the promise of something desirable (e.g., a prize, employment opportunity, etc.), and threat-based schemes where the victim is induced to pay money based on a threat of harm (e.g., arrest, release of intimate imagery, etc.). We also reviewed comments initially categorized as "Other" to determine what commonalities existed, and what new codes should be created. The final codebook includes 9 post-type codes and 39 scam-type codes.

After discussing changes to the codebook, the lead coder reviewed content marked with any of the codes that were split or merged (e.g., comments/posts that were marked with one of the "Other" codes). Any errors that the lead coder noticed were also corrected.

7.1.3 Ethical Considerations

Our research is not required to undergo IRB review, as we relied on publicly available data. Nevertheless, it is important to consider the ethical issues associated with social media research. Most users do not expect their posts to be analyzed as part of academic research [177], and we did not obtain consent from community members or moderators prior to conducting our research. In line with the recommendations of Fiesler et al., we considered the community's privacy norms before conducting our research [178]. As detailed in the results, many users on `/r/Scams` explicitly acknowledged the public nature of their posts and expressed a desire to help others by sharing their experiences. Members on `/r/Sextortion` discussed and shared research that

was based on posts to their community [85]. We believe this context supports the ethical use of these posts in research. The publication was shared with /r/Scams, /r/Sextortion, and /r/phishing with permission of the moderators, so that the studied communities may benefit from our findings [485, 486, 487].[4]

We chose to include deleted posts and comments. We felt that excluding removed content may introduce bias if posts about certain types of scams are disproportionately deleted. Since this content is available in public archives (e.g., ArcticShift) and we are not resharing it, we believe our analysis does not cause additional harm. We only discuss deleted content in the aggregate and do not individually reference posts or comments that were deleted as of July 9, 2025. Some comments on deleted posts are discussed, where the highlighted content remains available on Reddit.

Victims of crimes are a vulnerable population that may be targeted by criminals for revictimization (see Section 7.2.4 for a discussion of the scams targeting victims on Reddit). Moreover, scam victimization is highly stigmatized [127, 131] and users may not want to be publicly known as survivors of a scam. We take several steps to prevent the identification of the users we observed. First, we avoid mentioning the usernames of posters or commenters. Second, we will not publicly post our curated dataset, although the authors will provide an archived copy of the data upon reasonable request. Finally, most examples of posts or comments are described rather than quoted directly. Summarizing quotes involved editing original excerpts to replace the original text with synonyms or alternative phrases that convey the same idea. Some nouns were replaced with more general descriptors (e.g., a word like "cousin" or "brother" may become "family-member"). As recommended by Reagle [393], the lead author verified that Google or Reddit searches including the provided text did not return the highlighted post or comment.

There are two exceptions where we quote verbatim. We directly quote comments from the AutoModerator as these messages are not authored by individual users, and the excerpts cannot be associated with specific posts. The title includes a quote that several users have used in scam identification posts. As this quote cannot be tied to any single user, we leave it verbatim.

7.1.4 Limitations

We cannot verify the claims made by users, and some posts and comments may contain intentional or unintentional falsehoods (e.g., when a user misunderstands what has happened to them). Further, we reviewed only a subset of communities centered on the discussion of scams, and users also seek advice about scam victimization on other communities (e.g., /r/Advice, /r/personalfinance, etc.). The nature of discussions and quality of advice may differ in these communities. Our sample skews towards /r/Scams. This subreddit accounts for the majority of Reddit posts about scams, but sampling randomly limits comparative analysis across subreddits. We also exclusively focused on comments and posts in English, so non-English speaking users are not represented in our sample. Finally, qualitative coding is inherently subjective, and a different team may have identified different themes in the same data.

[4] /r/scammers moderators did not respond to a request to post about the study.

Table 7.1: An overview of the types of posts we observed. Percentages refer to the proportion of posts in each column assigned to each type code. Proportions for subreddits other than `/r/Scams` should be interpreted with caution due to the small sample size.

Post Type	`/r/Scams` (%)	`/r/Sextortion` (%)	`/r/scammers` (%)	`/r/phishing` (%)	Sample (%)
Scam identification	430 (32.8%)	2 (1.4%)	2 (4.8%)	5 (17.9%)	439 (28.8%)
Disc. of scam comms.	306 (23.3%)	2 (1.4%)	14 (33.3%)	4 (14.3%)	326 (21.4%)
Request for support	163 (12.4%)	93 (64.6%)	3 (7.1%)	3 (10.7%)	262 (17.2%)
Report of victimization	76 (5.8%)	26 (18.1%)	5 (11.9%)	1 (3.6%)	108 (7.1%)
Scambaiting	34 (2.6%)	3 (2.1%)	6 (14.3%)	1 (3.6%)	44 (2.9%)
Scam-related media	35 (2.7%)	1 (0.7%)	2 (4.8%)	4 (14.3%)	42 (2.8%)
Irrelevant	28 (2.1%)	4 (2.8%)	1 (2.4%)	2 (7.1%)	35 (2.3%)
Other	8 (0.6%)	3 (2.1%)	0 (0 %)	3 (10.7%)	14 (0.9%)
Media missing	231 (17.6%)	10 (6.9%)	9 (21.4%)	5 (17.9%)	255 (16.7%)

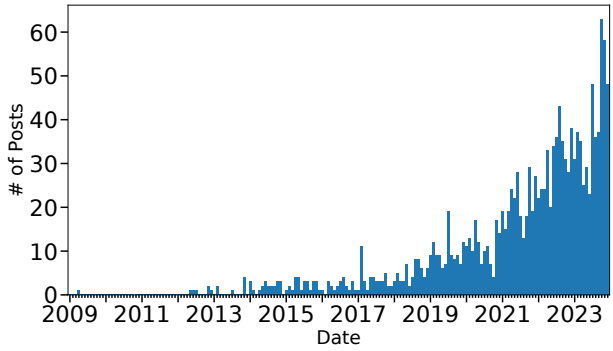


Figure 7.1: Frequency of posts per year within the sample of scam-related communities.

7.2 Results

In this section, we describe the results of our content analysis and evaluation of advice quality. We begin with an overview of the sample before discussing results with respect to each research question.

7.2.1 Overview

Figure 7.1 shows the frequency of posts in the sample over time. In line with the growth of Reddit, the majority of posts (55.0%) in the sample come from 2022 and 2023. The median number of comments per post was 5, with 170 posts (11%) receiving no comments. Table 7.1 shows the breakdown of the different categories of posts, broken down by subreddit. We found that Reddit is used primarily to help users identify scams, discuss scammers' tactics, and seek support to deal with scams.

Scam identification

Across all subreddits, a plurality of the posts we reviewed (439 posts or 23.8%) were seeking to identify whether something was a scam (*scam identification*). These posts featured an image of communications from a potential scammer or a description of an interaction that the poster was concerned about. In a typical example, a user made a post to `/r/Scams` asking whether a message they received was a scam. The post linked to an image of an SMS that threatened the recipient with arrest. The user received reassuring replies, with the first commenter stating that the government never sends messages threatening arrest and that the poster could safely ignore it. This result is similar to Hasegawa et al.'s study of non-expert security and privacy questions on a Japanese Q&A website (Yahoo! Chiebukuro), which found that the most common type of question sought to determine whether a user was facing a cyberattack [215].

In most cases, users posted scam identification questions before victimization. Only 23 posts in this category (5.2%) described some financial harm, and 19 posts in this category (4.3%) described the disclosure of personal information or intimate imagery to a scammer.

Discussion of scam communications

The next most common category was *discussion of scam communications*, with 326 posts (21.4%) assigned this code. This category is diverse, including any post in which users describe a scam they encountered or heard about but were not victimized by. Some of these posts were intended to warn others about a scam the user encountered. For example, one user made a post to `/r/Scams` where they described a phone call they received that attempted to collect on a fake debt. In explaining why they were posting, the user stated that they could not find a reference to the scam on the subreddit, so they posted to warn others. Posts designed to warn others about security threats were also noted in Bhagavatula et al.'s study of discussion of security and privacy on Twitter and Facebook, including posts warning about compromised accounts and a post about a Trojan being advertised on Facebook [46].

In some cases, users wanted to learn more about how a scam operated. In a typical example, one user made a post to `/r/Scams` where they asked about the purpose of a seemingly fraudulent job ad offering hundreds of dollars per day for less than eight hours of work.

Other posts in this category were intended to make fun of scammers' poor attempts at deception. For example, a user made a post to `/r/Scams` that linked to a fake receipt that was likely an attempt to initiate a customer service scam. The title jokingly admonished the scammer for being lazy.

Request for support

We coded 262 posts (17.2%) as a *request for support*, meaning that they asked for some type of support other than scam identification. In most cases, these posts related to specific victimization experiences, although a small number sought support with scam prevention. These posts were the majority of content we reviewed from `/r/Sextortion`, reflecting the community's primary purpose as a support group for people facing financial sextortion. We discuss the types of support that users requested in greater detail in Section 7.2.3.

Report of victimization

We coded 108 posts (7.1%) as *report of victimization*, meaning that they described users' experience with scam victimization without asking for support or asking to determine whether something was a scam. Reports most often described financial harm (56 occurrences, 51.9%), the sharing of intimate imagery with a scammer (23 occurrences, 21.3%), or the disclosure of personal information (14 occurrences, 13.0%).

Reports varied widely in detail, with some posts being extremely short. Several posts consisted of only titles that described an incident without going into further detail in the body or comments. Other posts were very detailed. For example, one user made a post to `/r/Scams` to describe their experience with a phishing scam. They explained that they received a message with a short code purporting to be from their bank, reporting an unexpected debit card transaction. After interacting with the text message, they received a phone call from a spoofed phone number matching their bank and were socially engineered to send a transfer to the scammer. They concluded by discussing how they immediately realized that it was a scam, but were told by their bank that they were unlikely to get their money back. They expressed shame and the feeling of being stupid. These and similar posts provide valuable insight into victims' experiences by explaining the mechanics of the scam, identifying the features that caused the scam to seem plausible (i.e., authentic-looking message content, spoofed phone number, etc.), and describing the emotional and financial impact on the victim.

As with posts that were categorized as discussions of scam communications, users often expressed a desire that their story would help others avoid the same scam. For example, one user wrote a post identifying a specific Reddit user who scammed them on Craigslist. They concluded by expressing the hope that their post would stop others from being harmed. This type of prosocial motivation has been observed in previous studies examining victims' motivation for reporting fraud to the government [128].

Especially on `/r/Sextortion`, posts in this category often seemed to serve the purpose of 'venting' about victims' fears or negative emotions. For example, in a post made to `/r/Sextortion`, a user who was victimized four months prior discussed their difficulty in recovering emotionally. They shared that the experience had destroyed their self-esteem and given them a constant fear of making decisions. They described that being extorted drove them to begin smoking and drinking.

Other post categories

The remaining categories did not provide insight into users' experiences with scams and were not analyzed beyond initial categorization. Posts related to *scambaiting* were relatively rare in the communities we investigated (44 posts or 2.9%). This result is unsurprising, as scambaiting content is not allowed under the rules of the largest subreddit we investigated, `/r/Scams`. Additionally, we coded 42 posts (2.8%) as *scam-related media*, consisting predominantly of links to news articles, blog posts, or YouTube videos warning users about particular scams. The *other* category (14 posts or 0.9%) was reserved for posts that were on-topic but did not fit into any of the other categories (e.g., general discussion of spam that was not related to a specific type

of scam). Finally, we coded a small number of posts as *irrelevant* (35 posts or 2.3%), including generic spam and other off-topic posts.

7.2.2 RQ1: Types of scams discussed

In this subsection, we describe the types of scams discussed on Reddit. The most frequently reported scams—including sextortion, fraudulent ecommerce sellers, and corporate impostor schemes—align with both government data and Reddit's user demographics. Scenarios that did not meet our definition of a scam were rarely mentioned. Across the sample, most scams were mediated by the internet or related technologies.

Type coding results

Ecommerce seller scams, sextortion, and corporate impostor scams were the most common types of scams in our dataset. Table 7.2 shows the frequency of the 10 most common individual type codes. The high prevalence of *ecommerce seller* scams is unsurprising, as online shopping scams are the second most common type of fraud reported to the FTC, and the most common type of fraud reported to the FTC by those under the age of 39 [498] (i.e., the primary age group that uses Reddit [198]). Bitlab et al. previously leveraged Reddit discussion from `/r/Scams` to investigate fraudulent ecommerce scams [50]. The inclusion of `/r/Sextortion` in our dataset is mainly responsible for the high proportion of *sextortion*. We categorized only 26 posts (1.9%) that were posted to other subreddits as discussing sextortion.

Corporate impostor schemes also appeared frequently. 56.2% of the occurrences of the corporate impostor code came from posts categorized as discussion of scam communications, as compared to only 3.5% of posts about sextortion and 19.3% of posts about ecommerce seller scams. This may reflect that the users of these subreddits found these scams easier to identify and avoid, or that victims of these scams were less likely to seek help on Reddit. Impostor schemes are the most common type of fraud reported to the FTC, with older age cohorts more likely to report these schemes to the FTC than younger adults [498].

Many corporate impostor messages were likely attempts at *phishing*, although we only classified posts as such where the scam message clearly indicated that personal information was required (E.g., by explicitly asking users to log in). For example, in a post to `/r/Scams`, a user shared that they received a message saying that a delivery service had tried to drop off a parcel but was unsuccessful. They described that the message instructed them to click a link to initiate a new delivery. They went on to say that they found the scam particularly insidious because they were waiting for a package that morning. They acknowledged that it was likely a coincidence but that this experience helped them appreciate how people could fall for this type of scam. We did not categorize this as phishing because the author did not specify that sensitive information was requested.

Our scam-type classification is necessarily uncertain for posts where users have not been victimized by the scam they discuss. We observed many communications likely tied to scam attempts, but we lacked certainty. For example, 17 posts (1.5%) discussed so-called *wrong number* scams. These are spam texts seemingly addressed to someone else that are often sent with the hopes that the user will respond, and a scammer will be able to form a relationship with the

Table 7.2: 10 most common scam type codes, excluding "unclear" and "likely not a scam." Percentages are calculated with respect to the 1,135 posts categorized as scam identification, discussion of scam communications, report of victimization, and request for support. Codes are not mutually exclusive; 352 or 31.0% of posts were assigned multiple type codes.

Scam Type	Type Category	#	%
Ecommerce seller	Commerce	145	12.8%
Sextortion	Threat	141	12.4%
Corporate impostor	Impostor	130	11.5%
Prize scam	Opportunity	102	9.0%
Employment scam	Opportunity	92	8.1%
Fake payment	Other	66	5.8%
Ecommerce buyer	Commerce	62	5.5%
Other threat	Threat	55	4.8%
Phishing	Impostor	53	4.7%
Investment scam	Opportunity	44	3.9%

victim. While often used as an entree to investment scams [170], these texts may be legitimate. Moreover, even assuming they are fraudulent, it is impossible to tell what type of scam will follow based on the initial text alone, hence why we created a separate category for this ruse.

Many posts were also difficult or impossible to classify. 218 (19.2%) posts were classified as having an *unclear* scam type, meaning we could not unambiguously classify them within our coding scheme. The majority of these (49.5%) were in posts categorized as scam identification or discussion of scam communications (30.7%). For example, one post to `/r/Scams` from 2019 asked about a robocall in Chinese they were receiving on a weekly basis from various numbers. It seems probable that this robocall is tied to some sort of scam, such as the Chinese consulate impostor scam observed by Prasad et al. around the same time period [382]. However, it is impossible to categorize the incident with just the information provided.

Discussion of non-scams

Some posts clearly discussed incidents that did not meet our definition of a scam (see Section 2.1). 122 posts (10.7%) were categorized as *likely not a scam*, with most of these posts being categorized as scam identification (62.3%). Neither the posts coded as likely not a scam nor the associated comments were analyzed beyond post type and scam type categorization (e.g., they were not assigned codes for harm, communication methods, technologies used, etc.).

This set of posts was diverse, but we noted some common archetypes. Several posts involved cases in which users were concerned that legitimate interactions were scams. For example, in one post categorized as scam identification made to `/r/Scams`, a user who was seeking to purchase marriage records asked whether a local government website was legitimate, stating that they had no idea whether it was real and that they do not trust things.

Other posts in this category discussed business practices that users found distasteful but did not seem to constitute criminal behavior. For example, one user made a post complaining about

an online shop that bought items at a discount from Lowe's but sold them at the MSRP on Amazon. These posts reflect the colloquial usage of the term scam but are not of interest in understanding cybercrime.

Finally, some users discussed account compromise, identity theft, or other issues that involved no human interaction and were not clearly connected to scam victimization, like receiving mail addressed to other people. While it is possible that these experiences stemmed from scam victimization (e.g., phishing that results in account compromise), these posts lack detail on any social engineering that may have occurred.

Technologies used by scammers

The majority of scams discussed in the sample were mediated by the internet or related technologies. Only 31 posts (2.7%) were initiated via an in-person interaction or postal services. While a large proportion of posts (194 posts or 17.1%) did not clearly specify how the scam was initiated, the most common communications technologies used to initiate scams were social media/social media messaging (221 occurrences, 19.5%), instant messaging (148 occurrences, 13.0%), and email (122 occurrences, 10.7%).

Unsurprisingly, the type of communication technology used to initiate a scam differed based on scam type. The majority of ecommerce seller scams were initiated via ecommerce websites or services (62.1% of posts specifying an initial communication method). In contrast, and in accordance with prior work [85], sextortion was most often initiated by social media/social media messaging (53.1% of posts specifying an initial communication method). The majority of corporate impostor scams were initiated via instant messaging (35.8% of posts specifying an initial communication method) or email (30.0% of posts specifying an initial communication method).

The most common payment methods mentioned as being used in scams were digital payment services (83 posts or 7.3%) and crypto assets (61 or 5.4%). Gift cards were used only rarely in the sample (20 posts or 1.8%). The relative frequency of different payment methods is in line with the relative frequency of payment methods in FTC fraud reports [498]. Digital payment services like PayPal or Zelle are particularly popular with scammers, as they often lack protections for users who willingly initiate a transaction, even when they are deceived. For example, in its documentation on Fraud and Scams, Zelle states, "Even if you were tricked or persuaded into authorizing a payment for a good or service someone said they were going to provide... you may not be able to get your money back" [564]. Crypto assets are popular with scammers for similar reasons, as transactions are typically irreversible and pseudonymous; however, their use is more challenging for users, as evidenced by the fact that victims were often instructed on using crypto assets as part of scams.

7.2.3 RQ2: Type of support requested

In this subsection, we examine the nature of scam-related support-seeking on Reddit. Users sought help for a wide range of scams, with sextortion emerging as the most common in our sample. While many posts described financial losses, few explicitly asked for help recovering assets. Instead, users often requested more open-ended support. Notably, many users sought reassurance that the scammer could no longer harm them or sought other emotional support.

These findings highlight the multifaceted nature of scam-related harm and the broad spectrum of support needs expressed by victims.

Types of scams in requests for support

Due to the nature of our sample, the plurality of posts categorized as requests for support discussed sextortion (103 posts or 39.3%). The remaining posts in this category were diverse, with no other single type of scam being asked about more than 25 times; scams involving ecommerce sellers (22 posts, 8.4%) and fake payments (19 posts, 7.3%) were the two next most common categories respectively. Moreover, 12 posts (4.6%) requested support unrelated to a specific scam and 20 posts (7.6%) requested support for things that were likely not scams.

Users typically requested support for scams perpetrated against themselves, but 30 posts (11.4%) requested support to help another person, like a family member or friend. Several of these posts related to older family members, like parents or grandparents, who were entrapped by a scam that the poster recognized but the victim did not. For example, one user made a post to `/r/Scams` describing how their father was being victimized by a scammer who had built up a relationship with him over Facebook Messenger before asking for increasing amounts of money. The poster had learned about the scam after the perpetrator requested tens of thousands of dollars for a fictional overseas investment. While the user's father hadn't sent the money, he was still in contact with the scammer. The poster requested help with preventing further harm, since confronting his father did not cause him to cut off contact. Difficulty in convincing victims they have been scammed has been noted by previous studies, especially with regard to romance scams [371, 544].

Users requesting support experienced a diverse set of harms. 84 of the posts categorized as requests for support (32.1%) mentioned some form of financial loss, 83 posts mentioned sharing of intimate imagery with a scammer (31.7%), and 42 posts (16.0%) mentioned the disclosure of personal information.

Financial loss is difficult to quantify, as users were often non-specific about how much money they lost, with 35 posts (41.7% of requests for support that mentioned a financial loss) not specifying the amount of money lost. Estimating based on conversion rates as of February 10, 2025, the median specified loss was $300, with values ranging from $4.99 to $175,000. This result is in line with the general distribution of fraud losses in FTC reports [498].

The most common types of information disclosed were generally not extremely sensitive. The most common items of information disclosed were a phone number (17 occurrences), name (12 occurrences), location (11 occurrences), and email address (9 occurrences). Only 6 posts mentioned disclosure of a Tax ID (e.g., social security number), 4 posts mentioned disclosure of bank account number, and 3 posts mentioned disclosure of one or more passwords. An additional 3 posts described account compromise without specifying what personal information had been disclosed. Our method likely undercounts the frequency of information leaks, as we only counted instances where users specifically discussed giving a scammer information. We did not infer that certain information was disclosed where this was not specified (e.g., many ecommerce seller scams likely involved disclosure of credit or debit card numbers).

Most victims of sextortion did not have their intimate imagery exposed. Only 15 or 14.6% of the posts classified as a request for support that discussed sextortion mentioned that their intimate

Table 7.3: The frequency and percentage of types of advice requested by Reddit posters dealing with scams. Codes are not mutually exclusive.

Support Sought	#	%
General scam remediation advice	118	45.0%
Reassurance about scam outlook	78	29.8%
Scam prevention advice	27	10.3%
Other users' experiences	27	10.3%
Understanding of victimization experience	22	8.0%
How to report scams	13	5.0%
Advice on financial recovery	13	5.0%
General emotional support	10	3.8%
Advice on collecting evidence	6	2.3%
Likely not a scam	20	7.6%

image was posted publicly or sent to someone that they knew. This is likely an overcount, as scammers may show victims fake screenshots of intimate images being shared to convince them to pay. For example, one `/r/Sextortion` commenter reassured a poster by stating that they were shown screenshots of the scammer sending a close family member their intimate images, but the family member never received anything. Still, even assuming that all of these cases of reported exposure were genuine, this result is in line with prior findings that most extortionists do not follow through on threats to release intimate images [85].

Scam remediation requests

Table 7.3 shows the frequency of different types of support that were requested. Most of these requests can be classified as seeking advice on *scam remediation*, referring to the need for scam victims to return to their pre-victimization state. The most common request type code was *general scam remediation advice* (118 posts, 45.0%). This category was primarily composed of non-specific requests for advice. For example, in a post to `/r/Scams`, one user described how their friend never received payment after selling a concert ticket on social media. Their friend did research after the experience, finding a number of social media comments warning about the user. They asked if anything else could be done, expressing the hope that their post would serve as a warning if nothing could be done.

We also included in this category specific questions that did not match the other themes we developed, but seemed to relate to recovering from scam victimization. For example, one user posting to `/r/Sextortion` asked how long they should wait before changing their social media account back to public visibility after being sextorted. As will be discussed in Section 7.2.4, changing privacy settings is a common strategy for mitigating the harm from sextortion.

Despite the high proportion of financial loss among requests for support, specifically asking about recovering lost funds (*advice on financial recovery*) was relatively rare, with only 13 posts. For example, one user asked about recovering funds from a Facebook Marketplace seller on

/r/Scams. They were manipulated by lies about the scammer's ownership of an expensive car into paying with Apple Pay, and they never received the item. They asked whether there was anything they could do to get their money back and called themselves stupid for completing the transaction. As discussed in Section 7.2.2, recovery from peer-to-peer payment services not intended for commerce is, unfortunately, often impossible.

Some users also requested support with other aspects of remediation like *how to report scams* (13 occurrences) or *advice on collecting evidence* (6 occurrences). Typically, requests about reporting focused on service providers in order to have the scammer de-platformed. For example, one user posted to /r/Sextortion asking about whether it was possible to have the perpetrator's Instagram account banned. Posts asking about collecting evidence most often related to trying to find the scammer. For example, a user posted to /r/Scams sharing their experience with a scam involving fake concert tickets. They concluded their post by stating that they wanted to identify the perpetrator. Their motive is not stated, but presumably, they hope to recover the lost funds or bring the scammer to justice.

Many users also posted to Reddit to remediate the emotional or psychological harm resulting from scams. We categorized 78 posts that requested support as seeking *reassurance about scam outlook*, meaning that they wanted to verify that they were safe from further harm. This was especially common in the context of sextortion or other threat-based scams. In a typical example, a user made a post requesting reassurance after encountering a scammer on a dating app. The scammer pretended to be an underage girl and then called the poster pretending to be her parents. The poster acknowledged that the scenario seemed fake, but they still wanted confirmation that they were not in danger. Users who disclosed personal information also sought reassurance about how the information they provided may be used by a scammer. For example, in a post to /r/Scams, a user shared that their significant other gave a government impostor scammer part of his tax ID, their address, and the name of his bank after he was threatened with arrest. They asked what the scammer could do with the information they now have. Hasegawa et al. observed similar questions on Yahoo! Chiebukuro, especially from users who began entering information into phishing forms but stopped before final submission [215].

Exclusively on /r/Sextortion, users also requested more *general emotional support* that was connected to scam victimization, but was not specifically related to the outcome of a scam (10 occurrences). For example, one user recounted their story of sextortion, concluding by stating that they felt stupid for paying when it did not even make them feel more secure. They said they wanted to find other victims to talk with who could make them feel better. These requests illustrate both the deep psychological harm that scams can do and the value that users find in seeking support on Reddit.

These findings mirror Wei et al.'s investigation of image-based sexual abuse, as they found that many victims of financial sextortion use Reddit for therapeutic support [539]. We extend this insight to a broader range of scams, demonstrating that victims commonly seek reassurance about their safety and the potential consequences of their victimization. The fact that many users turn to Reddit for reassurance and emotional support suggests a gap in accessible, formal support systems for scam victims.

Other types of support sought

Some common types of support requested did not clearly fall into the category of remediation. For example, some users requested that others share their experience with scams. This was often implicitly or explicitly tied to a request for scam remediation advice or emotional support. For example, in a post to `/r/Scams`, a member of the military recounted their experience with a scammer whom they paid thousands of dollars after the scammer threatened to get them kicked out of the military. They requested help from anyone who had experienced something similar. We coded this question as both a request for general scam remediation advice and a request for *other users' experiences*.

In other cases, users' requests for similar experiences seemed more rooted in curiosity than a desire to draw on others' experiences for support. For example, a user shared a detailed account of their experience with a survey impersonating Verizon and purporting to give a prize. They concluded by saying that they were posting to prevent others from falling for the scam and to see if anyone else had been victimized by a similar scam. In total, there were 27 (10.2%) questions tagged as requests for other users' experiences.

Users also sought advice on protecting themselves from scams. 27 posts were categorized as seeking *scam prevention advice*. Most often, these requests were related to spam texts and calls that a user received. For example, a user made a post to `/r/Scams` requesting advice on how to stop spam calls that they were receiving on a daily basis. Other users requested advice more generally, not tied to specific communications. For example, one user made a post to `/r/Scams` asking how they could protect their loved ones from scams. Similarly, a user posted to `/r/Scams` requesting advice on how to avoid being defrauded when selling an automobile online.

Along the same lines of informational support, some scam victims requested clarification on how or why a scam occurred. We categorized these requests as seeking an *understanding of victimization experience* (22 occurrences). For example, one user made a post to `/r/Scams` asking why they were getting job scam emails to their main email account that claimed to have found them on a job board. Their account with the job board had a different email attached, so they could not understand how their main email account was (apparently) being found through the job board. They further asked for scam prevention advice in the form of questions about how to stop these scam emails.

Sometimes, the request for understanding of victimization was directly tied to the need to remediate harm. For example, one user sought help with account compromise stemming from a ransomware attack on `/r/Scams`. Despite taking a series of account remediation steps (i.e., changing their email password, ending all logged-in sessions, and disabling POP/IMAP), the attacker still had access to their email account. They asked how the scammer could access their email and how to block the scammer from accessing their email in the future. We categorized this set of questions as both seeking understanding of victimization experience and general scam remediation advice.

7.2.4 RQ3: Support type and quality

In this section, we describe the type of support received on Reddit. A plurality of the comments we reviewed were intended to explain features of the scam a user encountered or provide reassurance to a victim. The most common types of other advice related to remediating harm from scams, although users also gave more general security advice. Most advice was predominantly non-technical, emphasizing interpersonal, procedural, or institutional actions such as blocking the scammer, contacting a bank, or reporting the incident. In contrast, technical advice—such as changing account settings or deleting social media accounts—was relatively rare and typically surfaced only in specific contexts.

Beyond the content of specific advice, we identified several broader themes that characterized how scam support was delivered. These included the important role that bots played in providing explanations and directing victims to helpful resources, distinct patterns of interaction and emotional support between `/r/Scams` and `/r/Sextortion`, and the ways that scam reports benefited not only the original poster but also other internet users who discovered the threads later. Together, these findings suggest that Reddit serves as a community-driven infrastructure for scam identification, emotional support, and harm reduction.

Overview of support process

Receiving support was a dynamic process, with users asking clarification questions and posters providing additional details about their experience as appropriate. 475 of the $1,883$ comments we reviewed (25.2%) came from the original poster of the thread. The majority of these comments provided additional details about a user's experience (276 occurrences of *additional detail(s)*, 58.1% of comments from an original poster). The content of comments posted by other users varied widely, with a large proportion unrelated to our research questions (102 occurrences of the *other/irrelevant* code).

The most common type of support given to victims was to identify or otherwise explain the type of scam a user encountered. We assigned 416 comments the code *explain scam*. We frequently assigned this code to cases where users simply identified that something was a scam, but we also applied this code to more specific explanations of a user's understanding of the likely mechanics of the scam. For example, in response to a user of `/r/Scams` who inquired about how an account could have been compromised after they clicked on a phishing link but did not provide any information, one user speculated that the scammer could have installed a keylogger or stolen the Steam login cookie from their browser. It is unclear whether this explanation is correct, but it reflects the type of informational support often given.

The second most frequent type of support we observed was *reassurance* to victims (254 occurrences). We most frequently applied this code to comments that offered encouragement related to the likelihood of harm. For example, in response to a post requesting support for victimization by sextortion, one user offered reassurance by stating that it was actually against the fraudster's best interest for them to post their intimate imagery, as it places them at greater risk of being caught. They further shared an anecdote from a former scammer who claimed that scammers deleted data from hard drives to avoid being caught with evidence. We categorized this comment as both *explain scam* and *reassurance*, as the commenter infers how sextortion

scams operate and uses these inferences to provide emotional support to the victim. We also applied the reassurance code, where users offered more general emotional support, such as one commenter's statement that a victim of sextortion would be able to overcome their experience. The prevalence of reassurance is unsurprising in light of the many posts requesting reassurance or other emotional support. Wei et al. similarly noted a prevalence of therapeutic support among posts discussing IBSA, including financial sextortion [539].

The remaining types of support can be broadly classified as *scam remediation advice* and *general security/scam avoidance advice* based on the apparent purpose of the advice. The following sections discuss the most frequent types of advice in each context. A number of other specific pieces of advice appeared more than once but cannot be discussed due to space constraints (see the Appendix E.1). In addition to these codes, many comments included advice that was difficult to classify, especially where the advice was specific to a user's situation. We coded 105 comments as recommending *other remediation step(s)* and 85 comments as recommending *other security/scam avoidance advice*.

Scam remediation advice

As discussed in Section 7.2.3, users most often sought advice on remediating harms stemming from scams, so most comments were made in this context. The most frequent type of remediation advice given was the recommendation to *cut off communication* with the scammer by blocking them or otherwise ignoring their messages (198 occurrences). This advice was especially pertinent in the context of sextortion or other threat-based scams. For example, in response to a 16-year-old user who requested support after being extorted by a government impostor, a user advised them to cut off communication immediately, explaining that the police would not ask for money over the phone. This recommendation was sometimes given alongside the explicit recommendation not to pay the scammer any money (74 occurrences of *don't pay scammer*). For example, a user who sought support for sextortion against their boyfriend was advised that he should ignore them and not send any money. The commenter explained that the extortionist's demands would only escalate and that paying them off permanently was impossible. This advice is generally well-founded and in line with recommendations from law enforcement [495].

Other common remediation advice was to *contact the financial institution/payment processor* associated with a scam (65 occurrences), or contact the provider of a service associated with a scam (55 occurrences of *contact service provider*). In the context of financial institutions/payment processors, this refers to things like disputing a transaction or locking a credit/debit card. These steps may allow a victim to recover lost funds or prevent the use of a compromised account. In the context of service providers, comments typically recommend that the victim report the scammer using the features built into services. This recommendation does not necessarily help the victim but does help platforms identify and remove scammers.

Recommendations to *contact law enforcement* (21 occurrences) or other government entities (10 occurrences) were less common. When reporting to law enforcement was discussed, it was occasionally described as pointless or ineffective. For example, one user made a post to `/r/Scams` asking how to file a sextortion complaint, stating that they wanted to harm the perpetrator legally. The first comments posted were dismissive of the idea, with one user stating that the scammer was likely untraceable due to using a stolen account. Another commenter jokingly

compared sextortion to learning one has a terminal disease, stating that the poster needed to accept the situation and move on. While commenters are correct that bringing a scammer to justice is unlikely, reporting to law enforcement still provides a societal good by making authorities aware of the most pressing threats, which may lead to broader action against common scams.

Victims were also sometimes encouraged to *contact personal connection(s)* (33 occurrences), like family members or friends. This was often suggested in the context of sextortion to alert others to the risk of intimate imagery being released. For example, one commenter advised a victim of sextortion that they should tell their friends to block the scammer and ignore any messages. They suggested that the poster could claim the scammer used a deepfake app to generate the nudes to protect their privacy. Especially where users expressed serious emotional distress, it was also suggested that users may reach out to close others who can provide emotional support, including therapists. For example, one prior victim of sextortion responded to a new victim requesting emotional support by referencing their personal experience and encouraging them to talk with someone they feel safe with, such as a therapist or friend. Victims of sextortion may be reluctant to discuss their experience with others due to the stigma associated with victimization; however, it can be helpful to mitigate the harm from exposure and cope with the trauma.

Users were relatively infrequently provided with technical remediation advice. This result is unsurprising, as most schemes we observed were technically simple, involving internet technologies primarily used for communication. The most common form of advice was especially given to victims of sextortion, who were advised to *change security/privacy settings* (45 occurrences). This was often phrased vaguely, such as one user's recommendation that another user should adopt the most stringent privacy settings available before logging back into social media. Users sometimes pointed to specific settings. For example, one commenter told a victim of sextortion that they should disable the ability for others to tag them in photos and turn off message requests on Instagram. Changing these settings would make it more difficult for the extortionist to harass the victim.

The next most common technical remediation advice was to change an identifier, such as phone numbers or social media usernames (44 occurrences of *change identifier*), or disable/delete one's social media account (43 comments). These serve a similar purpose as changing privacy settings, making it more difficult for a scammer to contact a victim, although they may also harm a victim's ability to interact with others.

General security/scam avoidance advice

Security and scam avoidance advice were most often provided in response to victims to help them avoid future harm. For example, users who lost money were often warned about the potential for revictimization by "recovery" scams. (52 occurrences of *watch out for recovery services/scams*). These scams target crime victims by falsely claiming to help recover lost money or prevent the release of intimate imagery [500].

Recovery scams seem to be prevalent on Reddit, as some users responded to warnings about recovery scammers by sharing that they had been contacted by these scammers. For example, a user who requested support after their family member was victimized by an investment scam shared that they had already received a message from a supposed ethical hacker who promised to help get the lost money back. Subreddit moderation seemed to be effective at removing com-

ments from scammers on public threads, with only 3 occurrences of comments advising users to contact recovery scammers. Still, this result demonstrates the risk of publicly seeking support for scam victimization.

Users were also commonly advised on how to avoid ecommerce scammers. This result is unsurprising, considering the large number of posts made related to ecommerce scams. We grouped the advice into two broad categories: advice on payments (36 occurrences of *advice on ecommerce payment*) and advice on selecting ecommerce websites (10 occurrences of *advice on where to shop*). With respect to payments, users were instructed to use different payment methods based on the context. Most commonly, users were told to avoid services like Zelle that lack fraud protection. As discussed in Section 7.2.3, these payment processors are often exploited by scammers, so steering people away from using them for purchases with strangers is appropriate.

Advice on where to shop was often vague, such as when one user recommended that people should only shop at well-reviewed websites. Users sometimes provided specific guidelines for identifying scam websites. For example, some users pointed to the age of domain registration as an indication of fraudulence. As fraudulent ecommerce websites are often short-lived [50], avoiding ecommerce websites with recently registered domains can be a useful heuristic to avoid scams.

Users were also advised not to click on links (6 occurrences of *don't click links*) or otherwise interact with unsolicited communications (12 occurrences of *don't answer communications from unknown sources*). For example, one minor who posted asking for support after receiving an unsolicited instant message containing pornographic content was told they should report the message and block the number that sent it. The same commenter also encouraged them not to click links from unexpected texts. This advice is commonly given to help users avoid phishing or other malicious websites [236].

Technical security best practices like multi-factor authentication (10 occurrences of *enable MFA*) or using unique passwords (6 occurrences of *use unique passwords*) were only rarely recommended. For example, a poster who requested advice after their family member fell for a customer support scam was told that their family member should use two-factor authentication and unique passwords as a precaution. Based on Redmiles et al.'s evaluation of security advice, these are considered by experts to be among the most important steps users can take to protect themselves from security and privacy harms [395]. In the aftermath of scams where personal information has been disclosed, these steps can help prevent account compromise.

Bots played an important role in the support process

Throughout the coding process, we noted that bots play an important role in providing support on `/r/Scams` and `/r/Sextortion`. In early 2018, the moderators of `/r/Scams` configured a tool called AutoModerator to provide explanations of common scam scenarios. For example, a user would type the command "!fakecheck" and the AutoModerator would provide a detailed explanation of the use of fake checks in scams: "...The fake check scam arises from many different situations... but the bottom line is... you receive a check... you deposit [a] check and see the money in your account, and then you use the funds to give money to the scammer... The bank will take the initial deposit back... and any money you sent to the scammer will come out of your

own personal funds" *(verbatim)*. Most often, this was used to explain the type of scam a user encountered in the context of the discussion of scam communications, scam identification, requests for support, or reports of victimization (i.e., comments coded as explain scam). The content of the explanations and the number of scenarios the bot could explain evolved over time. Some of the explanations provide advice on recovering from the scam. For example, an automated explanation of the advance fee scam from September 2021 stated, "...If you are involved in an advance-fee scam, you should attempt to dispute/chargeback any payments sent to the scammer... and you should ignore them if they attempt to contact you again" *(verbatim)*. Bots, and particularly AutoModerator, are widely used by communities across Reddit, although they are primarily used for administration tasks, such as screening posts for rule-breaking content [243, 302]. In this context, AutoModerator is being utilized as an informational tool that makes it simple for users to provide a standard explanation and basic advice for common schemes.

We also observed the use of AutoModerator to provide initial information to posters on `/r/Scams` and `/r/Sextortion`. Starting in early 2023 on `/r/Scams` and mid-2022 on `/r/Sextortion`, every post received an automated comment. On `/r/Scams`, this serves to warn users about the rules of the community and to encourage reporting of any scammers in the comments. This can be viewed as a form of "proactive moderation" that is designed to reduce rule-breaking content before it occurs [424]. On `/r/Sextortion`, these comments point users to a "new victim" guide that provides reassurance and resources to victims. In particular, it discusses the low likelihood of exposure with reference to a study of `/r/Sextortion` from the Canadian Centre for Child Protection [85], encourages users to seek professional mental health care if necessary, provides links to report sextortion to various law enforcement agencies and NGOs, and warns users about scammers that may attempt to revictimize them. While not necessarily comprehensive, this post serves as a useful starting point for sextortion victims to understand and remediate their situation. In total, 139 comments on the posts we reviewed came from the AutoModerator. This number includes comments unrelated to support (e.g., messages informing a user why their comment has been removed).

Differences in support between `/r/Sextortion` and `/r/Scams`

`/r/Sextortion` was generally more supportive than `/r/Scams`. While somewhat subjective, this difference can be illustrated through several coding results. `/r/Sextortion` users commonly offered to direct message (DM) users for emotional support. For example, one user responded to a sextortion victim who expressed suicidal ideation by encouraging the victim to message them and sharing that they also deal with depression, so they would not be judgmental. There were 25 occurrences of *you can dm me* on `/r/Sextortion` vs. 3 occurrences on posts from `/r/Scams`. Victims also seemed to be more likely to be chastised on `/r/Scams`, especially where they struggled to understand the deception they faced. In response to a victim of a fake payment scam who doubted that they had been scammed, a user wrote that they were foolish and called the scammers' ruses unbelievable. There were 24 comments assigned the code *chastise* on posts from `/r/Scams` vs. 1 comment on posts from `/r/Sextortion`. While we did not systematically analyze comments from scam identification posts, in the first round of coding, we also noticed that users were sometimes chastised for asking about scams that the commenter perceived as obvious. For example, one user made a post to `/r/Scams` asking about

whether a sugar mama was real or a scammer. The framing of the post indicated that the poster suspected that it was fake, but multiple commenters chastised them for even entertaining the idea that anyone would pay them for a relationship.

These comparisons are qualitative and should be interpreted with caution. The number of annotated posts from /r/Sextortion is smaller than that from /r/Scams, so the observed differences may not generalize to all posts in either subreddit. Nonetheless, the thematic contrasts suggest differences in how each community responds to scam-related disclosures. Beyond the fact that victims of sextortion were more likely to request emotional support, one potential explanation for the difference between /r/Scams and /r/Sextortion is the increased prevalence of prior victims providing support. Many users on /r/Sextortion mentioned their prior experience with sextortion when offering support. For example, one sextortion user began a comment offering reassurance to a minor victim by sharing that they were the same age and were also currently dealing with an extortionist. In total, we coded 102 comments as referencing a *similar experience* on /r/Sextortion as compared to only 43 comments /r/Scams. This greater level of victim participation in the support process likely leads to more comments coming from users who understand the experience of victimization and are less likely to blame victims.

This result is similar to Marshall's observations of the treatment of some victims in romance scam discussion groups on Facebook [310]. While some victims were treated with pity (the victims who were "deserving" of sympathy), others were made fun of as lacking common sense or being ignorant (the victims who were "undeserving" of sympathy). Marshall tied this to gender differences, with men generally being treated worse than women, and the extent to which victims were perceived as behaving immorally (e.g., if the victim was cheating on their spouse). While we did not observe the same gender dynamics, the overly critical treatment of victims may undermine the value that these communities have as a support group.

Support given to Reddit posters helped other internet users

Throughout all phases of the coding process, we noted several occurrences where a post about a scam led to discussion from other internet users who encountered the same scam, including requests for support not from the original poster (22 occurrences of *Non-OP request for support*). In response to a /r/Scams user reporting being defrauded by a Reddit user, three other commenters reported being scammed by the same person. Users were aware of this benefit of posting, with some posters deliberately adding the URL of a scam website or the name of an entity associated with a scam (e.g., username, company name, etc.) to a post to ensure that it was discoverable by other users. In a typical instance, a user requesting support with a crypto investment scam appended an edit to their post sharing the link to the scam website. They advised others not to click the link, but stated they wanted Google to recognize the URL as a scam. In recent years, Reddit has become increasingly prominent on Google [350], so it is unsurprising that the top results for scam websites include /r/Scams. In this way, Reddit serves as a de facto review website, protecting a wider range of users from harm.

7.3 Discussion

In this section, we summarize the results of our analysis and synthesize recommendations. In particular, we discuss the need for more automated tools to help users avoid scams, encourage future research based on Reddit scam discussions, and suggest greater law enforcement outreach on Reddit.

Through our analysis of four communities dedicated to scam discussion, we found that Reddit is used to help users identify scams, discuss the strategies used by scammers, and provide support to victims. The types of scams discussed on Reddit broadly reflect the types of scams most often reported to government entities, with a large proportion of posts focusing on ecommerce scams or corporate impostor scams (RQ1). Victims of scams often request general remediation advice, although many users also seek reassurance that they are safe from harm. `/r/Sextortion` especially serves as a support group for victims of financial sextortion (RQ2). Our analysis of Reddit comments suggests that these communities specialize in providing informational support and reassurance, with technical security advice rarely being needed or provided (RQ3).

Our results reveal challenges in moderating public support groups for scam victims. Prior work has noted many problems faced by volunteer moderators of online spaces, including an inability to keep up with the large volume of posts [243] and difficulties in protecting their communities from hate/harassment [469]. Uniquely, moderators of scam communities must ensure that victims are not re-ictimized after posting about their experiences publicly. The public nature of these communities lowers the barrier to receiving support, but also means that scammers can easily find users who may be vulnerable to false promises of financial recovery. On `/r/Sextortion` and `/r/Scams`, users were protected through proactive warnings about recovery scams and apparently vigilant moderation of comment sections to remove scammers' messages. These protective strategies are especially relevant to the scam context, but may also be applicable to communities focused on victims of other forms of crime, such as survivors of intimate partner violence [360, 427].

Future work should build on these insights by interviewing or surveying moderators to better understand the practices involved in managing these sensitive spaces and whether there are opportunities for improvement. A content analysis focused on moderation-related comments could also provide greater insight into the moderation of scam-related spaces, although looking at publicly visible actions alone is likely insufficient to understand the full breadth of moderator labor [291]. While the largest communities we investigated (`/r/Scams` and `/r/Sextortion`) seem to already be excluding scammers from direct participation, a study focused on the moderation challenges could reveal opportunities for policy changes and new tooling in scam discussion spaces.

Our results also highlight the need for automated tools to help users reason about scams. The prevalence of posts seeking to identify potential scams reflects the difficulty many users have in distinguishing between legitimate and fraudulent communications online. While Reddit's scam communities serve as a useful resource, we note several drawbacks to relying on crowdsourced scam advice. In addition to the risk of targeting by scammers, posters were sometimes chastised for their inability to recognize scams or blamed for being victims. While the best solution is to prevent users from encountering scams, tools that aid users in distinguishing between scams and benign messages could fill the gap left by current automated scam filtering techniques.

Generative artificial intelligence (GAI) models present a promising path towards providing automated, context-specific advice on scams. Prior work has shown that GAI chatbots are effectively able to answer security-related questions [154] and categorize phishing messages [263, 416]. These results suggest that a tuned chatbot may be able to provide much of the same informational support without exposing users to harassment or the potential of revictimization. Indeed, the use of AutoModerator on `/r/Scams` illustrates a basic form of automation in the support process. GAI tools could provide more specific, context-sensitive explanations than are presently possible with the template-based approach of `/r/Scams`.

The following two chapters focus on developing a GAI-based warning system that provides informational support for SMS scams. This system focuses on informational support; however, our results reveal the nuanced nature of scam-related help-seeking. Users often posed open-ended requests for support, rather than specific technical queries. Posters also needed emotional support alongside procedural advice about remediating scams. While AI chatbots are being increasingly adopted for mental health care solutions [213, 220], their effectiveness in responding to emotionally charged scam-related queries remains unclear. Those developing GAI-powered anti-scam interventions, including service providers such as Google [36], should consider our results in the design and evaluation of such tools. This includes assessing whether GAI-generated responses appropriately balance empathy and clarity, and testing how users respond to different explanation styles. Regardless of the quality of AI-generated support, online communities will remain important, as users may prefer receiving support from other humans.

We encourage future work based on Reddit scam discussion. Through their discussion, Reddit users have crowd-sourced a rich corpus of examples of scams that could be mined to create datasets of scam communications or measure patterns in scams over time. Our results indicate that Reddit discussion is especially useful for studying scams involving ecommerce sellers, sextortion, or corporate impostors. Posts categorized as discussion of scam communications are the most useful for this purpose, as they often include transcripts or images of scam communications. To our knowledge, only the study by Bitaab et al. has used Reddit discussion from `/r/Scams` for measurement purposes [50]. Prior work has used other social media platforms. For example, Nakano et al. developed CrowdCanary, a system to identify phishing URLs from English and Japanese reports posted on X (formerly known as Twitter) [337].

Measurement research of scam discussion on Reddit would not only be of academic interest, but also could give stakeholders, including the FTC, Anti-Phishing Working Group,[5] or Global Anti-Scam Alliance,[6] greater insight into emerging threats that are under-reported via traditional means. Any research must consider the risks to users discussed in Section 7.1.3. Moreover, researchers should also consider sharing their results with the communities they measure or otherwise including community members in the research process [178]. Considering the use of the Canadian Centre for Child Protection's analysis of victim narratives[85] in its "New Victim's Guide," we believe that users will appreciate the insight into their experience that research provides.

We also suggest greater outreach from law enforcement agencies and other government entities to communities on Reddit. While users were sometimes encouraged to report crimes—

[5] `https://apwg.org/`
[6] `https://www.gasa.org/`

most notably in the /r/Sextortion "New Victim's Guide"—there was a pervasive attitude of futility regarding the benefits of reporting. Some users who reported scams on Reddit expressed a desire to protect others [128], so communications from stakeholders could acknowledge the difficulties in bringing scammers to justice while emphasizing the broader impact that reporting has. Communications could also clarify the most appropriate authorities to report to for internet crime, as confusion around jurisdictional issues can pose a barrier to reporting, especially in federal states like the United States or Australia [129, 131]. While such advice exists in other forms on the internet (e.g., [517]), it is valuable to meet users where they are. Within the United States, this outreach would most appropriately come from national-level entities like the FTC or FBI, which have the jurisdiction and resources to investigate internet fraud.

Chapter 8

Investigating User Reasoning with Context-Specific SMS Scam Warnings

The previous chapters of this thesis have focused on understanding users' experiences with scams. In response to the evident need for more user-centric anti-scam interventions, this chapter shifts focus toward developing solutions to better protect users from harm. In particular, it presents the first in a sequence of studies meant to develop generative AI (GAI)–based contextual scam warnings for SMS scams.

As discussed in Section 2.1 instant messaging remains a popular venue for fraudsters to spread scams. According to the US Federal Trade Commission, text-message-based scams caused a reported $470 million of losses in 2024, which was a nearly 500% increase from 2020 [502]. A typical pipeline to combat these scams is to first automatically detect them, and then either block them from delivery or flag them with a warning. These systems (notably, Google Messages[1]) have opted for simple warnings that do not provide much detail about *why* a message is flagged. This approach runs contrary to best practices, which recommend that warnings provide users with the evidence necessary to help them make their decision [34, 396].

Historically, incorporating context-specific explanations has been challenging, due to the black-box nature of AI reasoning [526]. Generative AI (GAI) offers new opportunities to overcome these barriers. Modern language models can extract rich contextual cues, such as the named sender, brand mentions, or suspicious URLs, and articulate human-like explanations. With appropriate prompting, these models can highlight linguistic signals (e.g., urgency, mismatched URLs) and generate concise, context-specific warnings that may help users make better decisions. The preceding chapter showed that users frequently turn to online communities to identify scams and cope with victimization, an informal support process that is valuable but inconsistent and can expose victims to further harm. Automated advice may offer a way to deliver similar informational support at the point of risk. Indeed, Google has recently deployed tailored scam warnings [36].

Recent studies on explainable phishing detectors have shown that LLM explanations can improve task performance [90, 143, 295, 534]. However, these efforts have focused primarily on

This chapter is based on a paper originally published at the IEEE Symposium on Security and Privacy 2026 and was co-authored with Enze Liu, Alexandra Xinran Li, and Lorrie Faith Cranor [60]

[1] `https://messages.google.com`

the explanations themselves, evaluating their accuracy and linguistic quality, rather than on how such explanations function when embedded within security warnings. Yet in practice, explanations reach users through warning interfaces that must balance brevity, visual salience, and trust. How these design factors shape users' reasoning and behavior remains poorly understood in the context of SMS. We address this gap by asking:

1. How do users reason about instant messages from unknown senders in the presence of warnings?
2. How do different warning design choices, especially the integration of context-specific explanations of varying lengths, affect users' reasoning?

To answer these questions, we present a task-based interview study ($n = 20$) using high-fidelity mockups of the Google Messages app to examine how people interpret and respond to scam-related SMS warnings. Across interviews, participants primarily reasoned based on message features, such as URLs and linguistic errors, their prior experience, and expectations of "normal" communications. Warnings supported this reasoning in multiple ways: sometimes serving as a decisive signal, other times offering reassurance or highlighting specific cues and actions. Participants valued the context-specific direction and evidence generated by GAI, but differed in their preferred level of detail and reactions to AI disclaimers. Visual and interactive elements such as red iconography and one-tap "Report & Block" options further influenced perceptions.

This chapter makes two main contributions. Empirically, it presents the first in-depth qualitative study of how people reason about SMS messages in an inbox-style interface when exposed to multiple, realistic warning designs, including false positives. It shows how warnings influence decision-making in multiple overlapping ways. From a design perspective, it reveals that users prefer context-specific, action-oriented warnings, while highlighting how design parameters, including length, tone, AI disclaimers, and visual styling, can either support or undermine effective responses.

8.1 Methods

We conducted an online interview study with 20 US-based participants to understand their reasoning about SMS messages in the presence of warnings. Below, we describe our recruitment process, interview protocol, selection of benign and phishing messages, warning design choices, data analysis, ethical considerations, and limitations.

8.1.1 Recruitment

We primarily recruited participants from Pittsburgh, PA. To avoid alerting participants to our focus on warning design, the recruitment material specified that our study aimed to understand how users respond to messages from unknown senders. We posted physical posters on the campus of Carnegie Mellon University, at local libraries, and on utility poles around the city. To reach older adult participants, we distributed a digital version of the poster on the mailing lists of lifelong learning centers at Carnegie Mellon University and the University of Pittsburgh. We also distributed the digital poster to personal social networks and Slack channels. People were

encouraged to share the poster with others who may be interested in the study, even if they were not eligible themselves. Finally, we placed posters on the campus of the University of California, San Diego.

Recruitment materials directed potential participants to complete an informed consent form via a link on the poster. This form explained the procedure of the study and required potential participants to affirm their eligibility for the study (18 years of age or older, use an Android phone to send and receive instant messages, and be able to scan a QR code with their phone to view images). Eligible participants were directed to a survey to collect demographic and contact information for scheduling the interview. The demographic survey is available in the Appendix F.1.

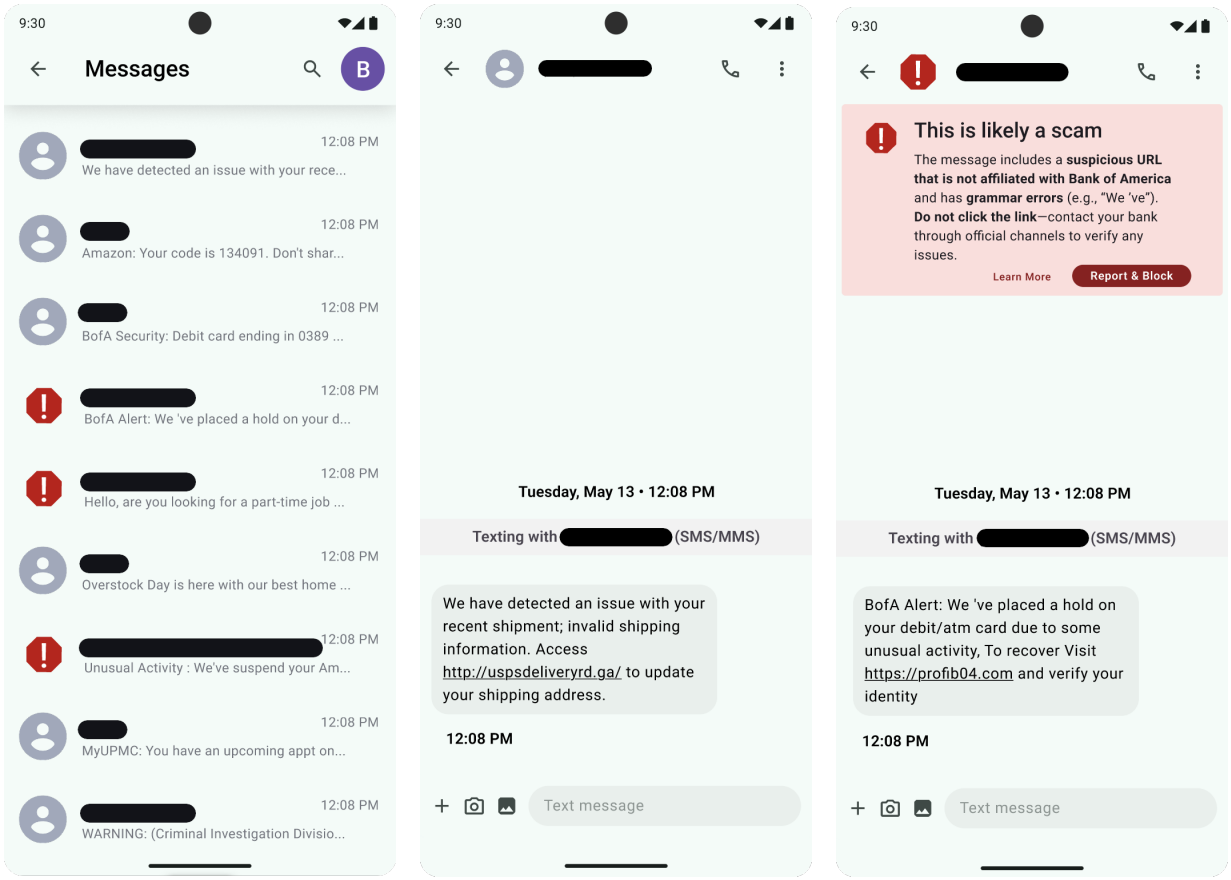
Not all those who filled out the survey were invited for an interview. We purposively selected participants based on demographic characteristics. We sought to ensure that our sample included individuals from a diverse range of ages and levels of computing education or work experience. Age is a frequently investigated demographic variable in studies on phishing vulnerability, although it is unclear to what extent it affects users' susceptibility [540, 570]. In total, 53 people completed the demographic survey, 26 people were invited to participate in an interview, and 20 completed the interview. At the conclusion of the study, we discarded the demographic and contact information of those who did not participate in an interview. Consistent with similar qualitative studies [470, 569], we planned to recruit 20 to 30 participants. The research team regularly met to discuss emerging themes and whether the most recent interviews had generated new insights. Upon reaching the twentieth interview, the team agreed that the recent interviews had not yielded new ideas, indicating that theoretical saturation had been reached and that additional interviews would likely not yield new insights [428].

8.1.2 Interview Procedure

All interviews were conducted over Zoom by a single author, who also took notes during the sessions. The interviews were semi-structured, meaning the interviewer asked a set of standard questions, along with follow-up questions tailored to participants' responses. The interviewer conducted four pilot interviews to practice the interview script and ensure that questions elicited informative responses. The pilot interviews are not included in our final analysis.

The final interview guide is available in Appendix F.2. Before each interview session, the interviewer briefly discussed the consent form to ensure that the participant agreed to being audio-recorded and understood the voluntary nature of their participation. The interviewer then asked a series of background questions (Q1 – Q5) about participants' mobile phone use.

The bulk of the interview time was dedicated to a message-processing task, using a within-subjects design. Each participant viewed a total of 12 text messages (six benign and six fraudulent) and was exposed to five different warning designs. The message stimuli and warnings are described in greater detail in subsequent subsections. The first two messages in each inbox were intended to warm up participants to the task: they were always one benign message and one scam message, with no warnings. The remaining messages were randomly ordered to control for order effects. The interviewer instructed participants that they were viewing messages received by a mobile phone user named Bob and gave some details relevant to the task (i.e., where he lives, the



(a) Inbox viewed by P8

(b) *Fake delivery notification*

(c) *Fake Bank of America alert* & long/confident contextual warning

Figure 8.1: Example screens displayed in the study. Redactions were not present during the study and are used to hide phone numbers.

name of his bank, and the delivery services he uses). Messages were presented through a Figma[2] mockup of the Google Messages app (examples in Figure 8.1), which is the most widely used messaging app on Android. For each message, participants were prompted to tap on the message and review it. The interviewer then asked them to recommend an action for each message and explain why (Q6 – Q9). To avoid priming for deception, we solicited open-ended recommendations rather than legitimacy ratings. Participants were not restricted to a limited set of options and were encouraged to recommend any action they might normally take. For messages presented with a warning, the interviewer prompted the participant to explain whether the warning affected their reasoning and to provide feedback on the warning.

To approximate real-world conditions, most participants viewed the inbox on their personal mobile devices. Due to technical issues with her phone, P20 reviewed the messages through a screen share of the mockup, which the interviewer controlled. P3 viewed all but one message on

[2]`https://www.figma.com`

their personal device. The final message in the inbox had to be viewed via a screen share due to an error in the implementation of the mockup, which resulted in a duplicate message being included.

After the message-processing task, participants were asked to provide general feedback on the warnings they saw and to share any other suggestions for improvement (Q10 – Q12). As the intended length of the interview was an hour, these questions were asked only in 14 interviews in which an hour had not elapsed before the conclusion of the messaging task. The average length of the recorded portion of interviews was 58 minutes and 4 seconds.

8.1.3 Message Stimuli

The complete details of all message stimuli can be found in Appendix F.3. We selected fraudulent stimuli from the SmishTank dataset [480], a crowd-sourced smishing dataset that has been used in prior studies [481]. The version of the dataset used contained messages received between March 2022 and November 2023. We chose messages to reflect a diversity of common scams and different social engineering tactics that users encounter, as reported by Tabassum et al. and Timko et al. [470, 480]:

- **Fake Amazon Account Alert**: Claims the user's Amazon account was locked after a login from Turkey and urges them to click a link. Sent from an email-to-SMS gateway. Contains grammatical errors.
- **Fake Bank of America Alert**: Claims the user's account was locked and directs them to a shortened link. Contains a minor grammatical error.
- **Fake Delivery Alert**: Claims that a delivery is on hold due to invalid shipping information. Directs the user to click a link impersonating USPS.
- **Employment Scam**: Offers a part-time Amazon sales job paying $30–$200 per day for one hour of work. Directs users to reply via Telegram or a UK-based WhatsApp number.
- **Fake T-Mobile Giveaway**: Claims to offer a T-Mobile gift and links to a shortened URL. Contains a small typographical error (missing dash in "T-Mobile").
- **IRS Arrest Threat**: Claims to be from the IRS and threatens arrest unless the user calls a southeastern US number. Includes spelling errors and awkward phrasing.

We selected benign messages that were legitimate counterparts to the fraudulent messages or messages that prior work suggests users may have difficulty identifying as legitimate, such as political campaign messages [470] and patient-doctor communications [530].

- **FedEx Notification**: Confirms a package delivery and links to an image of the shipment.
- **Healthcare Reminder**: Alerts the recipient to an upcoming appointment with a link to details.
- **Amazon OTP**: Provides a one-time passcode and includes a link for users who did not request it.
- **Overstock.com Promotion**: Advertises a 20% off sale with a shortened link to the promotion.

- **Bank of America PIN Alert**: Notifies the user that a card transaction was declined for missing a PIN and advises retrying.
- **Political Solicitation**: Urges support for a local candidate via a personalized message and link.

The doctor's appointment reminder and political solicitation were selected because they were relevant to Pittsburgh, PA, where the majority of participants were recruited.

8.1.4 Warning Designs

We evaluated five warning designs within our interview script. The first two warnings are employed by Google Messages and were selected to represent the status quo of generic SMS warnings. The last three warnings followed the same visual design with varying levels of contextual details and confidence. While other dimensions of warning design exist, we chose to focus on the level of contextual detail and confidence based on the results of Wang et al. [534]'s study of LLM-based SMS classification. Explanation length emerged as a common concern among participants in their work, with some feeling that the "short" explanation they presented was too long. By varying the level of contextual detail, we sought to directly investigate how different lengths of explanation affect a user's experience. Furthermore, we investigated different confidence levels to probe how framing affects users' trust calibration. This was motivated by Wang et al.'s findings on over-reliance on AI-generated explanations, in which 18 of 24 participants reversed their correct initial classification of phishing after reading an incorrect AI explanation. We sought to investigate whether more cautious framing could mitigate this over-reliance by signaling the system's uncertainty and encouraging verification. Examples of each notice are shown in Figure 8.2:

- ***Save contact notice***: Default Google Messages notice for unknown senders, offering options to save the number or report as spam.
- ***Spam warning***: Google Messages notice for messages flagged as spam, stating the message resembles other spam messages and offering an option to mark it as not spam.
- ***Generic scam warning***: Prototype warning stating a message is likely a scam, with no explanation but options to "Report & Block" or "Learn more."
- ***Contextual scam warning (short, uncertain framing)***: Prototype warning that a message may be a scam, with one AI-generated sentence of advice, a disclaimer stating that the message was generated by AI and may not be accurate, and options to "Report & Block" or "Learn more."
- ***Contextual scam warning (long, confident framing)***: Prototype warning that a message is likely a scam, with two to three AI-generated sentences of advice and options to "Report & Block" or "Learn more."

The design of the hypothetical warnings was based on best practices in risk communication [34, 547], emphasizing attention-grabbing color and iconography, clear hazard statements, and explicit recommended actions. Specifically, the warnings used a red stop-sign icon, headlines that explicitly included scam as a "signal word" [547], and GAI-generated explanation text paired with a one-tap "Report & Block" option to convey both the nature of the threat and the

recommended response. To ensure visual consistency with the simulated Google Messages interface, the design was implemented using the Material 3 Design Kit[3] and a color palette derived from the existing application. The prompts used to generate the contextual-warning text are available in Appendix F.4.

Apart from the first scam message that each participant saw, all scam messages were presented with one of the above warnings. The specific warning applied to each scam was randomly selected. One randomly selected benign message was presented with a false positive warning in the form of the *contextual scam warning (long, confident framing)*. The false positive was included to evaluate whether participants could identify benign content even when it was incorrectly labeled as fraud.

8.1.5 Analysis

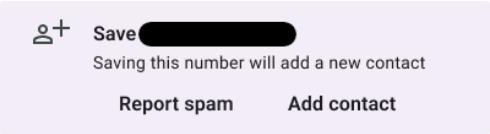
Interviews were audio-recorded and automatically transcribed via Zoom. The interviewer and another author reviewed the transcript to correct mistakes. The interviewer then divided the transcript into analysis units corresponding to sets of related questions and follow-ups. For each instant message, the analysis units were the action recommended by the participant, the participant's reasoning for the action, the effect of the warning (if applicable), feedback on the warning (if applicable), the participant's level of confidence, and whether they had received similar messages before. Analysis units were not always mutually exclusive, since interviewees often discussed topics across different parts of the interviews.

Analysis followed an iterative, inductive coding approach to identify themes. The resulting codebook is presented in Appendix F.5. The interviewer served as the lead coder and developed an initial codebook based on notes from all the interviews and an analysis of three interviews. Two other authors then reviewed the codebook and the coding of these three interviews, providing feedback and refining the codebook. This process was repeated with six additional interviews in groups of three until the codebook stabilized, with no major changes made. The lead coder analyzed the remaining 11 interviews independently. The secondary coders then reviewed the codes and provided feedback. As with previous rounds, this feedback was discussed and integrated to determine the final set of themes. Since coding was a collaborative process where we attempted to reach consensus, we cannot calculate inter-rater reliability [318].

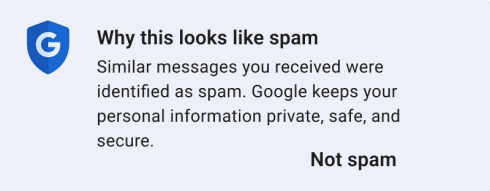
8.2 Ethics Considerations

The study protocol and materials were approved by our Institutional Review Board. All participants completed an informed consent form that explained the risks and potential benefits of their participation in the study. Participants were compensated $20 in the form of an Amazon gift card. While we presented real scam stimuli in our experiment, our mockups were designed such that it was not possible to click any links or copy text within messages. Moreover, all scam URLs were dead links that would no longer function if visited.

[3] `https://m3.material.io/`



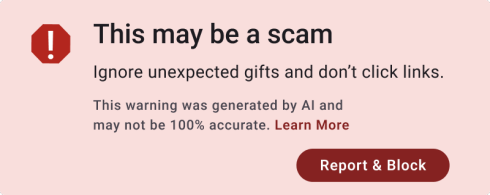
(a) Save contact notice



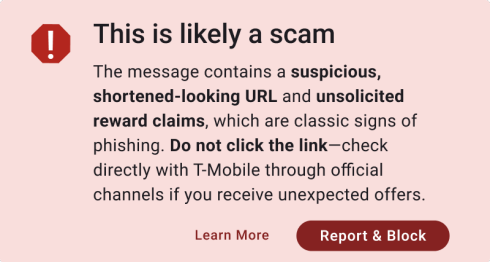
(b) Spam warning



(c) Generic scam warning



(d) Contextual warning (short, uncertain framing)



(e) Contextual scam warning (long, confident framing)

Figure 8.2: Examples of the warnings presented in the study on the T-Mobile Giveaway Scam

8.3 Limitations

Our study offers in-depth qualitative insights into users' instant messaging processing behavior and is not designed to draw quantitative conclusions. We provide code frequencies to contextualize the results; however, these should not be interpreted as statements about the frequency of particular sentiments in the general population. While we attempted to recruit a diverse sample through purposive sampling, we do not claim a representative sample (see Section 8.4.1). Moreover, most participants came from a single city in the United States. This may limit the generalizability of our findings, particularly for users outside Western cultures or those with lower educational attainment. Our study was conducted in a lab environment, so participants may not have reacted to the stimuli as they would in real life. Users' reactions to warnings are also subject to long-term habituation [17, 520] or other effects that may only manifest after repeated exposure. Due to time constraints, participants could view only a limited number of scam scenarios and warnings. Different scam scenarios may elicit different reactions and reasoning, especially considering that the contextual warning text would differ. Finally, while stimuli were randomized to control for potential order effects, participants' perspectives on later stimuli were likely influenced by their experience of earlier stimuli. Participants may also have developed cognitive fatigue due to the study's repetitive evaluations.

8.4 Results

In this section, we explain the results of our interview study. We start by describing our participants. Next, we discuss how participants reason about messages from unknown senders in the presence of different types of warnings. We conclude with an analysis of how various elements of a warning influence users' reasoning.

8.4.1 Participants

Table 8.1 shows the demographic characteristics of participants. Ten participants were between 18 and 34 years old, four were between 35 and 64, and six were 65 or older. In general, our sample was highly educated, with only four participants lacking a college degree. Prior work does not show a link between educational attainment and secure behavior [540], although differences may still exist. This skew toward higher education may impact the study's results, as above-average reading comprehension could affect the processing of warning information. Moreover, six participants reported having work experience or education in computing. This greater level of technical experience could result in an above-average ability to identify the scams within this study. Finally, all participants were white (15) or Asian (4), with two reporting Hispanic or Latino ethnicity, indicating limited racial/ethnic diversity in our sample. People of different racial and ethnic backgrounds may have different experiences with scams, which could affect participants' approach to the messages presented within the study. For example, immigrant communities are targeted by specialized immigration and government-imposter scams [375].

We also asked about participants' use of instant messaging. All participants discussed using instant messaging for *social* purposes, such as "catching up with friends and family" (P8). Most

Table 8.1: Interview participant demographics. A dash indicates that a participant opted not to respond to a particular question. Question text and categories are abbreviated for space. For tech work experience and education, U indicates that a participant was unsure. P17 and P18 additionally selected the option to self-describe, but the information provided was not relevant to race.

ID	Age	Gender	Race	Hispanic / Latino	Computing Edu./Exp.	Education	Income
P1	35 to 44	Woman	White	No	N / N	Bachelor's	$50k to $100k
P2	18 to 24	Man	White	No	Y / N	Bachelor's	> $150k
P3	25 to 34	Woman	Asian	No	N / N	Bachelor's	$25k to $50k
P4	35 to 44	Woman	White	No	Y / Y	Bachelor's	$100k to $150k
P5	18 to 24	Woman	Asian	No	N / N	Bachelor's	<$25k
P6	65 or older	Man	White	No	Y / Y	Graduate	-
P7	65 or older	Woman	White	No	N / N	Graduate	$100k to $150k
P8	25 to 34	Non-binary	White	No	Y / Y	Bachelor's	$25k to $50k
P9	25 to 34	Woman	Asian	No	N / N	Graduate	<$25k
P10	65 or older	Woman	Asian	No	N / N	Graduate	$50k to $100k
P11	18 to 24	Woman	White	No	N / N	High school	-
P12	35 to 44	Man	White	No	U / N	Graduate	<$25k
P13	65 or older	Man	White	No	N / N	Some college	<$25k
P14	65 or older	Woman	White	No	N / N	Graduate	$100k to $150k
P15	45 to 54	Man	White	No	Y / Y	Some college	<$25k
P16	25 to 34	Woman	White	Yes	Y / N	Bachelor's	$25k to $50k
P17	18 to 24	Woman	-	Yes	N / N	Bachelor's	<$25k
P18	18 to 24	Woman	White	No	N / N	Some college	$25k to $50k
P19	25 to 34	Man	White	No	N / N	Bachelor's degree	$25k to $50k
P20	65 or older	Woman	White	No	N / N	Graduate	$50k to $100k

also had experience with the types of messages presented in the study. For example, all participants except P13 reported receiving One-Time Passwords (OTPs) via instant messaging, either for authentication or phone number verification. All except P20, P7, and P5 received *finance* related messages, such as transaction alerts from their bank or receipts from businesses. All except P1, P11, and P5 had experience receiving messages from a *healthcare* provider, like appointment reminders or requests for online check-in. As discussed in the following subsection, these participant characteristics and prior messaging experiences are particularly important, as many participants referenced their personal experiences when reasoning about unsolicited messages.

8.4.2 Reasoning about Unsolicited SMS Messages

In this section, we discuss how participants reasoned about messages during the instant message review task. We first provide a brief overview of participants' assessments of the different messages and describe the actions they recommended in various scenarios. These results provide important context for understanding how the warnings impacted participants' decision-making. We then describe the features participants used to reach their decisions, with a focus on warning-related reasoning. Our results show that participants engaged in a layered reasoning process when evaluating unsolicited SMS messages. Scam messages elicited decisive defensive actions, whereas legitimate messages prompted conditional reasoning dependent on context and prior experience. Across both, participants reported that message features were the primary cues, while warnings influenced decisions by serving as shortcuts, confirmations, or action prompts.

Overview of Message Assessments

Table 8.2 summarizes participants' judgments of each message they saw while explaining their recommended actions. Participants were coded as identifying a message as *legitimate* when they clearly indicated that they believed a message was genuine (e.g., "I would say, 10 out of 10 this is... not a scam. They can be annoying. But it's legitimate.") In contrast, participants were coded as identifying a message as *suspicious* when their response suggested that they believed the message was fraudulent (e.g., "As soon as I read this I'm aware that this is a scam..."). Participants' evaluations were coded as *unclear* when they didn't make an unambiguous statement about whether a message was real or when their reasoning was unrelated to whether the message was genuine.

In general, participants performed poorly at correctly identifying legitimate messages as clearly safe to interact with. This result likely stems from the presence of false positive warnings and the artificial nature of the study environment, where message analysis was a primary task. Some participants noted the lab effect. As P2 observed while reviewing the Amazon OTP message "I'm in a place where... I kind of view these messages with [a] more critical lens." This type of issue is common among usable security lab studies [167, 449], and means that the absolute distributions of participants' judgments or recommended actions should not be interpreted as representative of real-world accuracy. However, the cues participants attended to, and their qualitative reflections on the warnings themselves, remain valuable indicators of how people interpret and make sense of warning information in context.

Table 8.2: The frequency of the different judgment types, the number of action recommendations that were conditional, and the most common features for each message. The numbers in parentheses show the frequency that each message feature was mentioned, regardless of whether the feature increased or decreased suspicion.

	User Judgement of Message				
Message	**Legitimat**	**Unclear**	**Suspicious**	**Condl.**	**Message features frequently mentioned**
Fake Amazon Account Alert	0	0	20	3	url features (15), sender information (10), linguistic features (8)
Fake Bank of America Alert	0	1	19	3	url features (16), request for action (11), linguistic features (10)
Fake Delivery Alert	0	2	18	6	url features (18), personal experience (7), communication channel (5)
Employment Scam	0	2	18	3	communication channel (15), absurd offer (13), sender information (7)
Fake T-Mobile Giveaway	0	2	18	2	absurd offer (12), url features (10), linguistic features (5)
IRS Arrest Threat	0	0	20	3	communication channel (14), linguistic features (11), sender information (9)
FedEx Notification	6	10	4	14	url features (11), personal experience (8), linguistic features (7)
Healthcare Reminder	10	4	6	9	personal experience (13), url features (11), linguistic features (9)
Amazon OTP	8	7	5	14	personal experience (17), request for action (8), url features (8)
Overstock.com Promotion	1	9	10	11	url features (18), unfamiliar organization (6), personal experience (5)
Bank of America PIN Alert	8	3	9	16	linguistic features (10), personal experience (8), request for action (7)
Political Solicitation	11	9	0	12	url features (12), personal experience (8), sender information (6)

Participants' action recommendations often defied mutually exclusive categorization. Many responded with multiple, conditional, or even contradictory actions. For example, P13 initially laughed at the Employment Scam, stating "I would block this." While explaining his action recommendation, however, he added that he might actually click the link within the message or attempt to find the job posting on the company's official website, "if I was looking... to work at some job like this." To capture the full breadth of decision-making, action frequencies reported throughout this section include all actions mentioned by participants.

Despite this complexity, qualitative differences emerged by message type. In line with their general assessment of scam messages as suspicious, participants typically recommended defensive actions. In the majority of cases, participants recommended reporting a scam message (76 occurrences of *report* out of 120 observations of scam messages) or blocking the sender (63 occurrences of *block* on scam messages out of 120 observations of scam messages). This result contrasts findings from Tabassum et al. [470], where over half of the participants stated that they generally would not report scam messages. As discussed later, this difference likely results from the presence of warnings that highlight the report option. Two exceptions are the Fake Bank of America Alert and the Fake Delivery Alert, in which participants often recommended contacting the relevant service externally to confirm there was no issue, even when they were confident the message was a scam. For example, upon viewing the Fake Bank of America Alert, P1 stated, "I would probably think that it is fake, but if I was concerned I would either go to my app or website and log in to verify or pick up the phone and call the bank." This indicates a preference for actions that guarantee safety, even when the deception is apparent.

For legitimate messages, participants frequently indicated that their responses would depend on external context not provided in the message, regardless of whether they believed the message was legitimate (76 occurrences of *conditional* out of 120 observations of legitimate messages). For instance, they might *ignore* a message or *interact* with a message based on the situation they were in. For the FedEx Notification, the majority of participants indicated that their decision would depend on whether they were actually expecting a package. P13 stated, "If Bob had actually ordered this package and was expecting it, I guess I would click through just to see... But if I hadn't ordered it... I wouldn't click." Similarly, most participants indicated that their response to the Amazon OTP message would depend on whether they were trying to access Amazon at the time. P19 stated, "if Bob is on amazon.com attempting to log in just type [the code]." When asked what they would do if Bob wasn't logging in at the time, they said, "I think I'd probably just delete this text because I would be concerned about clicking the link." For the Political Solicitation and Overstock.com Promotion, actions were frequently conditioned on individual interest or on whether Bob had signed up to receive these messages. P9 concluded about the Political Solicitation, "I don't think this warrants a reply, but if I wanted to read what they're trying to do and participate, that would be it." Tabassum et al. similarly observed that user decision-making about SMS messages was highly contextual, with particular life situations making messages more or less credible [470]. The frequency of conditional actions for all messages is shown in Table 8.2.

Cues to Legitimacy or Deception

Table 8.2 lists the most frequent message features cited by participants, many of which overlap with those emphasized in the contextual warnings (e.g., URLs and language features). While

the presence of contextual warnings may have increased the salience of these cues, participants identified them spontaneously, including before encountering any contextual warnings. These reasoning patterns have also been documented in prior studies [470, 481]. Below, we summarize the main themes. Where relevant, we indicate which warning condition was shown to contextualize user responses. Benign stimuli were viewed without warnings unless otherwise specified.

URL features were the most common indicator of deception across all messages. Participants frequently focused on mismatched or shortened links. For example, in the Fake Amazon Alert and Bank of America Scam, many participants noted that the links did not include any references to the companies they were supposedly associated with. As P9 said when viewing the message with the save contact notice, "there's the website that they've given to unlock the Amazon account. It's again not an Amazon website, so super suspicious." The vast majority of participants (17) found the shortened URL in the benign Overstock.com Promotion suspicious, especially given its lack of connection to the company. As P16 said, "The link looks weird. It doesn't look like it even has the name of the store there." Tabassum et al. similarly observed confusion around shortened URLs [470]. URLs with legitimate top-level domains also sometimes drew suspicion. For example, four participants indicated that the URL within the FedEx Notification was suspicious, despite the domain being FedEx.com. P14 referred to parameters in the URL, stating, "I don't know. I've never seen a message that had that T and the N and the underscore U S... I just get this feeling it's not right." Such comments indicate that URL appearance was a highly salient heuristic, although it was sometimes inconsistently applied, aligning with prior work on phishing URL judgments [9].

Linguistic features, such as grammatical errors or perceived awkward phrasing, were another frequent focus. For example, 10 participants noted grammatical errors in the Fake Bank of America Alert message when explaining their recommended actions. While viewing under the save-contact notice, P1 stated, "the text message looks careless and suspicious with the ... bad spacing and the bad punctuation." While commonly emphasized in anti-scam education, these types of cues can be unreliable, making well-crafted scam messages seem real or casting doubt on benign messages with perceived errors. For instance, the Healthcare Reminder message concluded with the statement that recipients could "Reply XRemind to opt out." Several participants (5) pointed to this as unusual or even suspicious. P19 stated, "It looks like a standard UPMC visit reminder, but it's got some typos at the bottom or confusing text, so it could be a scam either way."

Looking beyond specific message features, participants frequently discussed their *personal experience* and resulting expectations about what was normal. This code was the most common cue to legitimacy discussed by participants, and it was often paired with specific message reasoning codes. For example, 12 participants stated that their personal experience with healthcare appointment reminders made the Healthcare Reminder seem legitimate. When summarizing her reasons for recommending that Bob interact with the message, P4 stated, "It looks like other appointment reminders I've received. It has that little label at the beginning... because sometimes services share the short codes... It looks like the right format." This quote was coded as referencing both *linguistic features* and *personal experience*. Even when not referring specifically to *personal experience*, participants also discussed other specific expectations about messages. For many participants, the IRS Arrest Threat message was clearly ridiculous due to the *communication channel* (14 participants). As P6 stated when viewing the message with the contextual

warning (long, confident framing), "IRS doesn't send text messages. Period. End of discussion." Participants also pointed to the presence of *absurd offers* in the Employment Scam (13 occurrences) and the Fake T-Mobile Giveaway (11 occurrences).

Effect of Warnings

Participants saw warnings on half the messages they reviewed, including five of the scam messages and one legitimate message (120 observations). In approximately half of these cases, participants mentioned the warning in their reasoning (68 instances of *warning (unprompted)*), and the exact frequency varies by type of notice. Only five participants mentioned the *save contact notice* without being prompted. The *spam warning* was also mentioned by less than half of the participants, with only eight participants describing it as a factor in their reasoning without being prompted. In contrast, the *long/confident contextual warning* and *false positive* warning were both mentioned by 15 participants. These differences in unprompted mentions may suggest that certain warning designs were more salient or influential in shaping participants' reasoning.

In many cases, participants did not clearly attribute an effect to warnings. Nevertheless, we identified four main ways that the SMS warnings helped users' reasoning. First, the presence of the warning was sometimes an important piece of evidence about the legitimacy of a message (38 occurrences of *warning as red flag*). The degree of effect varied. For some participants, the presence of the warning seemed to be the primary deciding factor. For example, when evaluating the Bank of America Scam message with the *short/uncertain contextual warning*, P5 stated, "Well, I'm not even gonna read it... if my phone flags it as a scam, I will just like report and block." P10 also adopted a policy of following the guidance of warnings. In explaining why she would click the report and block button on the *false positive* warning for the FedEx Notification, she said, "What would they have to gain by warning you? Nothing, nothing. There's no ulterior motive. It's for your benefit that they warn you." P19 also demonstrated this type of reasoning strategy across multiple messages, explaining, "Well, I don't like being on my phone.... So one way to reduce the amount of time is to not even bother reading it and just click 'Report and block' immediately... If it really is a serious issue, then they're gonna find some other way to get in touch with me." In these cases, warnings functioned less as a supplement to user reasoning and more as a shortcut, where participants outsourced judgment to the system.

This tendency to defer to the system seems to extend to inaccurate alerts. Participants generally trusted *false positive* warnings, with a little more than half of the participants (11) recommending that Bob report and block the message with the *false positive* warning. Only four participants indicated that they believed the warning they saw was incorrect.

Other participants indicated that they viewed the warning as only one factor among many. When reviewing the Overstock.com Promotion with a *false positive* warning, P18 explained, "I do pay attention to [warnings]. If there's a warning that says this is likely a scam, I'm more likely to pay closer attention to what does the link look like? Any spelling mistakes? It's just more scrutiny." They subsequently explained that they had a *general distrust* of warnings due to bad experiences, "Typically, I have had an email... falsely mislabeled something as a scam... in general it just encourages more scrutiny." While also grouped under the same theme, this kind of reasoning places greater emphasis on users' judgment of the message in conjunction with the warning's content.

In a similar vein, participants sometimes described warnings as providing reassurance or as affirming a conclusion they had already reached (10 occurrences of *affirmed right decision*). For example, P9 explained that even before opening the Fake Bank of America Alert, the grammatical errors present in the inbox preview made it look suspicious: "Even before I opened it, I could see the exclamation mark, and I can see the 'we've placed a hold' so immediately it's like, Oh, this is not a real message. [The warning] just sort of confirms what I already knew." In the case of emotionally charged scams, like the IRS Arrest Threat, a warning may be emotionally reassuring. In explaining his decision to report and block, P2 stated, "Without the scam message I might be a little bit worried... I wouldn't actually call them, but it would be a little bit like concerning." These examples illustrate the value of warnings, even when they did not decisively change a participant's behavior.

Warnings also served to make certain actions more apparent or efficient (15 occurrences of *highlighted particular action*). This was especially noted as an effect of the *save contact notice* (6 occurrences). For example, when viewing the Bank of America Scam Alert, P1 initially stated that she would "probably just delete it" before seeing the notice and stating, "Oh, I guess there is a 'report spam'. I would probably click that. I like it when they make it easy." The effect of options was less frequently noted for other warnings, but it seems apparent based on participants' recommended actions. Participants recommended reporting and blocking a message in 54 of the 80 observations of warnings with a report and block option present.

Finally, some participants discussed how a warning highlighted specific evidence of fraud (9 occurrences *highlighted evidence*). This was noted for both the contextual warnings and the *save contact notice*. The contextual warnings served to highlight specific suspicious cues within the message. For example, P16 stated, "The pay being ridiculously... generous, for like an hour of work, I might have missed that if I didn't have the information on the message." The *save contact notice* served only to highlight the sender's information or that the number was not saved in their contacts. For example, P2 indicated that the notice influenced their reasoning when viewing the Fake T-Mobile Giveaway, sharing, "Often, like T-Mobile or your provider has already sent you messages... they're not an unknown or a new number."

While one might expect that incorrect evidence in a *false positive* warning would prompt users to question the warning's accuracy, some participants instead rationalized it. For example, P19 credited the evidence within the *false positive* warning as decisive, explaining "It says that... the link is pretending to be from FedEx. So, despite it saying FedEx.com, it would take you to a different location." This may reflect a misunderstanding of the differences between email and SMS phishing. In email, a hyperlink's visible text can differ from its actual destination, but SMS messages display the full URL directly, meaning the link shown is the true destination. This highlights a broader risk of explanatory warnings: when false positives include incorrect or irrelevant evidence, users may treat the added detail as reinforcing the warning's credibility rather than questioning it. In such cases, explanations can make inaccurate warnings more persuasive, rather than helping users recognize inconsistencies.

Taken together, these findings suggest that warnings influenced participants in multiple, sometimes overlapping ways: as decisive signals that replaced deliberation, as one factor within broader reasoning, as affirmations of suspicions or reassurance in stressful scenarios, and as cues that highlighted actions or evidence. While not always the primary driver of decision-making,

warnings consistently influenced how participants interpreted and responded to messages, underscoring their role as credible, context-sensitive supports in scam detection.

8.4.3 Effect of Warning Design Parameters

In this section, we discuss how participants reacted to the different warning design patterns, with a focus on the effect of the contextual explanations. Participants' reactions revealed several recurring design tensions that cut across warning types. They valued context-specific guidance that clarified what action to take and why, yet too much explanatory text risked overwhelming readers. While confident, cautionary framing encouraged warning adherence, AI disclaimers sometimes undermined trust by introducing uncertainty. Replicating prior results, we observed the importance of making it easy to take action (e.g., the one-tap "Report and Block" option) and found that strong visual cues, like the color red and hazard icons, enhanced salience and urgency.

Warning explanations

Participants' reactions to warning text reflected a central design tradeoff between direction and evidence: they wanted clear, actionable guidance on what to do, but differed in how much explanation they needed to justify that advice. Most appreciated the contextual warnings for offering concise, situation-specific direction (17 participants coded as saying *likes level of direction* in response to one of the contextual warnings). For example, P4 reacted to the *short/uncertain contextual warning* by saying, "I like that it gives a warning, and that it says to not click links and to contact your bank directly. That is very good advice." This type of feedback was sometimes given with the proviso that the speaker did not need the advice, but it would be *helpful for others* (9 occurrences). P1 commented on the *short/uncertain contextual warning* for the Fake Amazon Account Alert: "This one does tell you not to click the link and to verify via the official app or website, which I think is a good reminder for people who might not be familiar with... text... like an older adult... I'm thinking about my dad."

The *long/confident contextual warning* provided detailed evidence of fraud alongside directions. Participants' reactions to this explanatory content were mixed, with nine participants stating that they *like the level of evidence* on the *long/confident contextual warning*. P15 saw the warning last, on the Fake Bank of America Account Alert, and felt that it was the best warning, explaining, "It talks about the... specific URL that's not affiliated with a Bank of America. It talks about the grammatical error I saw, and it also specifically says not to click on the link and to contact the bank through official channels... this warning message has all the bases covered." An additional five participants stated that they liked the level of evidence presented in the false positive warning; however, this evidence was often misleading (e.g., indicating that an Amazon.com URL was suspicious).

Others felt that the longer text risked overwhelming users, preferring briefer explanations that conveyed the main point without extra reading. Seven participants indicated that they *dislike the level of detail* on the *long/confident contextual warning* or a false positive warning. For example, when asked for their thoughts on the warning, P19 stated, "It's a lot of words... I don't like phone reading in general.... I feel like that could just be in the 'Learn more' section. All I need to know is the recommended action." Other participants expressed dissatisfaction with the amount of text

in the long warning, but were less clear about what should be excluded. P7 stated that they liked the text for the Employment Scam, but felt like it may be too long: "It's a lot of text, but it is informative. My experience is that people often don't read messages that are too wordy."

For some who disliked the longer text, the *short/uncertain contextual warning* was more effective. However, by prioritizing brevity, the short warnings sometimes introduced confusion. P14 noted that the short contextual warning's advice not to click links seemed inconsistent with the presence of a "learn more" link: "It says, 'Don't click links.'... [but] where there's learn more that's clicking a link." This sort of confusion was also experienced by P10 and P7. The longer warnings more clearly referenced the link within the text message, which prevented this sort of issue.

The generic scam warning, which omitted explanatory text, similarly illustrated a tradeoff between clarity and brevity. Some participants appreciated its simplicity (5 occurrences of *likes level of detail*). P19 liked this warning best, explaining, "Fewer words is better." P20 initially expressed a preference for more text, but changed her mind upon seeing the warning with no text: "This is actually... a very effective message... what they're trying to convey to you is this is likely a scam. So pay attention..." More common was a desire for some explanation to be added (*dislikes level of evidence* or *dislikes level of direction*). For example, P8 reacted negatively to the warning, saying, "I mean, it's pretty unhelpful, like this is like saying, this is likely a scam, but it doesn't tell me why. So I'm like... it could be, but... how do you know that?" She went on to also recommend that the warning include "like a suggested action of 'Don't click the link' would be good."

Finally, opinions on the explanation from the *spam warning* were divided. Some participants felt the general explanation ("Similar messages you received were identified as spam") was a sufficient level of evidence to justify the warning (7 occurrences of *likes level of evidence*), while others desired a more specific explanation of why a message looked like spam (4 occurrences of *dislikes level of evidence*). Some participants also felt the warning lacked clear instructions for the user (3 occurrences of *dislikes level of direction*). Interestingly, several participants commented on the second sentence of the warning: "Google keeps your personal information private, safe, and secure." This boosted confidence in some, like P17, who directly contrasted it with the contextual warning AI disclaimer: "I think it's nice that it says that... Google keeps your personal information private, safe, and secure.... It suggests to me that I'm keeping some form of privacy... in a way that I think marking it as AI in the other spam message didn't." Others viewed the message as useless or deceptive, with P18 calling it "false" and P20 referring to it as "a little ad for Google." These tensions highlight how even brief, auxiliary text can influence users' perceptions of a warning.

Warning framing and confidence

Participants emphasized that the tone and confidence of a warning were also important in shaping their trust and intended response. Participants most often commented on warning framing and confidence in three contexts: the *save contact notice*, the *spam warning*, and the *short/uncertain contextual warning*. The *save contact notice* was criticized by several participants for lacking an explicit warning about the associated scam message (6 occurrences of *dislikes tone of warning*). P12 felt that the notice may encourage some people to "trust this number." P7 similarly felt the

warning was "trying to get you to add [the number] rather than reporting it as spam." A few participants also argued that the Google Messages *spam warning* suffered from a similar issue (4 instances of *dislikes tone of warning*). In recommending that the headline be changed, P16 explained "This is more like, 'Why, this could be a spam'.... It sounds a little more doubtful, so people might not pay that much attention to it." These reactions indicate that when warnings convey uncertainty or a weak cautionary tone, users may either ignore the message or interpret it as implicitly endorsing the sender.

The presence of an AI disclaimer in the *short/uncertain contextual warning* sparked polarized reactions, revealing a tension between transparency and authority. Half of the participants felt it was unnecessary or even undermined the warning (10 occurrences of *warning shouldn't include AI disclaimer*). P8 explained that seeing the note about AI generation made them feel they needed to double-check the message, "I don't think it's useful to be told 'this was generated by AI. It may not be a hundred percent accurate' because I mean it's already saying 'this may be a scam'... It's adding more layers of maybes." Only a few participants thought the AI disclaimer was necessary (3 occurrences of *warning should include ai disclaimer*). For P18, who discussed their negative experiences with email spam warnings, the AI disclaimer increased her trust in the warning by directly acknowledging potential inaccuracy: "I might click the report and block on this one just because it says... 'this was generated by AI and may not be 100% accurate,' which is my experience." For P6, the text was useful since it alerted him to a *privacy concern*. He shared, "They tell you the AI read the text message... That always bothers me... that somebody or something is reading my text messages. I think that's a privacy problem." These divergent responses underscore the challenge of communicating AI uncertainty in the warning context: while disclaimers may increase trust for some users by acknowledging potential inaccuracy, for others, they introduce additional ambiguity that can reduce the perceived authority of the warning.

Warning options

The options within the warnings were less often discussed than the textual features, but participants generally seemed to value ease of access to defensive actions. Replicating the findings of Tabassum et al. [470], several users commented negatively on the "Not spam" option present on the Google Messages spam warning (6 occurrences of *dislikes options*). These participants generally wanted the ability to confirm that a message was spam. In a typical example, P2 explained, "I wish it had a button that lets me immediately say that it is spam...It makes it a little harder to actually report this as spam and easier just to report it as not spam." In contrast, seven participants commented positively on the options available on the *save contact notice* or the four hypothetical warning conditions. When P15 was asked to provide feedback on the *generic scam warning*, he referred positively to the options, "there's a... report and block button right there to click on... easy peasy." In general, this suggests that users value the ability to take defensive actions from the message screen.

Visual design

Participants occasionally commented on the visual elements of the warnings, indicating the use of red color and the stop sign icon as influential. Some participants indicated they liked the red

icon and color in the four hypothetical warning designs (5 participants coded as saying *likes iconography* or *likes color* for the hypothetical warnings). In discussing the *contextual warning (short, uncertain framing)*, P5 said, "It stands out because it's red and it's like the first thing you see when you open up the message." The red octagon icon, which appeared in the inbox view for all but the *save contact notice*, reportedly influenced some users' perceptions before they opened the message. When discussing the IRS Arrest Threat, P9 explained, "Even before you open it, there are exclamation marks next to it. That sort of biases your viewpoint, unless you know the number." More decisively, in discussing the icon associated with the warning on the Fake Bank of America Alert, P8 stated, "I probably just wouldn't have opened this in the first place to be honest."

In contrast, the blue or purple coloring of the Google Messages notices was generally viewed negatively by participants who commented on it (5 participants coded as saying *dislikes color* or *dislikes iconography* for *save contact notice* or *spam warning*). Discussing the *spam warning* on the Fake Bank of America Account Alert, P1 stated, "It's also not in red. So it's not as like glaring to people... if someone's not really paying attention... they might just click that link." Similarly, P5 indicated that the blue color of the warning failed to capture her attention, saying, "the warning message... doesn't really... stand out... so I naturally gravitated towards the actual content of the message below." In sum, participants viewed the red visual elements as an effective cue that reinforced the warning's message and, in some cases, supported snap judgments about suspicious messages without the need for careful reading.

8.5 Discussion

In this section, we discuss our results, focusing on our novel findings regarding warnings that include context-specific explanations.

Our findings affirmed several patterns in how participants reasoned about unsolicited SMS messages. As a group, participants responded decisively to clearly scam-like messages (often recommending that they be reported or blocked), but adopted more conditional and uncertain strategies for legitimate messages. Participants commonly reasoned based on message features, their personal experience, and other expectations about what "normally" occurs. These results confirm prior work on SMS message processing [470, 481]. Our findings extend the existing literature by demonstrating how warnings actively support users' reasoning. They sometimes serve as a decisive signal that short-circuits further evaluation. In other cases, they acted as one factor among many or as a source of reassurance that confirmed existing suspicions. Warnings also highlighted particular actions or pieces of evidence, shaping both perception and behavior.

Our exploration of design features revealed some patterns that align with prior warning research. Participants favored visually salient designs that used strong color contrast and clear iconography to convey danger, particularly red elements that stood out against the blue and purple hues of the Google Messages interface. While the preference for red may be culturally specific [251], prior work supports the value of contrasting colors and distinctive iconography for capturing users' attention in warnings [248, 563]. Participants also preferred warnings that offered immediate defensive actions, such as "Report and Block," over options that appeared to

downplay risk (e.g., "Not spam"), mirroring the results of Tabassum et al. [470] and reflecting best practices [34].

Beyond visual and action-oriented features, participants also responded positively to warnings that included context-specific text. They often described these as clearer or more informative, with a particular emphasis on the ability to provide specific guidance and contextual evidence.

The designs we tested revealed a key tradeoff between **detail and readability**. Longer explanations conveyed clearer evidence but were visually dense, whereas shorter explanations were easier to scan but sometimes introduced confusion or contradictory cues (e.g., "don't click links" next to a "learn more" link). This tension may be especially consequential in mobile contexts, where users often interact under time pressure or partial attention. Although participants in our study were highly attentive by design, their comments about the dense text suggest that it could exceed attention limits in everyday use.

Improved formatting (e.g., increased white space or structured bullet points) could help mitigate this problem and preserve the benefits of contextual detail [548]. This may be challenging to control due to the stochastic nature of GAI output. Alternatively, contextual warnings could also be made configurable, allowing users to select a warning format based on their preferences. Designers could also offer expandable explanations, such as through "accordion" interfaces that reveal longer reasoning on demand. Prior work on notice design suggests that few users voluntarily access secondary text [7, 55]. Still, such layered designs could preserve attention and readability while allowing motivated users to access the richer, context-specific evidence that supports comprehension.

A second tension arose around **confidence and transparency**. Participants generally preferred warnings that appeared certain and directive. The AI disclaimer, especially the phrase noting that the warning "may not be 100% accurate," undermined perceived danger for many. At the same time, a few participants viewed the disclaimer positively. For them, acknowledging AI involvement increased transparency and trust, either by aligning their experiences with those of spam detection systems or by highlighting potential privacy implications. Balancing these perspectives is difficult: transparency is ethically important, but excessive qualification can reduce authority. A more neutral phrasing, such as "This warning was generated by AI," may preserve transparency without signaling excessive doubt.

These dynamics were also reflected in participants' responses to false-positive warnings. Even when a legitimate message was incorrectly flagged, most participants accepted the warning's authority and recommended reporting or blocking the sender. This behavior could reflect trust in the warnings developed over the course of the study, leading participants to reflexively follow them. This result could also be confounded by the study's design. For example, participants may have complied with false-positive warnings presented later in the study to reduce cognitive burden, rather than out of trust in the system. In many cases, deference to the system may be fail-safe: it is arguably better for users to heed an occasional false alarm than to overlook a genuine threat. However, this pattern also underscores the need for calibrated trust. Users should view warnings as informative signals rather than unquestionable truths, attending to the evidence they provide while recognizing that even contextual, AI-supported explanations can be wrong or even intentionally manipulated through prompt injection. These issues are common to all GAI systems [335], but are especially challenging to address in the context of warnings.

These results are most directly applicable to the Android instant messaging context; however, some may generalize more broadly. For example, while iMessage (the default SMS app for iOS) does not presently include scam or spam warnings, it does include a notice on all messages from unknown senders, which highlights the option to "Report Spam" [22]. This UI element may help highlight defensive actions in the same way as the *save contact notice* on Android, though the iOS implementation is more subtle, appearing as text embedded in the background rather than a distinct pop-up box.

Our results also must be interpreted within the limits of the study context. Participants viewed messages in a controlled, task-focused setting where attention to potential deception was unusually close. In everyday use, where users are multitasking and cognitively loaded, warnings may be overlooked or interpreted differently, meaning the reasoning patterns observed here likely represent a best-case scenario for user engagement with warning text. Future work should test context-specific warnings under more ecologically valid conditions. While real-world or longitudinal studies would provide the strongest evidence of how users respond to warnings over time, controlled laboratory settings can also approximate real-world constraints by introducing time pressure or divided-attention tasks. For example, Sherman et al. enforced a time limit in their evaluation of spam call indicators [439].

Chapter 9

A Real-World Evaluation of GAI-generated, Context-Specific Warnings

This chapter presents the next stage in evaluating the GAI warnings explored qualitatively in the previous chapter. We first briefly discuss three online experiments used to select a single context-specific warning design for evaluation. The resulting *long* warning features a bulleted list describing evidence of fraud and providing the user with concrete guidance.

We then developed an SMS application with a GAI warning system based on *SmishX* [534] to conduct a real-world evaluation comparing the *long* warning to a *generic* alternative. While prior work conducted in lab environments, including Chapter 8, suggests the value of contextual warnings [90, 143, 295, 389, 412, 534], no prior study has evaluated their effect in the real world. Specifically, we sought to evaluate the following hypotheses:

$\mathcal{H}_1$: Participants shown context-specific warnings will be less likely to click a link or call a number in a scam message than those shown generic warnings.

$\mathcal{H}_2$: Participants shown context-specific warnings will be more likely to report a scam message than those shown generic warnings.

$\mathcal{H}_3$: Participants shown context-specific warnings will perform better on a post-study phishing identification task than those shown generic warnings.

$\mathcal{H}_4$: Participants shown context-specific warnings will report a more positive perception of the warning system than those shown generic warnings.

If a warning design is better, it should be more likely to prevent users from interacting with messages ($\mathcal{H}_1$) and encourage them to take an active defensive action ($\mathcal{H}_2$). $\mathcal{H}_3$ follows from Roy et al.'s evaluation of explainable browser phishing warnings, which suggests that contextual warnings may improve users' ability to spot subsequent phishing even in the absence of the warnings [412]. Finally, $\mathcal{H}_4$ seeks to confirm that the positive sentiment observed in Chapter 8 remains when the warnings are viewed in realistic scenarios.

To evaluate these hypotheses, we recruited 91 participants who use Android smartphones to test our SMS application for two weeks. Using a mixed factor design, participants saw warnings for half of the study period, with the other half of the time serving as a control period to measure

This chapter is based on an in-progress collaboration with Ryan Li, Enze Liu, and Lorrie Faith Cranor.

their baseline behavior. Participants were randomized to see either the *generic* or *long* warning design. Over the two-week study period, we sent participants eight simulated phishing messages that impersonated various businesses and employed different URL-impersonation techniques. We also classified genuine messages that participants received from unknown senders. To test $\mathcal{H}_1$ and $\mathcal{H}_2$, we measured participants' interactions with these eight messages; to test $\mathcal{H}_4$, we surveyed them about their experiences at the midpoint and conclusion of the study. $\mathcal{H}_3$ was also evaluated in the conclusion survey through a phishing identification task.

In our experiment, we found that both warning designs were associated with reduced click-through on simulated scam messages compared with no warnings. This effect was significant for the *long* design, but did not reach significance for the *generic* design. While this is consistent with *long* being better than the *generic* design, there was no significant difference between the two conditions (failing to support $\mathcal{H}_1$). Participants seeing each warning design did not vary significantly in their odds of reporting a message (failing to support $\mathcal{H}_2$). Performance in the post-study phishing identification task did not vary significantly based on condition (failing to support $\mathcal{H}_3$). $\mathcal{H}_4$ was partially supported, with participants exposed to the *long* warning significantly more likely to rate the warning as more helpful to themselves or others. Further, qualitative evidence identified similar themes as Chapter 8, with participants highlighting the evidentiary value of the warning text and the ease of reporting with a one-click "Report & Block" button.

Ultimately, this chapter makes a significant empirical and methodological contribution by offering the first field experiment on context-specific, GAI-based warnings. These results are useful both for researchers seeking to conduct future SMS simulation studies and for those implementing GAI-based scam-explanation systems. While we were unable to conclude that GAI warnings are better at reducing harmful behavior than a *generic* alternative, trends in behavioral analysis and participants' positive sentiment suggest that this approach remains promising.

9.1 Pre-studies

We conducted three online experiments centered on SMS message review tasks to inform the design of the prototype warning system to be tested in the field study. For efficiency, we wanted to select a single GAI warning design to test against a single baseline warning without GAI text. We conducted the pre-studies sequentially, with each building upon the previous. We first describe the common features across all studies, then the specifics of each.

9.1.1 Overview

We used Qualtrics to administer the studies, and Prolific to recruit participants. We used quota sampling to ensure that the samples were gender-balanced and age-balanced (across ages 18 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, and 65+). Participants were required to be fluent in English, be located in the United States, and access the survey on a desktop or laptop device. Location was assessed using IP geolocation, and participants determined not to be in the United States were automatically excluded. All surveys concluded with demographic questions to contextualize participants' responses.

For all three studies, we took several steps to prevent low-quality responses. Participants were required to have completed at least 30 prior tasks on Prolific and have at least a 90% approval rating. We instructed participants not to use LLMs in completing the survey. To ensure compliance, Prolific's "Authenticity Check" [385] feature was enabled. We also used JavaScript event listeners to record whether participants copied or pasted text in the survey. Finally, all studies also included attention check questions, requiring them to identify the topic of a message shown in the simulated interface. Participants with low authenticity scores, who failed attention checks, or who repeatedly copied or pasted without explanation were not compensated and had their data excluded.

9.1.2 Pre-study 1

Materials for this study are available in Appendix G.2. The first study was a between-subjects experiment intended to test the effect of different warning designs on participants' purported behavior. Participants were instructed that the study aimed to evaluate how users interact with text messages in a novel texting interface. All participants viewed six messages in an interactive mockup of a text-message interface (see Appendix G.5 for example screenshots). Participants were asked to select the action they would take if they received the message (e.g., respond, report the message, click a link in the message) and to justify their response. Regardless of condition, the first two messages were a scam message (a USPS delivery scam) and a legitimate message (a healthcare reminder text), presented with no warnings. The remaining four messages were presented in a random order, with warnings displayed on scam messages based on the condition. Stimuli were a subset of those used in Chapter 8's interview study, with modifications to the healthcare message to make it non-specific to a particular location.

Participants were randomized to one of six conditions: a control condition with no warnings or one of five conditions with different warning designs:

- **Generic Warning**: A warning with the headline "This is likely a scam" and the option to "Report & Block" or "Learn More." The design features no explanatory text.
- **Spam Warning**: Based on the Google Messages Spam warning. The headline states"Why this looks like spam," and provides the option to "Learn more" or mark the message as "Not spam." The design features generic explanatory text about similar messages being classified as spam.
- **Long Warning**: A warning with the headline "This is likely a scam" and the option to "Report & Block" or "Learn More." The design features a bulleted list that explains the classification decision and what the user should do.
- **Accordion Warning**: A warning with the headline "This is likely a scam" and the option to "Report & Block" or "Learn More." The design features a short statement explaining the classification and what the user should do. Pressing "Learn More" reveals the bulleted list from the *long warning*.
- **Blocking warning**: A warning that is identical to the *long warning*, except that the design presents a box in the center of the screen, preventing the user from seeing the text message.

The "Learn more" option is replaced by a "Proceed anyway" option, which causes the warning to disappear.

All warning conditions featured a light red warning box with a red warning triangle icon in the top-left corner to indicate a hazard. The explanatory text within the *long*, *accordion*, and *blocking* warnings was generated with GAI, using the system described in Section 9.2.2. The design of the text content was informed by the prototypes tested in Chapter 8, with modifications. Specifically, we prompted the GAI system to use bulleted text for longer explanations, we tested the option to view the longer warning text via an accordion interface within the shorter warning, and used a "confident" headline across all warning designs (i.e., one saying "This is likely a scam" rather than "This may be a scam"). The *blocking* design was inspired by findings from browser security research that suggest such designs are more effective at encouraging compliance [7, 397].

After the message review task, participants in the warning conditions answered a set of nine Likert-scale questions assessing various subjective aspects of the warning design, their trust in AI to analyze text messages accurately, and their willingness to have AI analyze text messages. For reference, participants were shown a previously unseen scam message with a warning during this task. Participants were also asked whether they would make any changes to the design they viewed and (if so) to explain what changes they would make. Finally, they were asked to view all five warning designs and select their favorite.

Actions in response to scams were coded into two binary outcome variables: *interact* (clicking or replying to a message) and *dismiss* (reporting, deleting, or exiting a message). All other actions were coded as false for both variables. This yielded three observations per participant for each variable. For each outcome variable, we modeled the data using a mixed-effects logistic regression with participant as a random effect. We compared three nested models via likelihood ratio tests (LRT), with each successive model adding one fixed effect. For each outcome, the first model was a base model that included only the random intercept for the participant and a fixed effect for message identity. The second model added a binary fixed effect for the presence of a warning. The third model replaced the binary warning indicator with a six-level categorical fixed effect representing which specific warning condition was shown. If warnings are effective, we expect the second model to provide a significantly better fit than the base model. If the specific warning condition matters, we expect the third model to provide a significantly better fit than the second. This structure was chosen to first validate that warnings had a significant effect in the survey context and then to determine whether the specific condition mattered.

We analyzed Likert-scale sentiment by translating each scale point to its corresponding integer value (1 to 5) and using the Kruskal-Wallis test to determine whether ratings tended to be higher in one or more conditions.[1] The Kruskal-Wallis Test is a non-parametric alternative to the Analysis of Variance (ANOVA) that does not require the data to be normally distributed and is thus suitable for ordinal data [269].

Detailed results of the statistical analysis are in Appendix G.2.3. Average completion time was 13 minutes and 19 seconds, and the final sample size was 353 participants. For *dismiss* actions, model 2, considering the presence of warnings, achieved the best fit, with the LRT showing no significant improvement by model 3. Controlling for message, the presence of the

[1]Scale points were labeled with their integer values when participants took the survey to help participants view points as equidistant.

warning was associated with 3.36 times greater odds of a *dismiss action* ($\beta = 1.212$, $p < 0.001$, $S.E. = 0.354$). For *interact* actions, models failed to converge due to quasi-complete separation. The second scam message (an employment scam) induced only a single *interact* action. This result suggests that the second message was clearly fraudulent for the majority of participants, leaving little opportunity to observe the effect of warnings. Given this and other limitations of the study design, such as the lack of complete randomization of message stimuli, we conducted two additional studies.

9.1.3 Pre-study 2

Materials for this study are available in Appendix G.3. Pre-study 2 was conducted to validate a broader range of realistic scam stimuli. Before the study, we designed 10 service-impersonation messages by adapting genuine messages from various businesses (e.g., Amazon, Discover Bank, ID.ME, T-Mobile, etc.). The fraudulent versions used URL-impersonation techniques documented in prior research [274, 401]. We purchased the domain names referenced in the messages and configured them to redirect to an error page. The stimuli can be found in Appendix G.3.2.

In contrast to the other studies, participants were explicitly recruited for a study on "how people evaluate text messages." Participants saw a set of 10 messages (five legitimate, five fraudulent) using the same mockup interface as in Pre-study 1, but without warnings. Stimuli were randomized so that each participant saw either the fraudulent or the legitimate version of each message. For each message, participants classified it as fraudulent or legitimate on a five-point scale (from "Definitely fraudulent" to "Definitely legitimate") and justified their answer. After reviewing all messages, participants answered demographic questions and indicated which (if any) of the ten services they use.

Ratings were coded into a binary outcome variable indicating whether a fraudulent message was correctly identified as such: responses of "Definitely fraudulent" or "Probably fraudulent" were coded as correct (1), and all other responses as incorrect (0). This yielded five observations per participant, one for each fraudulent stimulus seen. We compared two mixed-effects logistic regression models using a likelihood ratio test: a null model with only a random intercept for participant, and a model that adds a fixed effect for message identity.

Detailed results of this analysis are in Appendix G.3.3. We recruited 200 participants who completed the survey in an average of 15 minutes and 5 seconds. The message model provided a significantly better fit than the null model, indicating that the correct identification of scams varied significantly by message content. In general, no message was as obviously fraudulent as the employment scam tested in Pre-study 1. A message impersonating T-Mobile that requested payment for an overdue bill was the easiest to identify as fraudulent, with 74% of participants rating it as fraudulent. A message requesting confirmation of a dinner reservation and a message urging the recipient to register for a vaccine were the hardest to identify as fraudulent, correctly recognized by only 12.0% and 37.5% of participants, respectively.

9.1.4 Pre-study 3

Materials for this study are available in Appendix G.4. Pre-study 3 followed the same procedure as Pre-study 1 but used stimuli validated as not obviously fraudulent in Pre-study 2, employed

complete randomization of message order, and included one legitimate message paired with a false-positive warning. Participants viewed 10 messages in random order (five scam, five legitimate) and indicated what action they would take in response to each. As in Pre-study 2, participants saw either the legitimate or fraudulent version of each message topic. The first legitimate and first scam messages shown to each participant were always presented without a warning. For participants in a warning condition, the remaining scam messages were displayed with the corresponding warning, and one randomly selected legitimate message was presented with a false-positive warning.

We also excluded two of the warning conditions evaluated in Pre-study 1: the *blocking* and *spam* warnings. The *spam* warning was the least likely to be selected as participants' favorite design (26 participants in warning conditions, 9.0%). The *blocking* warning was also rarely selected as a favorite (43 participants in warning conditions, 14.9%) and was rated as more annoying than the other designs. This left the *long* and *accordion* designs as the primary conditions of interest. We also retained the *generic* warning, which was rated poorly (selected as a favorite by 29 participants in the warning conditions, or 10.1%). It was retained as a baseline condition without GAI-generated explanatory text.

We performed sentiment and action analysis in the same manner as Pre-study 1. With the inclusion of false-positive warnings, we fit an additional set of nested mixed-effects logistic regression models to the legitimate-message stimuli.

Detailed results of this analysis are in Appendix G.4.3. Our final sample included 350 participants who took an average of 16 minutes and 59 seconds to complete the survey. For *dismiss* actions on scam messages, model 3 (warning condition) provided the best fit, significantly improving on model 2. While all warning designs significantly increased the odds of dismissal relative to the control, post-hoc, pairwise comparisons between the individual warning conditions (Tukey-adjusted) showed that the *long* design was associated with significantly greater odds of a *dismiss* action than both the *accordion* design ($\beta = 0.6524$; $S.E. = 0.236$; $z = 2.763$; $p = 0.016$) and the *generic* design ($\beta = 0.5865$; $S.E. = 0.239$; $z = 2.455$; $p = 0.038$). *generic* and *accordion* did not significantly differ ($z = 0.285$; $p = 0.956$; $S.E. = 0.231$).

For *interact* actions on scam messages, model 2 provided the best fit. This suggests that all warning designs were equivalent in reducing *interact* actions. Controlling for message identity, the presence of a warning was associated with substantially lower odds of interaction ($\beta = -1.9716$; $z = -8.103$; $p < 0.001$; $S.E. = 0.2433$).

For legitimate messages, model 2 (warning presence) also provided the best fit for both outcomes. Warning presence was associated with greater odds of *dismiss* actions ($\beta = 1.0761$; $z = 6.735$; $p < 0.001$; $S.E. = 0.1598$) and lower odds of *interact* actions ($\beta = -1.7436$; $z = -5.217$; $p < 0.001$; $S.E. = 0.3342$). This suggests that all three warning designs suppressed interaction and increased dismissal at approximately equal rates for false-positive messages.

Based on these results, we selected the *long* and *generic* warning designs for further testing in a real-world field study. The *long* warning is the GAI-based warning with the highest dismissal rate, while the *generic* warning serves as a baseline without GAI text. We discuss the implementation of these designs into a full prototype in the following two subsections, first describing the SMS client and then describing the server that performed AI classification and collected study metrics.

9.2 Prototype implementation

In this section, we describe the implementation of a prototype warning system. The design of this system was made to support the 2×2, mixed factor experimental design described in Section 9.3.1. The system consists of two components: an Android SMS application installed on participants' phones and a remote server backend. The prototype SMS application interacts with the user and collects telemetry. The server performs message classification using a modified version of *SmishX* [534], sends the results to the Android application, sends simulated phishing messages to be injected into the frontend, and receives and uploads metric information to a Postgres SQL database.

9.2.1 Android Frontend

We forked QUIK,[2] an open source SMS application, to provide a fully functional base for our prototype. We largely preserved QUIK's basic functionality. We made four main changes to QUIK: adding the ability to duplicate RCS conversations, implementing two scam-warning designs, instrumenting the app to collect behavioral metrics, and enabling the injection of simulated messages.

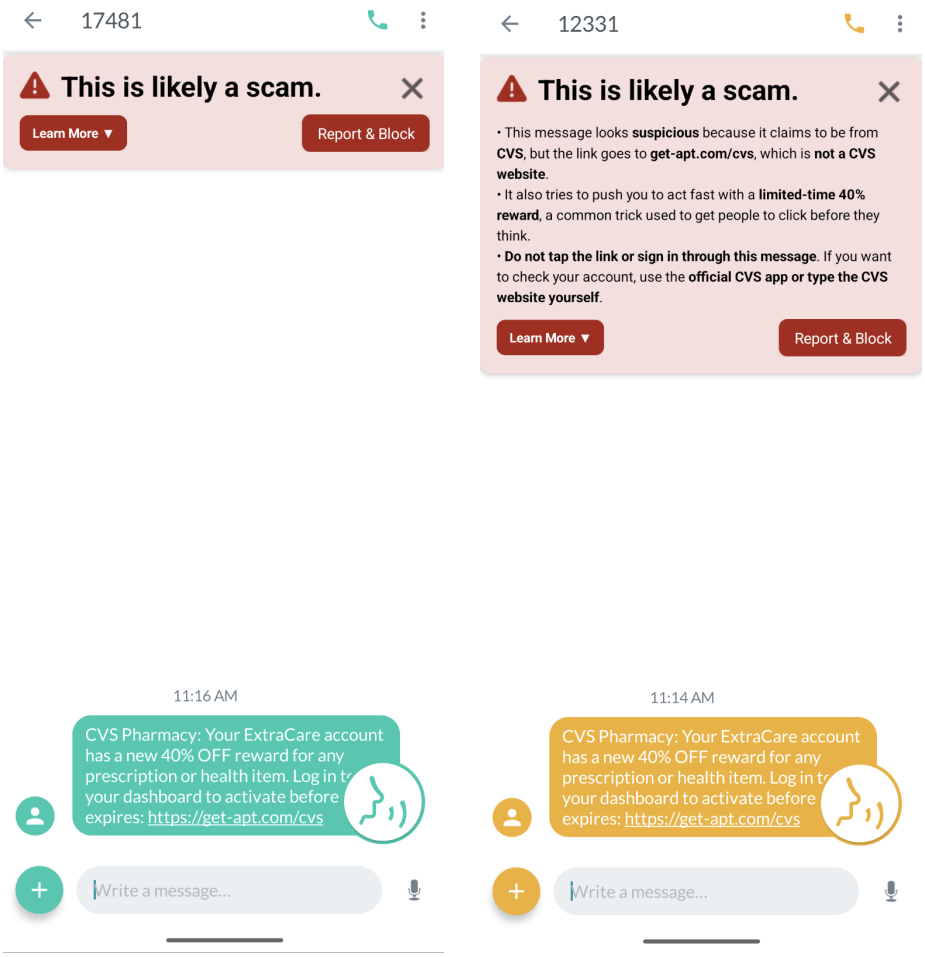
We begin with the RCS limitation, which motivated the first of these changes. QUIK does not support RCS features. RCS (Rich Communication Services) is a new text messaging protocol that enables various advanced features, including typing-indicators, emoji reactions, high-quality media sharing, end-to-end encryption, and advanced group chats [208]. While RCS is an open protocol, there is currently no support for RCS on Android outside of the closed-source Google Messages app. To aid participants in transitioning to QUIK, we implemented the ability to duplicate RCS chats. We also updated the about page in the app to include the researchers' contact information.

The primary addition was the implementation of scam warnings. Based on the pre-study results, we implemented two designs (see Figure 9.1): *long* and *generic*. Messages received from unknown senders (i.e., those not in the user's contact book) were sent to our server for classification, as described in the next subsection. We chose to classify messages from unknown senders under the assumption that most scams would come from senders not in a participant's contact list, and to minimize the collection of personal text messages. While a message is being classified, it is held for up to 180 seconds. Once the classification was returned from the server or 180 seconds had elapsed, the message was released, appearing to the user and creating a push notification. If warnings were not active for a participant during a given time period, the message was still classified but not withheld from display.

The app was instrumented to collect the following metrics:

- **App usage**: cumulative number of times the app is opened; cumulative amount of time the app has remained opened; and cumulative number of conversations of any kind that are opened.
- **Warning interactions**: "Learn more" button clicks; warning close or dismiss button clicks (x in the top right corner); and "Report & Block' button clicks.

[2] `https://github.com/quik-sms/quik`



(a) The *generic* warning design

(b) The *long* warning design

Figure 9.1: Screenshots of the QUIK app with our scam warning design displayed on the CVS scam message.

- **Real unknown-sender and simulated message interactions**: opening of messages; clicks on links; deletion of messages or threads; and replies to messages.

App usage metrics are collected solely to confirm participant activity during the study period. Simulated message and warning interactions are the primary focus of the experiment, allowing us to understand how participants respond to warnings in our simulated messages. Real unknown sender message outcomes were collected to support future analysis of how participants respond to messages received naturally; the present chapter analyzes only behavior in relation to the simulated stimuli.

The app was designed to collect and store data on the Android device, then periodically upload it to the server. The metrics are uploaded to the server upon first opening the application, on a daily schedule, and once at the end of the 14-day study. Metrics were also uploaded whenever the app sent a message for classification.

The final addition to the application was the ability to "inject" simulated messages. Such injected messages appear in the app as normal messages, but are not actually sent over the cell network. This was done in order to ensure delivery of messages, in light of Rahman et al.'s observation that many messages in their simulation were not delivered [387].

9.2.2 Server Backend

The server acts as a backend for the Android application, classifying messages, sending simulated messages, and recording metrics. We focus here primarily on classification.

Our warnings were generated using SmishX, an LLM-based SMS classification system originally developed by Wang et al. [534]. The system takes an SMS message as input and extracts any URLs it contains. It takes a screenshot of the destination and also converts the page's HTML to markdown using Jina AI Reader.[3]. Both the screenshot and the page content are summarized by LLM calls. It extracts brand names from the text and uses the Google search API to retrieve the top five URLs associated with the name(s). Finally, it also conducts a WHOIS lookup for each URL. The resulting inputs are fed into an LLM with a detailed, chain-of-thought prompt instructing it to classify messages as phishing, spam, or benign and output a JSON classification report. This classification report is turned into user-friendly output through a final summarization prompt.

We made three main changes to SmishX. First, we modified the classification prompt by adding a statement about phone numbers in text messages ("Please be aware that both phishing and legitimate messages may include phone numbers. Legitimate messages will include numbers officially known to be associated with a business...').[4] This was done to account for scams that request call-backs at illegitimate numbers. We also changed the classification prompt to determine if the message was considered phishing. The original prompt specified whether the message was phishing or spam; however, we wanted to focus only on generating warnings for clear phishing or other scams.

Second, we modified the final summarization prompt to output "...two versions of a simple and easy-to-understand explanation to appear within a warning,..." if the message was deter-

[3] `https://jina.ai/reader/`

[4] The deployed prompt included minor typographical errors that are corrected here. See section G.1.7

mined to be a scam. Specifically, the prompt specified that the first version should be "...a single, short sentence that gives clear, context-specific direction on what the user should do with the message..." and the second should be "...no more than three bullet points, focusing on whether the message seems suspicious or safe..." and providing "...a simple reason to support your conclusion, including clear evidence such as a suspicious website link or an urgent tone in the message..." The specific components of the explanations (evidence and specific direction) were chosen based on the elements that participants in the interview study in Chapter 8 most appreciated. The first short explanation was used as the initially visible text in the *accordion* warning tested during the pre-studies. The second, longer explanation formed the text of the *long* warning, which we ultimately tested in the user study.

Both modified prompts are available in Appendix G.1.7.

The third and final modification we made to SmishX was to change from using GPT 4 to GPT 5.4.[5] This upgrade was made to take advantage of the newer model's generally improved performance [152]. We did not re-test our modified version of SmishX on its original evaluation dataset, and acknowledge that these modifications may have altered its classification performance.

Upon receiving a message for classification, the message content, the sender's area code (or the whole number in the case of short codes), and the hash of the sender's phone number were stored in a Postgres SQL database. If the message has not been classified before, the server runs the message contents through SmishX. If our SmishX implementation does not finish within the server's timeout period, the server signals the app to release the held message. If the classification completes within the timeout period, and warnings are enabled for the client during that week, the classification decision is sent to the client and displayed with the message. For participants in the *long* cohorts, the generated explanation is sent along with the classification.

To support the experimental design described in Section 9.3.1, eight simulated scam messages were defined on the server, each with associated *long* warning text. This warning text was pre-generated using SmishX to ensure consistency in stimuli. Upon installing the app, a random schedule was generated for each participant, scheduling the eight simulated messages to be sent at some time between 7:00 AM and 5:00 PM every one to three days, with four messages in the first week and four in the second. The timing was stored on participants' phones, and the phone would query the server for the message/warning content at the appropriate time. If the server was unavailable, the phone would repeatedly retry at 10-second intervals.

9.2.3 Runtime Configurations

We generated four versions of the client APK, one for each cohort (i.e., week one long warnings, week one generic warnings, week two long warnings, and week two generic warnings). These APKs were distributed using Qualtrics, with even randomization during the onboarding survey. Each APK was administered in accordance with the crossover design, with warnings either stopping or starting on the seventh day of participation.

[5]Specifically, the first pre-study used `gpt-4o-2024-08-06`. The third pre-study and main user study pinned GPT to `gpt-5_4-2026-03-05`.

9.3 User Study Methods

We conducted a two-week field experiment to evaluate the warning designs selected in the pre-studies. We recruited 91 Android users through Prolific and through physical and digital flyers distributed in Pittsburgh, PA. Participants installed our prototype SMS application as their default messaging client and used it throughout the study. Over the two weeks, we sent each participant eight simulated scam messages impersonating well-known services, four in each week, and recorded how they responded. Participants saw warnings on messages classified as scams during one week of the study and no warnings during the other, with the order randomly assigned; those who saw warnings were shown either the *long* or the *generic* design. Participants also completed four surveys: a screening survey, an onboarding survey, a midpoint survey after the first week, and a conclusion survey at the end of the study. The subsequent subsections describe the experimental design, survey instruments, recruitment, statistical analysis, ethical considerations, and the study's limitations.

9.3.1 Experimental Design

To evaluate our hypotheses, we employed a 2×2 crossover experimental design. The between-subjects factor was the warning design: participants were randomized to either see the *generic* warnings or the *long* warnings. The within-subject factor was the presence of warnings: half the time, participants saw scam warnings; the other half, they saw no warnings. The order was randomly assigned. In effect, this means participants were assigned to one of four cohorts: week 1 classification with *long* warnings; week 1 classification with *generic* warnings; week 2 classification with *long* warnings; and week 2 classification with *generic* warnings. This design was chosen to increase the power of the analysis for a given sample size [53].

Throughout the experiment, we sent participants eight simulated scam messages (four per week). All messages were service impersonation messages, derived by taking genuine service messages and inserting a fraudulent URL or phone number (see Section 9.1.3). We describe the stimuli here. Complete versions are available in Appendix G.1.8.

- **Amazon**: states that a package couldn't be delivered, providing a link for more information. The domain name utilizes a homoglyph substitution (`amązon.com`).
- **Bank of America**: states that a purchase has occurred at a store outside of the US. Recipients are encouraged to call a phone number with an area code in Michigan if they don't recognize the purchase. The message also includes an instruction to reply "STOP" to opt out. In the second and third pre-studies, this message impersonated Discover. We swapped the message to Bank of America, which is a much larger bank.
- **Chili's (OpenTable)**: states that a reservation has been made at Chili's on July 31st. The message instructs recipients to click a link to cancel. The domain name includes the name OpenTable (`confirm-opentable.com`), a prominent online reservation service. In the second and third pre-studies, this message referred to a small local restaurant. We adapted it to mention Chili's to be more broadly relevant.
- **CVS**: states that a 40% coupon is available in the recipient's ExtraCare account. Recipients are encouraged to log in at the link before the coupon expires. The link contains CVS in

the URL path (`get-apt.com/cvs`). In the pre-studies, this message was framed as a vaccine reminder. We adapted this to refer to a coupon, as some qualitative responses indicated vaccination may be an undesirable lure.

- **Docusign**: states that "Sarah Connolly" sent an invoice to be reviewed and signed. The URL contains a long string of subdomains beginning with `docusign.net`. The actual domain name is `etz5811-eiue348wi0li27dh8jtkku.xyz`
- **ID.me**: states that the recipient's identity is being used to log in to ID.me. It requests that the recipient click a link to upload a photo or report unauthorized access. The domain contains the words ID and me (`id-me.us`).
- **Indeed**: states that a recruiter from Apple has viewed the recipient's profile, and advises the recipient to click a link to log in now. The true domain is `apply-now-jobs.com`, but it begins with `indeed.com@`.
- **Venmo**: States that Jack Sims sent $44.62 for dinner. Instructs the recipient to click a link to log in and accept. The domain contains the name Venmo (`venmoapp.com`). The message in the pre-studies was framed as a request for money. We adapted this version to be a transfer to the recipient to be more appealing.

We chose to focus on service impersonation messages with minimal cues to deception to make the fraudulent stimuli difficult to spot, so that the effect of the warnings could be seen. We developed the URLs based on prior research on human URL parsing, focusing on strategies that are difficult to identify correctly [274, 401]. All messages appeared to be sent from unused 5-digit shortcode numbers.

Due to an error in stimulus preparation, the final sentence of the *long* warnings for the Bank of America and Chili's messages named the institution the warning text was originally generated for (Discover and the local restaurant, respectively) rather than the one shown. Both errors were confined to the *long* condition and to that one sentence.

9.3.2 Survey Instruments

Throughout the experiment, participants completed four surveys: screening, onboarding, midpoint, and conclusion.

Screening Survey

The screening survey verified that potential participants met the eligibility criteria and were willing to participate. We first asked participants two multiple-choice questions aligned with the study's stated eligibility criteria: which OS their phone used and how often they used the built-in texting app. Participants who indicated they did not use an Android phone or text at least once a day were screened out.

We then asked participants to rate their familiarity with 14 businesses. Four of these businesses were the subject of simulated scams sent during the experiment (Amazon, Bank of America, Indeed, and Venmo), two were fake businesses ("Nexis Pay" and "Second Pacific National Bank"), and the remaining businesses were irrelevant businesses in the same industry as the relevant businesses (e.g., eBay, Discover Bank, PayPal). To proceed to the full study, participants

had to indicate that they had used at least one of the four services that were impersonated by the simulated scam messages. This criterion was not told to the participants. Participants who reported using one of the fake businesses were screened out. We used this question to ensure that at least one of the simulated messages would be relevant to participants. We asked about only a subset of the impersonated businesses to reduce the chance that participants would recognize the relationship between the study and the scam messages. We selected the four businesses we asked about based on those we felt were most important for participants to use, so the message would feel congruent.

The screening survey closed with three yes/no questions that required participants to affirm their willingness to meet the participation requirements. Specifically, we asked if they were willing to install an app not on the Google Play Store, disable RCS, and have their text messages analyzed by AI. The question about RCS specified that this would disable read receipts, typing indicators, enhanced group chats, and end-to-end encryption. Participants were required to say yes to all three statements to proceed to the full study.

We compensated participants for screening by paying them $\$0.50$ (on Prolific) or entering them into a drawing for a one $\$50$ giftcard per 100 participants. We sent the consent form for the full study to participants who were determined to be eligible and then directed them to the onboarding survey. Participants not recruited on Prolific were asked for their phone numbers for this purpose. The full screening survey is available in Appendix G.1.3. The average length of the survey for participants was 6 minutes 20 seconds.

Onboarding Survey

The onboarding survey collected participant background/demographic information and instructed participants on installing the study app. We first asked participants from Prolific to provide their phone numbers so that we could identify their phones when they connected to our server. We then asked participants more detailed questions about their use of mobile phones, including the approximate age/manufacturer of their primary smartphone, their phone carrier, which instant messaging apps they used, and what they used instant messaging for. We then asked participants to describe their gender, age, race/ethnicity, household income, level of educational attainment, and whether they had education or experience in computing-related fields. The demographic questions all provided an option to decline to answer, and participants were assured that their responses would not affect their eligibility. We did not have specific hypotheses about the effect of phone usage or demographics, but rather collected this data to contextualize the sample.

The remainder of the survey consisted of installation and basic configuration instructions for the app. Participants were first provided with a QR code and a download link for the study app and then guided on installing it. Because the app was not on the Google Play Store, the process was complicated. Participants were required to change the default settings to allow external app installations and then dismiss several security warnings. After the app was installed, participants had to go into settings and specifically allow the app to use restricted permissions (see the Android documentation for more details [197]). Participants also had to disable RCS due to the lack of support for the protocol in QUIK. The instructions included screenshots of each warning participants would encounter, as well as a video showing how to enable restricted settings. Partic-

ipants were assured that the Android warnings did not indicate that the app contained malicious content.

We paid participants who completed the onboarding survey and were verified to have kept the app in use for at least 24 hours $15 via Prolific or a gift card. The text of the questions in the onboarding survey is available in Appendix G.1.4. The average time to complete the survey for participants in the study was 15 minutes and 58 seconds.

Midpoint Survey

Approximately 7 days after participants installed the app, we sent them a midpoint survey. This survey was primarily meant to capture participants' general perception of the app and identify any serious bugs that needed to be fixed. We did not call specific attention to the presence of warnings, but expected that participants who were exposed to warnings in the first week might spontaneously mention them in their feedback. We also expected that, in line with $\mathcal{H}_4$, participants' quantitative ratings of the app may be higher in the condition with the *long* warning. We refer to these sentiment questions as the *unprimed* sentiment questions.

Participants first completed the System Usability Score (SUS), a common measure of software usability [69]. We then asked participants whether they had experienced any "technical issues, bugs, or unexpected behavior" and whether there were any specific features they liked. Participants who said yes to either question were asked to explain their response. Finally, we asked participants to provide any additional feedback in a paragraph response box.

The survey closed with a question asking if participants wished to continue to the second week of the study. Participants who indicated that they did not wish to continue participating were debriefed (see Section 9.3.7) and given the opportunity to withdraw consent/request deletion of all collected data. They were also provided with instructions on uninstalling the study app, re-enabling RCS, and resetting their default messaging app. Only a single participant voluntarily withdrew at the midpoint.

We compensated participants who completed the midpoint survey $\$25$ via Prolific or a gift card. The text of the questions in the midpoint survey is available in Appendix G.1.5. The average time to complete the survey was 5 minutes and 19 seconds.

Conclusion Survey

Approximately 14 days after starting the study, participants were sent the conclusion survey. The survey began with the same questions from the midpoint survey.

To test $\mathcal{H}_3$, participants next completed a phishing identification task using the same interface from the pre-studies (see Section 9.1). All participants viewed the same eight stimuli (four benign, four fraudulent) and were asked to classify each as fraudulent or legitimate on a five-point scale (from "Definitely fraudulent" to "Definitely legitimate"). The stimuli shown are located in Appendix G.1.9. The messages were designed to resemble those sent during the experiment; all fraudulent messages were service-impersonation messages based on real messages and employed various URL-impersonation techniques.

To test $\mathcal{H}_4$, participants were next asked a series of warning-specific sentiment questions. These questions were substantially similar to those asked in pre-study 1 and 3 (see Section 9.1).

We first showed participants a screenshot of the study app with the appropriate warning above the simulated Bank of America scam message, and asked whether they recalled seeing the warning during the study. Participants who said "no" were directed to the debrief portion of the survey, without answering sentiment questions. Participants who said "yes" or "I'm not sure" answered seven Likert-scale questions assessing various subjective aspects of the warning design and their trust in AI to analyze text messages accurately. After the seven Likert scale questions, participants were asked to briefly justify their ratings.[6] We then asked participants whether they would make any changes to the design they viewed and (if so) to explain what changes they would make. Finally, they were asked to view the alternative warning design, rate it on a five-point scale (from "Much worse" to "Much better"), and justify their answer.

The survey provided a debriefing to participants. The debrief specified that the study focused on warning design and that we hypothesized GAI warnings would be more effective than generic warnings. It also explained that we sent them simulated scam messages and linked to a Google Drive document containing them.

After the debrief, we asked participants about their perception of the simulated messages prior to the debrief to evaluate the ecological validity of the experiment. This question provided a list of possible options, ranging from complete awareness ("I noticed an increase in scam messages, and I thought some or all of the scam messages were sent as part of the study") to complete lack of awareness ("I did not notice an increase in scam messages."). As detailed in the Appendix, the list of options was expanded after the first two participants completed the survey, based on qualitative feedback from the midpoint survey. We also added an open-ended question asking participants to justify their response at this point. The responses informed our interpretation but were not formally coded.

After the debrief, participants were asked whether they now wished to withdraw from the survey and have their data deleted. They were assured that their compensation would not be affected, regardless of their choice. Only a single participant chose to withdraw. The remaining participants were shown the complete list of messages we collected during the study, organized into groups by message sender. Each message and sender was associated with a unique ID. Participants could enter the ID of specific messages or senders they wanted deleted in a text box. Participants could also write "delete all messages" (or something similar) to delete all collected messages. The final screen of the survey provided instructions on uninstalling the study app, re-enabling RCS, and resetting their default messaging app.

We compensated participants who completed the conclusion survey $\$35$ via Prolific or a gift card. The text of the questions in the conclusion survey is available in Appendix G.1.6. The average time to complete the survey was 29 minutes and 57 seconds.

9.3.3 Recruitment

Prior to the experiment, we used GPower 3.1 [169] to estimate the required sample size to detect a small behavioral effect ($f = 0.1$) with 80% power.[7] The result of this analysis was a sample size of 70. We hypothesized that we would see a 30% attrition rate, meaning that we would

[6]The justification question was added after 4 participants completed the survey.

[7]We modeled the study as a repeated measures ANOVA, with the following parameters: $f = 0.1$, $\alpha = 0.05$, $\beta = 0.80$, $n_{groups} = 2$, $n_{measures} = 4$, Corr. among repeated measures $= 0.75$, and $\epsilon = 1$. This analysis does

need to onboard 100 participants to achieve 70 full study participants. As will be discussed, in practice, participants who successfully installed the study app were unlikely to withdraw.

We recruited the first group of participants via physical and digital flyers that contained a link/QR code to the consent form for the screening survey. The flyer was posted in public spaces in Pittsburgh, PA (e.g., utility poles, bus stops, community boards) and on the campus of Carnegie Mellon University. A digital copy of this flyer was sent to various community organizations in Pittsburgh, PA, some of which shared the flyer on their mailing lists or social media groups. The digital flyer was also shared on the researchers' personal social networks via private posts. Finally, the digital flyer was briefly posted to a Reddit community dedicated to paid research opportunities (`/r/PaidStudies`) and an associated Facebook page. However, these advertisements were removed after approximately 48 hours due to a large volume of fraudulent or low-quality participants (see Section 9.3.4 on fraud prevention). We contacted participants who were determined to be eligible via SMS and sent them the consent form and onboarding survey for the full study.

From May 31st to June 25th, 2026, 284 participants completed the screening survey via the digital or physical flyers, 122 (43.0%) of which were not automatically screened out. 75 (61.5%) of these respondents were invited to participate, and 58 (77.3%) subsequently completed the consent form. Of these participants, 23 (39.7%) completed at least part of the study. 21 of these participants completed the full study.

We recruited the majority of participants via the crowdworking platform, Prolific. We obtained approval from Prolific to run our study, as is required for studies that request personally identifiable information on their platform. We recruited a small cohort of participants on June 23rd, 2026, to test Prolific's viability for the study. The remaining participants were recruited on June 30th, 2026.[8] Eligible participants were automatically redirected to the consent form and onboarding survey for the full study.

In total, 168 participants completed the screening survey via Prolific, 88 (52.3%) of which were not screened out. All of these participants completed the consent form, and 69 (78.4%) completed at least part of the study. 61 of these participants completed the full study.

Across both recruitment sources, 92 participants completed at least part of the study; however, one participant withdrew after debriefing. This results in the final sample size of 91, 81 of whom completed the full study. This exceeds the pre-study estimate of 70 participants, so we did not recruit additional participants. The recruitment materials used in the study are available in Appendix G.1.1.

9.3.4 Fraud Prevention

We took several steps to help protect data quality. We enabled Qualtrics' built-in duplicate-response detection and IP geolocation features. Respondents who had previously taken the survey or were located outside the US were automatically excluded.

not perfectly match the experimental design, as GPower does not support modeling GLMMs. Nevertheless, it was chosen to provide a lower-bound estimate of the required sample size.

[8] A single additional participant was automatically recruited on Prolific on July 2nd, 2026, after a previously enrolled participant failed to matriculate to the full study.

After posting to the Reddit and Facebook group discussed in Section 9.3.3, we experienced several challenges. First, we noted respondents who "straightlined" the business familiarity portion of the survey. These respondents indicated that they used all of the listed businesses except the fake businesses. Subsequent participants recruited via flyers who answered in this manner were not invited to participate in the full study. We also noted a number of users connecting to the server with international or fake phone numbers (e.g., `+15551234567`). These phone numbers did not match any of the numbers provided in the screening survey, so we could not identify which potential participants were responsible for them. We hypothesized that fraudulent respondents were using VoIP numbers to appear to be US-based and receive the onboarding survey, then installing the app on an emulator or non-US device to connect to the server. We subsequently began to use the Twilio Phone Number Lookup API[9] to check whether the numbers provided in the screening survey were registered as VoIP numbers rather than genuine cell numbers. Only real cell numbers could be used with the study app, so participants who provided VoIP numbers were not invited to the full study. We discarded data from fraudulent participants who connected prior to the implementation of this procedure.

Separately, to ensure that participants who enrolled in the full study continued to use the app, a member of the research team regularly monitored the metrics database. We checked for participants whose phones had not connected to the server for 24 hours, and for those who had connected but were not recorded as interacting with the app (e.g., not opening the app). Participants who were detected as inactive were sent a test message from a study number (see Appendix G.1.2). Because this number was not in the participants' contacts, receiving this message should trigger the app to connect to the server for classification. We then contacted participants who failed this test and, if appropriate, withdrew them from the study. Two participants who were contacted in this manner reported issues with the app continuously crashing. They were directed to reinstall the app and were then able to remain in the study. Nine other participants were withdrawn from the study. The survey responses and behavioral observations from prior to the period of inactivity were retained, unless the participant requested deletion after being debriefed (see Section 9.3.7).

9.3.5 Statistical Analysis

Based on our hypotheses, we conducted a series of pre-planned analyses similar to those used in Pre-studies 1 and 3 (see Section 9.1). After reviewing the data, we added an additional set of post-hoc analyses. We first discuss how we prepared the behavioral data for analysis, and then describe the pre-planned and post hoc statistical analyses separately.

All statistical analyses were performed using R version 4.3.3 and several packages. Of note, mixed-effects regressions were fit with `lme4` (version 1.1.35.1) using `glmer()` with the `bobyqa` optimizer for logistic regressions and `lmer()` for the regression on SUS scores. Post-hoc pairwise comparisons were computed with `emmeans` (version 1.10.2). All behavioral analyses in this chapter concern the eight simulated scam messages. Participants also received genuine messages during the study, some of which were classified as scams and displayed with warnings. Because the content, timing, and volume of these messages varied uncontrollably across partic-

[9] `https://www.twilio.com/en-us/user-authentication-identity/lookup`

ipants, they cannot be treated as experimental stimuli, and we do not analyze responses to them here. We plan to report the classification and handling of genuine messages in future work.

Data Cleaning

For all statistical analyses of behavior, we reduced interactions with each message to the single 'worst' action each participant performed. For example, if a participant opened a message, clicked a link in it, and then pressed the "Report & Block" button, they would be recorded as having clicked the link. While clicking a link does not necessarily mean that someone will provide sensitive information in a phishing attack, it is the first step towards being victimized and is a common metric in phishing simulation studies (e.g., [227, 272, 273, 414]). We also did not want to design a simulated phishing webpage to collect personal information, to reduce the risk of alarming participants who provided information. We did not consider replying to messages hazardous, since most did not request a direct reply. The Bank of America message included the phrase "Text STOP to quit," but responding in this manner mainly risks alerting a fraudster that the phone number is active.

Data from some participants required special handling. Reinstalling the study app could reset the configuration, resulting in some participants not receiving proper randomization and receiving the same messages twice. In addition to the two participants who reinstalled the app after being instructed to do so (see Section 9.3.4), $P58$ and $P82$ apparently also reinstalled the app without being instructed to do so. We considered only the behavioral observations these participants saw prior to reinstalling the app. One participant received early disclosure of the presence of simulated scam messages (see Section 9.3.7). All of their behavioral observations were excluded from the analysis because the disclosure occurred at the start of the experiment.

After this filtering, 647 simulated scam messages remained for analysis.

Pre-planned Analyses

To test $\mathcal{H}_1$, that the context-specific warning would result in fewer participants clicking links or calling phone numbers within simulated scam messages, we compared two mixed-effects logistic regressions with clicking the link within the message (including the `tel:` link within the Bank of America message) as an outcome variable and participant as a random effect. The models included only observations in which participants opened the simulated message, because the warning was not visible unless the message was opened. The first model included only the random intercept for the participant and an eight-level categorical fixed effect for message identity. The effect of message content on phishing susceptibility is well established [142, 349, 570], so a model that includes it was deemed a reasonable baseline. The second model added a three-level categorical fixed effect representing which condition the observation was collected under: control (i.e., participants seeing no warning), *long*, or *generic*. For this analysis to support $\mathcal{H}_1$, the second model must fit significantly better than the base model, and a post-hoc pairwise comparison between the warning conditions must show that participants in the *long* condition were significantly less likely to click links or call numbers than those in the generic condition. We compared models rather than fitting the condition directly so that the improvement in fit due

to the condition could be tested formally against a baseline that already accounts for message identity.

To test $\mathcal{H}_2$, that the context-specific warning would result in more participants reporting simulated scam messages, we performed the same GLMM analysis, except with pressing the "Report & Block" as the outcome variable. Since the "Report & Block" button was only visible under warning conditions, these models were fit to the subset of observations in which participants opened the scam messages and saw a warning. Accordingly, the model testing the effect of specific warning conditions only compared the *generic* and *long* warning conditions. For this analysis to support $\mathcal{H}_2$, the second model must fit significantly better than the first, and the coefficients must show that participants in the *long* condition were significantly more likely to report scam messages than participants in the *generic* condition.

To test $\mathcal{H}_3$, that participants shown the context-specific warnings will perform better on a post-study phishing identification task, we analyzed participants' responses to the conclusion survey phishing identification task using Signal Detection Theory (SDT) [455]. For scam stimuli, we defined *hits* as selecting that a message was "Probably" or "Definitely fraudulent", while *misses* were selecting that a scam message was "Probably" or "Definitely legitimate." For legitimate stimuli, *correct rejections* were selecting "Probably" or "Definitely legitimate" while *false alarms* were selecting "Probably" or "Definitely fraudulent." From these rates, we computed sensitivity (d') and response bias (c) for each participant. Stimuli where a participant selected "I'm not sure" were excluded, as they could not be clearly categorized (45 or 13.9% of scam message ratings and 69 or 21.3% of legitimate message ratings). To account for *hit* rates of 0 and *false alarm* rates of 1, we performed the loglinear correction [455]. For d' we performed a one-sided Wilcoxon Rank-Sum (Mann-Whitney U) with continuity correction, testing the hypothesis $\mathcal{H}_{3a}$ that participants who saw the *long* warning would have greater discrimination ability than participants exposed to the *generic* warning. Our overall hypothesis did not yield a directional prediction for c, so we ran a two-sided Wilcoxon Rank-Sum test with continuity correction, evaluating $\mathcal{H}_{3b}$ that there is a significant difference in response bias between participants exposed to the *generic* and *long* warnings. The Wilcoxon Rank-Sum test is a non-parametric alternative to the t-test that does not require the data to be normally distributed [307]. When d' and c are computed from a small number of observations, the values are non-continuous, violating the normality assumption of a standard t-test. Continuity correction is necessary because the resulting ties violate the assumption of an exact null distribution underlying the standard Wilcoxon Rank-Sum test.

To test $\mathcal{H}_4$, that participants shown context-specific warnings will report a more positive perception of the warning system than those shown generic warnings, we evaluated a number of sub-hypotheses corresponding to the different sentiment questions. For unprimed sentiment questions, we fit a pair of mixed-effects regressions with participant as a random effect: linear regression for SUS scores and logistic regression for the question identifying whether participants noticed any features they liked. These analyses were designed to test the following subhypotheses:

$\mathcal{H}_{4a}$ Participants shown context-specific warnings will report higher usability (SUS) scores than those shown generic warnings.

$\mathcal{H}_{4b}$ Participants shown context-specific warnings will be more likely to state that they recalled a feature they liked than those shown generic warnings.

For the logistic outcome, we grouped observations in which participants responded "I'm not sure" ($n = 12$) with the "No" option. We did this so that our analysis captured instances in which participants unambiguously noticed a good feature.

As with the behavioral hypotheses, we compared two nested models via LRT for each outcome. The first model was a null model that included only the random intercept for the participant. The second model added a three-level categorical fixed effect representing the condition under which the observation was collected: control (i.e., seeing no warning), long, or generic. For $\mathcal{H}_{4a}$ to be true, the SUS model including the warning condition must fit significantly better than the model including only the random effect, and a post-hoc pairwise comparison between the warning conditions must show that participants in the *long* condition reported higher SUS scores than those in the *generic* condition. For $\mathcal{H}_{4b}$ to be true, the good-features model, including the condition must fit significantly better than the model including only the random effect, and a post-hoc pairwise comparison between the warning conditions must show that participants in the *long* condition were more likely to identify features they liked than those in the *generic* condition.

For the primed sentiment questions, we translated the Likert-scale sentiment by mapping each scale point to its corresponding integer value (1 to 5) and used the Wilcoxon rank-sum test (Mann-Whitney U) with continuity correction to determine whether ratings in one condition tended to be higher than in the other [307]. In line with our overall directional hypothesis, we tested the following sub-hypotheses:

$\mathcal{H}_{4c}$: Participants exposed to the *long* warning will rate it as **more** useful to themselves than participants exposed to the *generic* warning.

$\mathcal{H}_{4d}$: Participants exposed to the *long* warning will rate it as **less** annoying than participants exposed to the *generic* warning.

$\mathcal{H}_{4e}$: Participants exposed to the *long* warning will express **more** trust in the AI warning system than participants exposed to the *generic* warning.

$\mathcal{H}_{4f}$: Participants exposed to the *long* warning will rate it as **more** useful to the average person than participants exposed to the *generic* warning.

$\mathcal{H}_{4g}$: When asked to consider the alternative design, participants who were exposed to the *long* warning will rate the *generic* warning as **worse** than participants exposed to *generic* warning when viewing the *long* warning.

For several of the sentiment questions, our overall hypothesis did not yield specific directional hypotheses. For these, we test the following two-sided hypotheses:

$\mathcal{H}_{4h}$: There will be a difference in the extent to which participants exposed to the *long* warning and participants exposed to the *generic* rate the warning as grabbing their attention.

$\mathcal{H}_{4i}$: There will be a difference in participant ratings of the length of text between participants exposed to the *long* and *generic* warnings.

$\mathcal{H}_{4j}$: There will be a difference in ratings of how easy or difficult it is to understand the warning between participants exposed to the *long* and *generic* warnings.

We analyzed the question asking participants whether they would make any changes to the warning design, using Fisher's exact test to assess whether responses were significantly associated with the condition [181], evaluating $\mathcal{H}_{4k}$ that participants exposed to the *long* warning would be less likely to recommend changes than participants exposed to the *generic* warning.

The results of pre-planned analyses were not corrected. While multiple subhypotheses were evaluated for each main hypothesis, each was specified in advance and answered a distinct question.

Post-hoc Analyses

As will be discussed in Section 9.4, an unexpectedly large portion of the simulated messages were not opened by participants. We fit a series of mixed-effects logistic regressions with whether a message was opened as the outcome variable and participant as a random effect. We compared two nested models via LRT. The first model was a base model that included only the random intercept for the participant and a fixed effect for message identity. The second model added a three-level categorical fixed effect indicating whether classification was inactive, active with the generic warning, or active with the *long* warning (i.e., the warning that would have been shown if the message had been opened). As discussed previously, the warning was not visible until messages were opened; however, it may be that recent exposure to warnings makes one more or less likely to open subsequent scam messages. If this is the case, then we expect the second model to fit significantly better than the first.

For all behavioral models, including the pre-planned analyses and the post-hoc message-opening analysis, we conducted a post-hoc robustness check that incorporated two-level fixed effects for the study week (a "period" effect) and the week in which warnings were observed (a "sequence" effect). The purpose of this analysis was to confirm that our substantive conclusions were not artifacts of unmodeled ordering effects [53]. For each outcome, we began from the best-fitting model identified in the corresponding analysis described above, then compared a model adding these two parameters via LRT. If these unmodeled order-related effects explain a significant portion of the variance, we treat the model that includes period and sequence as the primary model for that outcome and flag the discrepancy in our reporting; for all other outcomes, we retain the original model as the primary model.

Finally, post hoc, pairwise contrasts were computed for some pre-planned regression results. For directional hypotheses, these contrasts were one-sided. Multiple comparisons were corrected using the Tukey method within the `emmeans` package [163]. Other post-hoc tests were not corrected and should be interpreted as exploratory.

9.3.6 Qualitative Analysis

Open-ended responses from the survey were coded by a single author, using an inductive approach [279]. The coder analyzed sets of responses together: unprimed feedback (i.e., from the midpoint survey and the start of the conclusion survey) and primed feedback (i.e., participants' explanations for Likert-scale responses and requests for changes). The coder first open-coded responses, using labels that captured aspects of feedback relevant to the research questions (e.g.,

for unprimed sentiment, whether warnings were mentioned). After initial coding, these labels were consolidated and standardized to create codes in four main categories:

- **Effect of Warning on Actions**: Participants' perception of how the warning designs affected their behavior.
- **Warning Feedback**: Negative and positive feedback about the warnings, including specific recommendations for changes.
- **QUIK Good Features**: Positive feedback about features that are unrelated to the warning system
- **Bugs or Complaints**: Negative feedback unrelated to the warnings, especially bugs or crashes.

The codebooks are available in Appendix G.1.10. All codes could be applied across both unprimed and primed feedback; however, certain codes were more common in unprimed or primed feedback. For example, discussion of non-warning-related features was more likely to appear in the unprimed questions. Given the nature of the data, the coder decided to adopt codes from Chapter 8 to capture similar feedback. The resulting codebook thus includes both novel codes and previously developed themes, although it should be stressed that this analytical choice was made after viewing the data rather than as *a priori*, deductive choice.

We viewed single coding as acceptable, since qualitative analysis is not the primary basis for our findings. Due to the subjective nature of qualitative coding, especially when conducted by a single author, quantitative or semi-quantitative descriptors (e.g., "many," "few," "several") should not be interpreted as general statements about the frequency of certain perspectives in the population.

9.3.7 Ethics

This experiment was approved by our IRB. Participants completed informed consent forms for the screening survey and, separately, for the full study prior to onboarding; both covered study procedures, data collected, and participant risks.

The study used incomplete disclosure: participants were told we were testing a novel SMS application that classifies text messages using generative AI and that we might send them messages, but were not told that the study focused on warnings or that the messages would include simulated scams. This was necessary because disclosing our focus would have altered participants' natural behavior. To preserve participants' right to informed participation, we debriefed them on the study's full nature after completion and gave them the opportunity to withdraw consent and request deletion of all collected data (see Appendix G.1.6). Participants who withdrew early received a disclosure message and were asked to notify the research team if they wished to have their data deleted (see Appendix G.1.2). One participant who completed the full study withdrew consent and requested data deletion. None of the participants who withdrew early requested data deletion.

One participant received an early disclosure. He contacted the research team after becoming alarmed by the first message he received (the Bank of America scam message). He feared that his data had been compromised. He chose to remain in the study after the disclosure.

Collecting SMS traffic poses several risks if the data were accessed by someone outside the research team. Beyond general privacy exposure, an attacker who gained access to the database during the study could use a participant's phone number to request authentication codes sent via SMS, potentially compromising their accounts. SMS messages are unencrypted by default, and this risk is not unique to our data collection: any tool that has access to a user's messages, such as notification-mirroring apps on a desktop computer, poses a similar risk. Nevertheless, we took several steps to minimize this risk: data was sent to the server over HTTPS; at rest, data were stored on a password-protected server accessible only via the university network; phone numbers were removed from the database as soon as identity was verified, reducing the window in which an authentication attack was possible; and OpenAI received only message contents, never sender/receiver numbers.

To reduce risks in retained and released data, we took steps to address two distinct concerns: preventing participant identification and limiting potential harm if identification were to occur despite these protections.

To reduce identification risk, phone numbers, Prolific IDs, and Android hardware IDs were redacted or replaced with randomly generated numeric strings after the study. Messages were scanned with Presidio,[10] a Python package for identifying PII in text, and a research team member reviewed each message to redact identifying information, including things Presidio missed (e.g., contextual identifiers like naming someone as a company founder). We plan to release text message data in a disaggregated form, meaning specific threads will not be grouped together by participant.

To limit harm from sensitive content even if a participant were identified, we redacted the full contents of highly sensitive messages, including those that appear to discuss potentially criminal acts or intimate relationships. As discussed in Section 9.3.2, participants were given the opportunity to review the collected messages and could request deletion of any or all messages during the conclusion survey. We deleted messages from participants who didn't complete the study, since they had no opportunity to review their data.

9.3.8 Limitations

Our study has several important limitations. While the phishing simulation methodology provides greater ecological validity than traditional lab studies, participants may still have behaved differently than they would when encountering a scam message outside a study context. Notably, 26 of the participants who completed the full study (32%) indicated that they "thought some or all of the scam messages were sent as part of the study." While this may be an overestimation of the true proportion due to hindsight bias [408], it suggests that a sizable proportion of participants became aware of the simulation, and may have behaved unnaturally. Participants' behavior may also have been affected by the use of a new, unfamiliar messaging app; several participants mentioned missing RCS features such as read receipts, typing indicators, and emoji reactions, though it is unclear how this may have influenced their responses to the simulated scam messages.

We also do not claim a representative sample (see Section 9.4.1). By design, participants in our study had to be Android users with regular access to the internet and be willing and able

[10] `https://presidio.dataprivacystack.org/project_transition/`

to install an app via a non-standard, multi-step process. 30 participants reported prior work experience or education in computer science, IT, or a related field (34% of all participants). More generally, the sample was highly educated and biased towards younger participants (see Section 9.4.1 for a more detailed description of participant demographics). iOS users may experience and respond to scam messages differently, given differences in rich messaging features across platforms. More generally, participants from different backgrounds may experience scams differently.

While the repeated-measures design offered greater statistical power for a given sample size, it may also introduce unmodeled carryover effects that could confound the estimated main effects [53]. Moreover, some participants were exposed to warnings on messages outside the study's simulated stimuli, either from genuine scam texts they happened to receive or from the classification system misidentifying a legitimate message as fraudulent. These classifications likely affected their perception of the system. Our experiment may be underpowered, increasing the risk of false-negative results in statistical testing. Our pre-experiment power analysis calculated the necessary sample size based on behavioral effects alone, assuming that all participants would receive and open all eight messages, which turned out not to be the case. Additionally, while each participant contributed up to eight behavioral observations, only one or two sentiment observations were made (at the midpoint and conclusion), so the repeated-measures structure that gave our behavioral models their statistical power was largely unavailable for these outcomes. Lastly, our qualitative analysis was conducted by a single coder (see Section 9.3.6), so the themes we report should be read as illustrative of the range of participant experience rather than as evidence about its distribution.

9.4 Results

In this section, we describe the results of our experiment, beginning with a brief discussion of the participant sample to contextualize them. We then describe participants' behavior, performance in the phishing identification task, and warning-related sentiment, with special attention to our hypotheses.

9.4.1 Participant Characteristics

Table 9.1 shows an overview of the demographics of the sample and their condition assignments. Of particular note, our sample for this experiment was highly educated, which may introduce similar limitations to those observed in the interview study in Chapter 8. Specifically, this skew toward higher education may impact the study's results, as reading comprehension could affect the processing of warning information. Users with lower reading comprehension may experience the warnings differently. The sample also skews young, with 60% of participants aged 25-44 and only 1 participant over 65. As discussed in Chapter 3, the specific role of age in phishing susceptibility is unclear, but, regardless, this trend means the perspectives of older adults are underrepresented. Finally, as discussed previously, a large portion of the sample (34%) reported past education or work experience in computing, IT, or a related field. This greater level of

Table 9.1: Summary of participant demographics collected via the onboarding survey and cohort assignment. Includes all 91 participants who completed part of the study. Some demographic category names are shortened for space, but the complete text can be found in Appendix G.1.4.

Age (Years)		Race/Ethnicity		Education	
18 to 24	11%	American Indian or Alaska Native	0%	Less than high school	1%
25 to 34	26%	Asian	7%	Graduated high school	10%
35 to 44	34%	Black	11%	Some college education	24%
45 to 54	20%	White	82%	Associate's degree	9%
55 to 64	8%	Self-describe	1%	Bachelor's degree	36%
65 to 74	1%	No response	2%	Degree beyond bachelor's	20%
75 or older	0%				
Gender		**Cohort**		**Household Income**	
Man	51%	Week 1 Long Warnings	30%	Less than $25,000	10%
Woman	45%	Week 2 Long Warnings	23%	$25,000 to $49,999	25%
Non-binary	2%	Week 1 Generic Warnings	27%	$50,000 to $99,999	38%
Self-describe	1%	Week 2 Generic Warnings	20%	$100,000 to $149,999	16%
				More than $150,000 per year	9%
				No Response	1%

technical experience could result in an above-average ability to identify the scams within this study.

9.4.2 Behavioral Analysis of Simulated Messages ($\mathcal{H}_1$ and $\mathcal{H}_2$)

We analyzed actions in two stages: those taken from the notification or inbox, where no warning was visible, and those taken after opening a message, where a warning could be shown. We first discuss inbox-level actions. Figure 9.2 shows the proportion of participants who opened each message alongside the other actions that participants could take without opening. Unexpectedly, only 425 (65.7%) of the simulated scam messages that participants received were opened. QUIK enables users to reply to notifications without opening the underlying message. Additionally, from the inbox screen, users can delete or archive messages without opening them. Archiving was not recorded as a metric, so we cannot determine the proportion of participants who did so.

The logistic regression model predicting opening that incorporated the type of classification that was active **did not** achieve significantly better fit than the base model that considered only message factors ($p = 0.549$, $\chi^2 = 1.1982$, df $= 2$). However, the model that incorporated period and sequence effects achieved a significantly better fit than the base model ($p = 0.003$, $\chi^2 = 11.815$, df $= 2$), though this improvement was driven by the period effect rather than the sequence effect. The model parameters are in Table 9.2. For message identity, the CVS impersonation message was significantly less likely to be opened than the arbitrarily selected baseline of the Bank of America message, perhaps because (even if assumed to be real) it seemed more like spam than a message requiring interaction. Conversely, the ID.me message was significantly more likely to be opened than the Bank of America message, which could be due to the per-

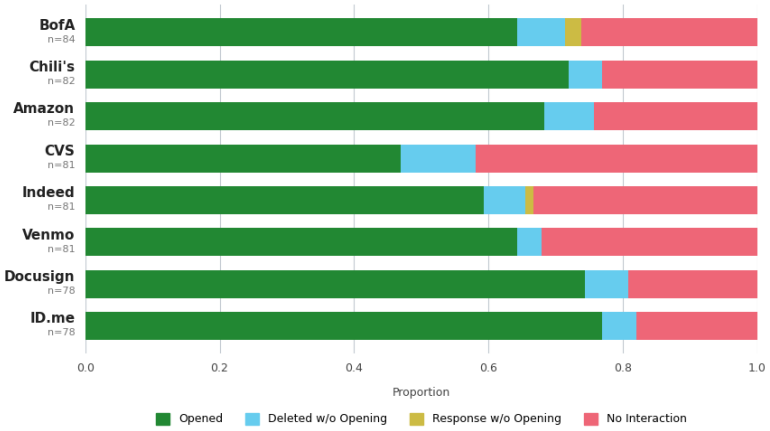


Figure 9.2: Inbox/notification behavior by message.

ceived urgency of one's ID being used without permission. Holding message identity constant, participants were significantly less likely to open messages in the second week than in the first week. 69.6% of the messages sent in week 1, compared to 61.3% of the messages sent in week 2, were opened. This suggests that participants became more vigilant or adapted to the simulated stimuli over the course of the study.

In support of this interpretation, some participants spontaneously reported noticing more spam or scam messages (coded as *potential recognition of study design*). For example, P30 stated in their midpoint survey response "I have received some unexpected spam... it's unusual for me to receive unsolicited messages..." Similarly, P72 stated in their midpoint survey response, "i think im getting more spam than usual, ive just been ignoring it or archiving it." These quotes suggest that participants noticed and adapted to increased spam, although they did not explicitly attribute it to an experimental manipulation. More directly, when asked about their perceptions of the simulated messages at the conclusion of the study, only 13 participants (16% of those who completed the study) indicated that they "didn't notice an increase in scam messages."

Figure 9.3 shows the proportions of each measured action taken after a message was opened, grouped by message and condition. In addition to the modeled variables (clicking and reporting), we also recorded whether a message was deleted (37 or 8.7% of opened messages) or responded to (12 or 2.8% of opened messages). We did not record the contents of replies to simulated messages, in line with our promise not to record the contents of messages that participants sent. Given that most replies were to the Bank of America message, which included an instruction to reply STOP to opt- ut of texts, we think it's likely that participants were replying with "STOP" or a similar phrase. Rahman et al. observed participants responding to messages in this manner during their phishing simulation study [387]. If this is the case, replying may indicate that the participant did not recognize that the message was a scam. Still, the danger of this behavior is less than clicking a link.

Narrowing in on our hypotheses, we found that warnings had an effect on participant behavior, but our specific hypothesis $\mathcal{H}_1$ **was not supported**. The model incorporating the effect of warnings fit the data significantly better than the base model that considered only message factors

Table 9.2: Mixed-effects logistic regression predicting whether a simulated scam message was opened. The Bank of America message is the reference level for message identity, week 2 for study period, and week 2 classification for sequence. Odds ratios are e^{β}. $N = 647$ observations from 90 participants; random intercept variance $= 3.679$ ($SD = 1.918$).

	β	OR	$S.E.$	z	p
(Intercept)	0.7914	2.21	0.4624	1.711	0.087
Message identity (ref. = Bank of America)					
Chili's	0.6380	1.89	0.4296	1.485	0.138
Amazon	0.2984	1.35	0.4211	0.709	0.479
CVS	−1.2578	0.28	0.4195	−2.998	0.003**
DocuSign	0.8384	2.31	0.4387	1.911	0.056
ID.me	0.9441	2.57	0.4492	2.102	0.036*
Indeed	−0.3303	0.72	0.4137	−0.798	0.425
Venmo	0.1444	1.16	0.4208	0.343	0.731
Order effects					
Sequence (warnings in week 2)	0.7900	2.20	0.4822	1.639	0.101
Period (week 2)	−0.6653	0.51	0.2255	−2.951	0.003**

*$p < 0.05$; **$p < 0.01$

Table 9.3: Mixed-effects logistic regression predicting whether a participant clicked a link (or phone number in the Bank of America message) after opening a simulated scam message. The control condition (no warning) is the reference level for warning condition, and the Bank of America message for message identity. Odds ratios are e^{β}. $N = 425$ observations from 84 participants (i.e., only those opened); random intercept variance $= 3.760$ ($SD = 1.939$).

	β	OR	$S.E.$	z	p
(Intercept)	−4.5882	0.01	1.1912	−3.852	< 0.001***
Warning condition (ref. = control)					
Long	−1.4741	0.23	0.6416	−2.298	0.022*
Generic	−1.0882	0.34	0.6173	−1.763	0.078
Message identity (ref. = Bank of America)					
Chili's	1.8225	6.19	1.1818	1.542	0.123
Amazon	2.6136	13.65	1.1623	2.249	0.025*
CVS	−0.3705	0.69	1.6088	−0.230	0.818
DocuSign	1.7901	5.99	1.1889	1.506	0.132
ID.me	2.1246	8.37	1.1663	1.822	0.069
Indeed	0.5023	1.65	1.3666	0.368	0.713
Venmo	1.4966	4.47	1.2235	1.223	0.221

*$p < 0.05$; ***$p < 0.001$

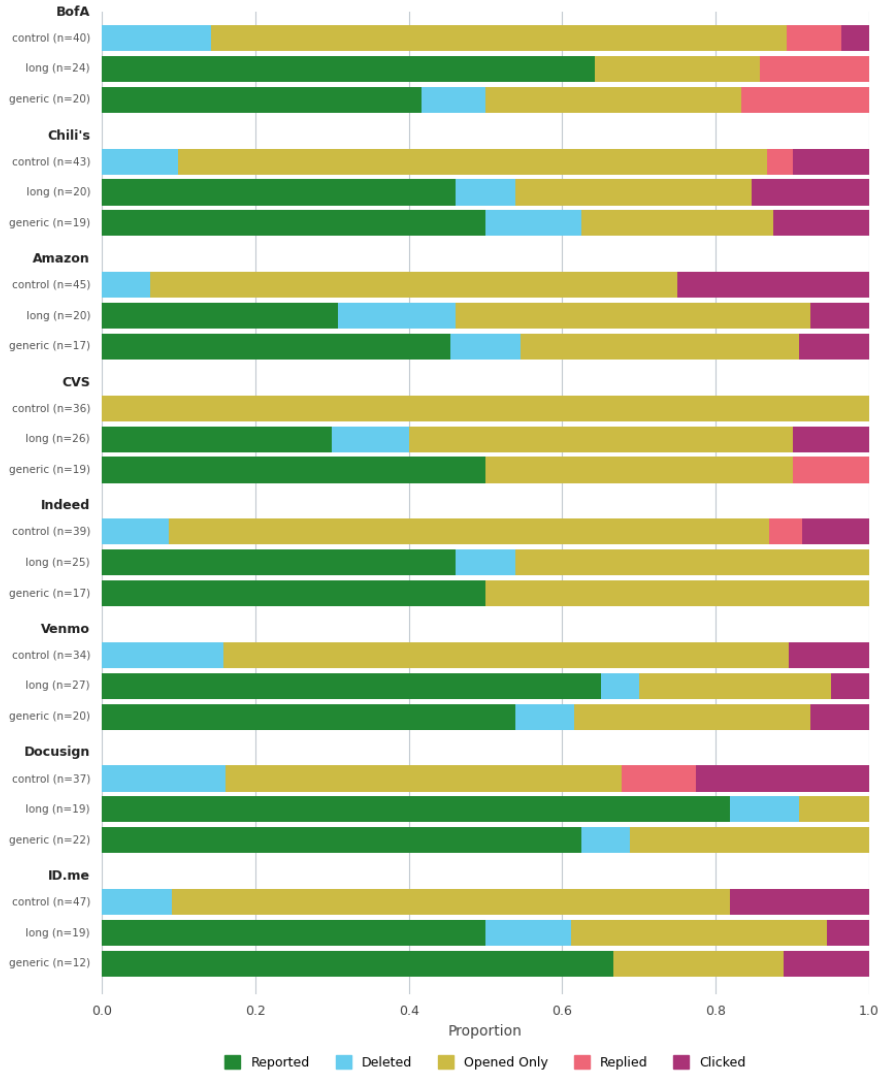


Figure 9.3: Message outcomes by warning condition.

($p = 0.012$, $\chi^2 = 8.8771$, df $= 2$). Table 9.3 shows the model parameters. Within this model, the *long* warning significantly reduced clicking, holding the message constant. The *generic* warning also reduced clicking, although the effect did not reach the $p < 0.05$ significance level. This may be due to the lower number of observations with the *generic* warning. Critically, a post-hoc comparison between the two conditions did not show that the *long* warning was associated with significantly lower click rates than the *generic* warning ($z = -0.456$, $S.E. = 0.847$, $p = 0.324$).

This result could mean there is truly no difference, but it may also reflect limitations of our experimental design. Most notably, our experiment was likely underpowered due to low message-opening rates and uneven participant assignment across conditions. Clicking was also generally rare, with messages receiving few interactions even in the control condition. This leaves less room for one warning to outperform the other. Ultimately, the results are consistent with *long* being more effective than *generic*, but we cannot draw this conclusion.

For message factors, only the Amazon impersonation message differed significantly from the arbitrarily selected baseline of the Bank of America message. Participants were significantly more likely to click the link within this message, holding the warning condition constant. This result may be due to the greater contextual relevance of Amazon to most participants. In the screening survey, 89 of our participants (98%) reported using Amazon in the last year, as compared to 61 (67%) for Venmo, 44 (48%) for Indeed, and 11 (12%) for Bank of America. The fake Chili's reservation message and ID.me alert also had coefficients strongly suggesting that their links were more likely to be clicked, but did not reach $p < 0.05$ significance.

In contrast to message opening, the study period did not seem to significantly impact participants' link-clicking behavior. The post-hoc logistic regression incorporating sequence and period effects did not fit the data significantly better ($p = 0.381$, $\chi^2 = 1.932$, df $= 2$).

Table 9.4: Mixed-effects logistic regression predicting whether a participant pressed "Report & Block" after opening a simulated scam message with a warning displayed. The Bank of America message is the reference level. Odds ratios are e^{β}. $N = 211$ observations from 72 participants (i.e., only those messages that were opened and displayed with a warning); random intercept variance $= 7.120$ ($SD = 2.668$).

	β	OR	$S.E.$	z	p
(Intercept)	-0.3095	0.73	0.7110	-0.435	0.663
Message identity (ref. = Bank of America)					
Chili's	0.3803	1.46	0.8922	0.426	0.670
Amazon	-0.6936	0.50	0.9156	-0.758	0.449
CVS	-1.0240	0.36	0.9691	-1.057	0.291
DocuSign	1.4939	4.45	0.9151	1.633	0.103
ID.me	0.5249	1.69	0.8819	0.595	0.552
Indeed	0.6693	1.95	0.9272	0.722	0.470
Venmo	0.9672	2.63	0.8466	1.143	0.253

$\mathcal{H}_2$ was also not supported. In this case, the logistic regression modeling reporting that incorporated the effect of the warning condition did not fit the data significantly better than the base model considering message factors alone ($p = 0.969$, $\chi^2 = 0.0015$, df $= 1$). The model parameters are in Table 9.4. None of the messages varied significantly from the Bank of America baseline. The variance of the random intercept is notably higher than the models for opening and link clicks, suggesting that there was more individual variance in reporting behavior. In other words, some participants reported most scams, while others reported few or none.

As with link clicks, the study period did not seem to significantly impact participants' reporting behavior. The post-hoc logistic regression model incorporating sequence and period effects did not fit the data significantly better ($p = 0.189$, $\chi^2 = 1.7263$, df $= 1$).[11]

We did not directly ask participants about their actions in response to scams. Still, qualitative evidence explaining actions did emerge incidentally in sentiment questions. For example, in their

[11]Because this model included only observations where participants saw warnings, study period and sequence are equivalent. As such, only one of the two variables was included.

midpoint survey response, P76 discussed their appreciation for the *long* warning, explaining "...I really like that the app explained why [the message] seemed like a scam, because a few were from places I do visit (Chilis for example, and it is my birthday soon, so I almost assumed it was a surprise from someone else)..." This quote supports the effect of the evidence within the warning and also highlights the importance of contextual relevance as a factor in susceptibility.

Discussion of the impact of warnings on shaping actions more often emerged in the conclusion survey, where participants were directly asked to explain their ratings of the warning design. Participants offered similar insights to those of the interview participants in Chapter 8. Some highlighted the warning as having an important role in decision making, such as P56, who stated that the *long* warning "helped me quickly evaluate the message" or P43 who described the *long* warning as providing "...the chance to pause and give a little more consideration to what was received and how I wanted to respond." (both coded as *warning as red flag*). Others suggested that the warnings mostly provided confirmation of a correct initial judgment, including P34 who saw the *generic* warning: "The warning for me served as reassurance for what I already knew..." (coded as *affirmed right decision*). Participants also discussed warnings alerting them to the ability to block messages (coded as *highlighted particular action*) or, in the case of the *long* warning, highlighting evidence of fraud (coded as *highlighted evidence*). Some participants discounted the impact of the warnings on their behavior, especially because they felt they could generally spot scams (coded as *I don't need a warning*). P68 stated in response to the *long* warning, "I think I'm already pretty aware of how to identify text scams..." Ultimately, while these results should be understood as reflections summarizing interactions over a two-week period, these comments provide valuable triangulation for the results from Chapter 8's lab study.

9.4.3 Phishing Identification Task ($\mathcal{H}_3$)

$\mathcal{H}_3$ was not supported. Specifically, the tests for both $\mathcal{H}_{3a}$ (that participants exposed to the *long* warning would achieve greater sensitivity d' than participants exposed to the *generic* warning) and $\mathcal{H}_{3b}$ (that there would be a difference in response bias between the conditions) failed to achieve significance. For $\mathcal{H}_{3a}$, $W = 803$, $p = 0.5288$. For $\mathcal{H}_{3b}$, $W = 704.5$, $p = 0.3177$.

Across the sample, sensitivity was poor with a median $d' = 0.206$ ($S.D. = 0.677$, min $= -1.81$, max $= 2.43$), reflecting that the typical participant had discrimination ability only marginally better than chance. Median response bias $c = -0.319$ ($S.D. = 0.610$, min $= -1.28$, max $= 1.28$) indicates that the typical participant had a bias towards labeling things as scams. This sort of scam bias has been noted in previous studies with phishing detection tasks [275, 419] as well as in Chapter 6's study on YouTube giveaway scams.

9.4.4 Warning Sentiment ($\mathcal{H}_4$)

$\mathcal{H}_4$ was partially supported, with participants exposed to the *long* condition rating the warning system more favorably than participants in *generic* condition across some but not all of the sentiment questions.

Unprimed Feedback

Statistical tests on unprimed sentiment questions (i.e., those which did not specifically ask about warnings) did not support the hypotheses. Specifically, for $\mathcal{H}_{4a}$ (that participants would report higher SUS ratings after being exposed to the *long* warning as compared to the *generic* warning) the model considering which condition participants were exposed to in the prior week did not fit the data significantly better than the base model with only a random effect for participants ($\chi^2 = 1.1013$, df $= 2$, $p = 0.577$). This suggests that the inclusion of warnings, *generic* or *long*, did not significantly impact participants' SUS ratings. Median SUS scores in the conclusion survey were 80 ($s = 16.8$, min $= 30$, max $= 100$), which is above the typical benchmark score of 68 [234]. A post-hoc one-sample Wilcoxon signed-rank test confirmed that this difference was significant with a large effect size ($p < 0.001$, $V = 2656$, $r = 0.521$). This seems most attributable to the QUIK app itself, considering the lack of a detectable impact from warnings.

Table 9.5: Mixed-effects logistic regression predicting whether a participant reported noticing a feature they liked, across the midpoint and conclusion surveys. The control condition (no warnings in the preceding week) is the reference level. Odds ratios are e^{β}. $N = 164$ observations from 83 participants; random intercept variance $= 6.356$ ($SD = 2.521$).

	β	OR	$S.E.$	z	p
(Intercept)	-1.1302	0.32	0.4975	-2.272	0.023*
Warning condition (ref. = control)					
Long	1.5850	4.88	0.6326	2.505	0.012*
Generic	0.6031	1.83	0.6344	0.951	0.342

*$p < 0.05$

In contrast, warnings had an effect on whether participants recalled a feature they liked, but $\mathcal{H}_{4b}$ was not supported. The model incorporating the warning condition fit the data significantly better than the null model ($\chi^2 = 7.6492$, df $= 2$, $p = 0.0218$). The parameters for this model are shown in Table 9.5. The *long* warning had a positive coefficient, indicating that participants had higher odds of reporting they saw a good feature after viewing the *long* warning as compared to seeing no warnings. The coefficient for the *generic* warning was also positive but did not reach the $p < 0.05$ level of significance. Critically, the post-hoc contrast between the two warning conditions was not significant ($z = 1.248$, $S.E. = 0.787$, $p = 0.106$), indicating that participants were not significantly more likely to report seeing a good feature after viewing the *long* warning as compared to the *generic* warning.

Given the significant effect of warnings (if not the specific condition), it is unsurprising that warnings were commonly mentioned in participants' descriptions of good features they saw. Much of the feedback was relevant to both warning conditions. Some participants highlighted that just the presence of the warnings was helpful (coded as *warning*). For example, P59, who saw the *generic* warning, stated "I liked the spam/scam warning. I appreciate the warning when I get a spam call, so getting something similar for texts is nice..." Others discussed the utility of the "Report & Block" button, such as P34, who described the "one tap block feature" in the *generic*

warning as "the most useful feature of the app" (coded as *likes options*). Some participants also valued how well the warnings got their attention (coded as *attention-grabbing*). P56, who saw the *long* warning, stated "I LOVE that it tells you when a text is likely spam. It is super helpful and catches my attention."

Participants exposed to the *long* condition commented on the novel presence of an explanation within the warnings (coded as *likes level of details*, *likes level of evidence*, or *likes level of direction*, based on the nature of the feedback). P35 discussed their appreciation for the *long* warning, explaining "I got a few suspicious texts and immediately getting a notice that popped up that... explained WHY they thought it was a scam was incredibly helpful..." As in Chapter 8, other participants reported that they did not need the warnings but viewed them as *helpful to others*. For example, P4 stated "It flagged possible spam messages and explained why they looked like spam. Most cases I think I would've noticed it myself but still helpful... especially for more vulnerable groups... like non-tech-savvy folks (eg my elderly mom)."

Descriptions of bugs or other negative feedback were dominated by complaints about the lack of Rich Communication Services (RCS) functionality (coded as *missing RCS features*), such as read receipts, emoji reactions, and high-quality media sharing. Some participants also experienced difficulty sending or receiving messages, especially in group chats (coded as *Issues w/ sending or receiving messages*). For example, P26 shared in their conclusion survey that they "could not get images from a fellow Android user when in a group text." These issues likely also stem from the transition away from RCS, as group messages do not seem to transition cleanly between RCS and SMS. Pre-study testing showed that Google Messages on other phones will continue to attempt to send messages over RCS even after the protocol has been explicitly disabled.

A small number of participants alluded to the warnings only working part of the time in their unprimed feedback (coded as *potential recognition of study design*). For example, P50 identified this as a bug in their conclusion survey response, stating "Originally when I started using QUIK it alerted me to potential spam messages... I then noticed that this type of alert stopped, but I still would occasionally receive such messages..." This feedback highlights another way in which the within-subjects study design may have affected our results, by making participants feel the system was less reliable.

Primed Feedback

Primed feedback questions (i.e., questions that directly asked about the warning design) were only asked of participants who selected "Yes" or "I'm not sure" when asked if they recalled seeing the warning during the experiment (63 participants or 78%).

Several of the specific primed hypotheses were supported:

$\mathcal{H}_{4c}$ Participants exposed to the *long* warning rated it as significantly more useful than participants exposed to the *generic* warning, with a small effect size ($\bar{x}_{\text{long}} = 4.0$, $\bar{x}_{\text{generic}} = 3.5$, $W = 622.5$, $p = 0.035$, $r = 0.229$)

$\mathcal{H}_{4f}$ Participants exposed to the *long* warning rated it as significantly more useful to the average person than participants exposed to the *generic* warning, with a small effect size ($\bar{x}_{\text{long}} = 4.5$, $\bar{x}_{\text{generic}} = 4.2$, $W = 632$, $p = 0.019$, $r = 0.262$)

$\mathcal{H}_{4h}$ The two conditions significantly varied. Participants exposed to the *long* warning rated it as significantly more effective at grabbing their attention than participants exposed to the *generic* warning, with a small effect size ($\bar{x}_{\text{long}} = 4.7$, $\bar{x}_{\text{generic}} = 4.3$, $W = 632.5$, $p = 0.029$, $r = 0.276$)

$\mathcal{H}_{4i}$ The conditions significantly varied. Participants exposed to the *long* warning were more likely to rate the text as too long as compared to the *generic* warning, with a moderate effect size ($\bar{x}_{\text{long}} = 3.3$, $\bar{x}_{\text{generic}} = 3.0$, $W = 677$, $p < 0.001$, $r = 0.460$). The mean of 3.0 indicates that, on average, participants exposed to the *generic* warning thought the warning was "Just the right length."

Conversely, the other sub-hypotheses for primed feedback were not supported:

$\mathcal{H}_{4d}$ Most participants rated both warnings as not at all annoying on average ($\bar{x} = 1.2$). Ratings of annoyance for the *generic* condition were not significantly higher than the *long* condition ($W = 506$, $p = 0.594$).

$\mathcal{H}_{4e}$ Most participants expressed moderate trust in AI to accurately classify messages ($\bar{x} = 3.2$). Ratings of trust for the *long* condition were not significantly higher than ratings in the *generic* condition ($W = 499.5$, $p = 0.482$).

$\mathcal{H}_{4j}$ Participants rated both warnings as very easy to understand on average ($\bar{x} = 4.8$). Ratings of the two warnings did not differ significantly ($W = 514$, $p = 0.715$).

$\mathcal{H}_{4g}$ When asked to consider the alternative warning design, participants rated the alternative as slightly worse on average ($\bar{x} = 2.9$). Participants who were exposed to the *long* warning did not rate the *generic* warning as significantly worse than participants exposed to *generic* warning when viewing the *long* warning ($W = 479$, $p = 0.408$).

$\mathcal{H}_{4k}$ 24% of participants stated they would make changes to the warning they saw during the experiment. The proportion of participants who recommended changes in the *generic* condition was not greater than in the *long* condition ($p = 1$).

Much of the positive open-ended feedback in the primed questions was similar to that in the unprimed questions. We thus focus on the negative aspects of this feedback and on responses that contextualize these Likert-scale results. The positive results for $\mathcal{H}_{4c}$ and $\mathcal{H}_{4f}$ provide quantitative evidence for the more general observation that *long* participants expressed an appreciation for details within the *long* warning.

As indicated by the significant result for $\mathcal{H}_{4i}$, some participants felt the *long* warning design had too much text. For example, P75 explained their Likert-scale responses by saying "The warnings were clear, but a little long." Some recommended a design similar to the *accordion* design, which was eliminated during the pre-studies (coded as *request for accordion*). Along these lines, P21 recommended the *long* warning be modified to have "...maybe 1 sentence saying this looks suspicious and do not call that number, with the reason details in the Learn More section." This type of suggestion underscores that many participants see the value in the context-specific text, even if they feel the *long* has too much content.

The length of text, or more precisely, the size of the *long* warning, may explain the result of $\mathcal{H}_{4h}$. Some participants referenced the size of the warning box (*likes/dislikes size*), not just the length of text. One participant directly described the size of the *long* warning as *attention-grabbing*: in discussing why they believed the *generic* warning was "Much worse," P60 pointed

to both the lack of an explanation (coded as *dislikes level of evidence (generic)*) and the smaller size (coded as *dislikes size (generic)*): "This doesn't tell why it's likely to be a scam.... [the *long* warning] takes up a lot more of the screen. Taking up the screen is good to draw attention..." Participants more often highlighted the size of the *long* warning negatively, such as P24, who said "I think that the warning design takes up too much space on the app." Still, this kind of feedback suggests the longer text made the warning more visually prominent.

The null result for $\mathcal{H}_{4j}$ may be viewed as a positive for the *long* warning. It suggests that participants did not view them as significantly more difficult to understand. This result is, of course, limited by the inherent difficulty of accurately reporting one's own ability to understand something.

The null result for $\mathcal{H}_{4g}$ is somewhat confounded by the fact that many participants in both conditions apparently believed that the "Learn more" button in the *generic* warning would lead to the same text present in the *long* warning (coded as *thinks learn more is accordion*). For example, when viewing the *long* warning as an alternative, P6 rated it as "Neither better nor worse," explaining, "I didn't ever click on the 'learn more' button during the study, but I can imagine these details being exactly what I would want to see if I did click on that button..." P65, who saw the *long* warning during the study, requested an accordion-style warning when asked what changes they would make. Upon seeing the *generic* warning, they stated: "This is the suggestion I just said! A smaller box where you could collapse the message. Looks alot less clunky..." This confusion almost certainly stems from using a single screenshot to represent the warning, which failed to capture its interactive behavior.

We hypothesized that the AI-generated explanatory text might lead to greater trust in AI's ability to classify phishing messages in general; however, the null result for $\mathcal{H}_{4e}$ suggests that this was not the case. Participants commented both positively and negatively. In the positive case (coded as *accurate*), participants described it as providing true positives and true negatives. For example, P80 explained their ratings of the *generic* warning by stating, "Multiple scams were identified, and regular automated texts were not falsely flagged." Some of the negative feedback (coded as *inaccurate*) focused on false negatives and is therefore confounded by the study design. In other cases, participants reported false positives. P19 stated "9/10 times the warning was correct. I found it false flagged new numbers too easily, which was a problem..."

A comprehensive evaluation of the classification results is beyond the scope of this chapter; however, we investigated P19's messages as a case study in light of their feedback. They received two real messages that were classified as phishing and appeared with warnings: one providing an Amazon gift card and another soliciting participation in a medical research study. The former was unambiguously a false positive, as it linked to a genuine Amazon domain and was, in fact, sent as compensation for this study. This classification is puzzling and was also observed for several other participants. The misclassification of the research-recruitment message seems more explicable, as the linked URL was a screening survey requesting various pieces of personal information (e.g., name, date of birth, zip code). P19 nonetheless opened this URL despite the warning.

An unreliable warning system is one that users may learn to ignore. Our qualitative impression, based on observations during the study and participant feedback, is that the system was accurate in most cases. Nevertheless, P19's experience illustrates a failure mode: messages requesting personal information from unfamiliar but legitimate senders can resemble phishing

closely enough to be misclassified. Ultimately, we plan to conduct a more systematic evaluation of the classification data in future work.

Requests for changes were diverse, although many focused on elements of the feedback already discussed, such as the text length or the size of the *long* warning. One interesting suggestion that came up in multiple responses was for an in-line warning within or adjacent to the SMS bubble (coded as *request for in-line warning indicator*). For example, P29 stated "Some people may not look at the top of the [s]creen, so i would include a warning sign next to the text as well, or some kind of flag." This recommendation is backed by some evidence in the email domain, such as Petelka et al.'s finding that warnings embedded within a link tool tip resulted in a significantly lower click-through rate than traditional banner warnings [379]. Future work should investigate the effect of different warning positions in the SMS context.

9.5 Discussion

In this section, we summarize our key results and their implications before describing methodological takeaways. We also discuss future work needed to further evaluate SMS scam warnings, especially for potentially vulnerable populations.

9.5.1 Key Takeaways

Our experiment provided the first evidence for warnings' effects on real-world SMS processing behavior, but did not show a behavioral difference between a *generic* warning and a *long* warning with AI-generated text in reducing clicking on phishing links ($\mathcal{H}_1$) or reporting phishing messages ($\mathcal{H}_2$). We also did not observe a significant learning effect from exposure to the GAI warnings, with participants exposed to the *generic* and *long* both performing similarly poorly on a post-study message identification task ($\mathcal{H}_3$). Participants exposed to the *long* warning did express more positive sentiment across several questions, rating it as more helpful both to themselves and to others than participants who saw the *generic* warning ($\mathcal{H}_4$). The *long* warning was perceived as more effective at grabbing participants' attention, but was also considered by some to have too much text.

These results suggest that participants' positive perceptions in Chapter 8 of warnings with AI-generated explanations extend to short-term, real-world deployments, but we were unable to demonstrate a behavioral or learning impact. From this result, we do not conclude that AI-generated explanations are unhelpful. In addition to the issues of statistical power and the well-educated, tech-savvy bias of our sample, we also note that the stimuli were single-message, service-impersonation scams in which the decisive cue is a URL. While URL-based phishing is extremely common, "reply-based" or conversational SMS scams are a growing threat [304]. Indeed, Google markets scam detection in Google Messages against this class of attack, with a prototypical exchange showing the warning appearing after a "wrong number" text transitions into an investment pitch [36]. Warnings against these attacks may depend more heavily on explanation, since no single message is decisively fraudulent and the recipient has typically already replied by the time a warning could fire. Future work should evaluate the effects of GAI warnings against these types of attacks and consider possible modifications to the explanation-generation

architecture. Our system evaluated each message independently, with no memory of prior messages in a thread, so it could not have produced an explanation referencing an extended exchange.

Simulating these attacks is itself a challenge, requiring methodological innovation. SMS phishing simulation, as practiced here and by Rahman et al. [387], is one-shot in the sense that a single message is delivered, and the response to that message is recorded. Studying conversational scams would require repeated interaction, which raises engineering and ethical difficulties beyond those we encountered.

The substantial number of participants who processed messages without opening them highlights a gap left by warnings that are visible only after opening. While deleting a message from the inbox is safe, replying via a notification, at a minimum, alerts the fraudster that the phone number is active. Including some level of warning in the notification or inbox, most simply as a warning symbol, could alert users to the presence of a hazard before they open the message. Inbox warning symbols are present for spam warnings within Google messages, and some participants in Chapter 8's interview study reported that this symbol alone may be sufficient to prevent them from opening a message. Future work should explore the real-world effect of such symbols on SMS processing behavior.

9.5.2 Methodological Reflections

This chapter also makes a significant methodological contribution as the first study to measure real-world interactions with instant messaging warnings and only the second study to simulate SMS phishing [387]. Our experience provides several valuable takeaways for future research.

Some are specific to Android. First, it seems that sideloading APKs is difficult for most users, based on the high attrition between consent and installation. As discussed briefly in Section 9.3, Android implements security warnings at the browser and operating system to prevent users from haphazardly installing APKs outside of the Google Play store. Even with clear, step-by-step instructions, participants may become confused, alarmed, or simply feel that participation is not worth the effort. For apps that cannot be made public, future researchers should consider using the Android internal testing feature. This allows developers to distribute an app to up to 100 users specified by an email list.[12] This manner of installation is still higher friction than a typical app installation, but would significantly lower the barrier to entry for participants.

Our experience also highlights the difficulty many contemporary Android users have using SMS-only messaging applications. Users missed RCS features and, more significantly, experienced bugs during the switch between protocols. Ideally, future studies would use apps with RCS implemented; however, this is easier said than done. While RCS is an open standard, the code for running it on Android is part of the closed-source Google Messages app. Private correspondence with the QUIK maintainer indicates that they are working on supporting the RCS specification; however, it is unclear when this feature will be completed, if at all. From a usability and ecological validity perspective, the best approach would be a study conducted via Google Messages, potentially using large-scale telemetry, as in the seminal study on browser warnings by Akhawe and Felt [7]. Barring direct collaboration with Google, the Android Accessibility-

[12]We considered using this feature, but were unable to do so due to our reliance on an older version of the APK format, which Google would not accept for distribution.

Service API [196] provides extensive capabilities, including reading and overlaying the screen, that could enable the implementation of some limited prototypes.

More broadly, our results also illustrate trade-offs in the design of phishing simulation studies. The high volume of scam messages and the within-subjects design were chosen to maximize power, but also were noticed by some participants. Adopting a full between-subjects design would require recruiting more participants, but would give each participant a consistent experience. Beyond the cost of compensation, our experience with Prolific suggests that recruiting larger sample sizes is not impossible. Sending fewer messages would reduce power and raise concerns about generalizability, given the variability in message effects. Extending the study over a longer period could make them less noticeable to participants, while risking greater attrition and potentially requiring greater compensation. In our experience, once participants installed the app, attrition was low, suggesting the feasibility of a longer study period.

9.5.3 Towards Inclusive SMS Warnings

As observed in Chapter 3, people with visual impairments (PVI), especially those who use screen readers, have a significantly different experience of warnings. Thus, testing mobile warnings with PVI is an important avenue for future work. Hypothesizing about considerations is difficult, as, to our knowledge, no prior study has investigated PVI's text message processing behavior. Still, some insights from desktop email warnings may be transferable. For example, when browsing emails, PVI may use keyboard shortcuts to jump directly to the email body, thereby missing warnings that appear only at the top of a page [562]. Mobile screen readers, namely iOS VoiceOver and Android TalkBack, could enable a similar browsing pattern if a user taps directly on a message body after opening an SMS. A warning with a fully blocking design and/or incorporating auditory alerts would be more salient under this browsing pattern. Analogously, Yu et al. tested a desktop email phishing warning design that incorporated an auditory alert, a fully blocking warning design, and a keyboard shortcut to read the warning content, finding that it was significantly more noticeable than the existing Gmail warning design [562]. The previously discussed suggestion of placing warnings in-line with content is an alternative that would also make the warnings perceivable to those who skip to content directly. Future work should explore both how PVI experience SMS scams and consider the accessibility of prevailing warning designs.

It is also important to consider how people with varying levels of cognitive ability and different levels of educational attainment experience AI warnings. As discussed in Section 9.4.1, our study sample was highly educated and may have had an above-average ability to comprehend the warning text. Future work should more directly evaluate the understandability of AI-generated security advice, rather than by self-report (e.g., via the Smart Cloze procedure used in Redmiles et al.'s evaluation of online security advice [395]) and recruit a sample with greater educational attainment diversity. Ultimately, the participants most likely to benefit from an explanation are precisely those our recruitment methods and installation requirements filtered out. Thus, more inclusive future work is essential to understanding whether these warnings work for the people who need them most.

Chapter 10

Conclusion

Over the previous chapters, this thesis has explored the causes of scam victimization and approaches to mitigating harm, with a focus on at-risk or vulnerable populations. A key observation is the role of contextual factors that are external to the potential victim in explaining harm. Chapter 3 presented a systematic literature review of prior research focused on at-risk users with original results on scams. This review highlighted the role of personal circumstances, such as resource constraints and underserved accessibility needs, and revealed gaps in the current literature.

The following four chapters present empirical studies focused on different populations and highlighting different risk factors that augment scam vulnerability. Chapter 4's investigation of privacy experiences on dating apps among closeted LGBTQ+ users highlights the manner in which marginalization of a population can lead to extortion. Chapter 5 explores minors' use of crypto assets, highlighting profit-centric motives for investing in crypto assets that may be rooted in teens' lack of financial resources. The chapter also highlights how teens' reliance on their parents or other adults increases the risk of harm. Chapter 6 compares teen and adult susceptibility to YouTube giveaway scams, and unexpectedly finds that teens were not more susceptible to these scams. Supporting the role of external factors over intrinsic susceptibility, the teenage participants were more likely to search for one of the fraudulent search terms, meaning the scams were effectively demographically targeted. Chapter 7 analyzed help-seeking among users of Reddit, with the results showing how the service fills gaps in real-world support structures by helping users identify scams and seek remediation advice.

Taken together, the preceding chapters show that scams are a serious problem and that users need help identifying them, especially when the content of a scam exploits their present context. The final two chapters explore a novel intervention: SMS scam warnings with AI-generated, context-specific text. Chapter 8 presents an initial formative study that led to the evaluation of a design in a field experiment in Chapter 9. In a highly educated, young sample, the warning with AI-generated text significantly reduced clicking on links and phone numbers in scam messages relative to receiving no warning, while the generic warning showed a similar but non-significant effect. There was no significant behavioral difference between the conditions, but participants exposed to the GAI text rated it as more useful to themselves and others. More work is needed to explore how these warnings may help against different scam types and among users with a broader range of backgrounds.

The remainder of this chapter considers the methodological approaches of the preceding studies, their implications for US policy, and the questions this work leaves open.

10.1 Methodological Approaches to Understanding Scams

This thesis employed a diverse range of empirical methods, each offering a distinct perspective on digital safety and accompanied by its own limitations. Here, I discuss the chapters in methodologically related groupings, highlighting their pertinent characteristics.

Chapters 4 and 5 approached the topic of scams indirectly, focusing on research questions that were not directly about fraud. Chapter 4 surveyed closeted LGBTQ+ users about their general privacy experiences on dating services rather than about victimization directly, and framed those experiences through the institutional and social privacy concerns advanced by Raynes-Goldie [392] and, in the online dating context, Lutz and Ranzini [305]. Chapter 5 similarly asked how minors access and use crypto assets, as well as what harms they experience while doing so, rather than asking how they are defrauded. Investment scams emerged as one of several outcomes. Broad descriptive work of this kind is valuable for identifying problems that warrant closer attention, but neither chapter can speak to prevalence. Nevertheless, establishing that these harms occur is a precondition for designing studies to more broadly estimate their prevalence or test interventions against them.

Chapter 5 also shares an important methodological feature with Chapter 7. Both draw on public social media posts rather than direct human-subjects research, enabling understanding of populations that are difficult to reach. Chapter 5 focuses on minors who use cryptocurrency, a group that is both very specific and ethically difficult to recruit. Moreover, the posts revealed behaviors that users may be unwilling to share with researchers, such as underage gambling and speculative trading. Chapter 7 provides qualitative evidence on scam victims' support-seeking, helping to pierce the veil of persistent underreporting in more direct studies [1, 18, 231, 391]. Reddit users are a specific subset of the population, so these results should not be generalized too broadly. Posts are also written to a public audience rather than to a researcher, which may encourage embellishment.

The two Reddit studies further illustrate an evolution in ethical practice regarding data presentation. While not releasing the curated dataset or identifying posters by username, Chapter 5 still reproduced posts verbatim. This practice effectively highlights particular pseudonymous posters by enabling identification through reverse searches of the content. Influenced by research by Reagle [393] and Fiesler et al. [178], Chapter 7 adopted the practice of summarizing and generalizing the content of posts, making it much more challenging, if not impossible, to identify specific posters.

The choice to obfuscate posts comes at a cost. It makes it challenging for a reader to determine whether the analysis faithfully reflects the original content and removes the subjects' voices. Researchers working with public posts should weigh this tradeoff, considering the potential harm to human subjects against the loss of precision. In retrospect, this protection was most warranted in Chapter 5. The likelihood that anyone would trace a quoted post back to its author is low, and we are not aware of any case in which this has caused harm. Still, the posters were minors who could not have consented to participate in the research, and some described conduct

that could carry consequences. Where the cost of a safety measure is modest, and the person bearing the risk has no input, the measure is worth implementing.

Chapters 8, 6, and 9 each represent direct evaluations of how users interact with scams, but illustrate tradeoffs in ecological validity and qualitative depth. Chapter 8 adopted a task-based interview approach, which was arguably the least authentic of the three studies. Participants were in the presence of a researcher and were specifically directed to look at the text messages and warnings. However, this approach generated detailed reflections on how participants perceived each warning design and each text message. Chapter 6 adopted a survey-based approach, which still involved role play, but allowed for recruiting a larger sample and for comparing groups directly. The survey was also structured to more closely resemble ordinary use by first presenting search results and then video clips, though participants knew their behavior was being observed and faced no real consequences for engaging with a scam. Chapter 9 offered the greatest ecological validity of the three studies. Participants received simulated scam messages on their own devices, during ordinary use, without knowing which messages were part of the study. This design also constrained what could be measured, thereby limiting qualitative insight to post hoc reflections. This reflects a general challenge in usable security research, where the conditions that permit careful measurement of an intervention are rarely the conditions under which it must work [148].

10.2 Policy Implications

The results of this thesis lead to two main recommendations for future policy development, with a focus on the US legal context. These deal with the regulation of 1) the platforms that distribute scams and 2) the payment providers that enable scammers to profit. These are rooted in the recognition that the best way to stop scammers is to eliminate their ability to reach victims or to profit. The third subsection contrasts these approaches with alternatives under consideration that place the burden on consumers.

10.2.1 Regulating platforms that deliver scam content

First, it is important to use policy levers to motivate service providers, such as social media websites, to combat scams. Internet scams are notoriously difficult to target through law enforcement action, especially because of their transnational character [65]. Much of the burden thus falls on service providers to prevent scammers from using their platforms to contact victims. Service providers already do a great deal to combat scams, including by providing consumer education (as discussed in Section 2.4.2) and automated filtering (as discussed in reference to YouTube in Section 6.4). Still, there are cases where providers' economic interests are not aligned with efforts to prevent fraud. For example, a 2025 Reuters report revealed internal estimates from Meta that it earns $7 billion from "higher risk" advertisers that are suspected to be perpetrating scams [230]. Meta disputes the claimed number as "overly-inclusive," but it underscores how platforms directly profit from fraud. More recently, Meta has announced it is cutting back on human moderation in favor of AI tools, which it claims are more effective at combating digi-

tal safety threats [527]. This imbalance between a growing threat and declining investment in moderation suggests the need for regulatory intervention to protect consumers.

Crafting laws or regulations to incentivize better enforcement is challenging and could raise free speech concerns. For example, multiple versions of a "Kids Online Safety Act" (KOSA) have been introduced in the 119th US Congress that require protections against content that may be harmful to minors, including deceptive content like scams [107]. This could give regulators a lever to penalize platforms for enabling scams targeting children, such as the giveaway scams described in Chapter 6. However, this benefit is speculative as discussions of its application have focused on other potentially harmful content, like violence or sexual material. Moreover, the bill has been criticized by organizations such as the Electronic Frontier Foundation (EFF) and the American Civil Liberties Union (ACLU), especially for the potential that the bill could enable government censorship by allowing the FTC or state attorneys general to sue companies if minors see content on topics like LGBTQ+ issues or depression [21, 332]. Critics argue that this legal risk will lead platforms to remove such content entirely, rather than risk falling afoul of the law. As an intervention specifically against scams, the benefits of this approach appear modest relative to the risks posed by its breadth.

The TAKE IT DOWN Act (S.146) was passed into law in May 2025 and is more narrowly targeted at crimes involving non-consensual intimate imagery (NCII), including sextortion [258, 514]. Among other provisions, it makes it a federal crime to threaten to publish real or AI-generated intimate imagery "for the purpose of intimidation, coercion, extortion, or to create mental distress." The law also mandates a take-down procedure for non-consensual intimate imagery on user-generated content websites and apps. This law therefore directly addresses the sextortion threats observed among closeted LGBTQ+ dating app users in Chapter 4. The take-down procedure provision went into effect on May 19, 2026, with the FTC responsible for enforcement [508]. The law has also been criticized by the EFF for its potential to enable censorship, potential to undermine end-to-end encryption, and failure to account for bad-faith requests [333]. While these concerns are substantive, its focus on a single, well-defined category of content seems to make it less risky than KOSA.

Platforms that facilitate scams may also be targeted under existing law, especially the FTC Act. The 1914 law established the Federal Trade Commission with the mandate to "prevent unfair methods of competition and unfair or deceptive acts or practices in or affecting commerce" and "seek monetary redress and other relief for conduct injurious to consumers." In practice, the FTC's mandate is broadly applied and could include cases against platforms that knowingly allow the dissemination of scams. Indeed, the advocacy group Consumer Reports called on the FTC and state attorneys general to take enforcement action in response to the Reuters reporting cited above on Meta [70]. Meta has subsequently been sued in D.C. court by the Consumer Federation of America (CFA) under the D.C. Consumer Protection Procedures Act (CPPA), which permits public-interest organizations to sue for deceptive practices [353]. Meta is also being sued by 29 states, although these claims generally relate to child safety rather than scams [246]. Ultimately, these piecemeal efforts fail to constitute a coherent regulatory framework and underscore the need for a more comprehensive policy.

10.2.2 Regulating platforms that process scam payments

Second, it is important to target the payment methods used in scams. While many payments are transmitted through means that are challenging to regulate—especially cryptocurrencies or gift cards—as observed in Chapter 7, many payments are made through centralized financial institutions, especially via apps like Zelle. As was also discussed in Chapter 7, banks are generally not liable in the US for payments intentionally made by a user who has been deceived [564], whereas unauthorized transactions made without the victim's involvement are often covered [172, 491]. The result is that banks lack a direct economic incentive to prevent scams, and consumers are left without the potential for remediation.

Although there is no US regulatory framework to hold banks accountable for scam transactions, there have been legal cases targeting financial institutions over scams. The Consumer Financial Protection Bureau (CFPB) sued JPMorgan Chase, Bank of America, Wells Fargo, and Early Warning Services, which operates Zelle, in 2024.[1] The lawsuit argued that Zelle was designed and operated with poor safeguards for scams, with the service performing inadequate identity verification and failing to block repeat offenders [77]. This is philosophically similar, though not legally identical, to fines against firms that fail to adequately protect consumers from data breaches (e.g., Equifax [499]). The legal basis for the suit was the Consumer Financial Protection Act, which authorizes the CFPB to ensure "consumers are protected from unfair, deceptive, or abusive acts and practices" [491]. The lawsuit was subsequently voluntarily dismissed upon new leadership taking over at the agency [96]. The New York State Attorney General subsequently sued Early Warning Services, alleging similar claims under different legal authority [79]. The case survived a motion to dismiss on July 20, 2026, allowing it to move to discovery [80]. While the case has not been adjudicated, this ruling suggests that courts may hold financial service providers liable for failing to adequately prevent fraud.

An instructive contrast to the lack of US payment reimbursements for scams is offered by the UK. In October 2024, the Payment Systems Regulator introduced mandatory reimbursement for Authorized Push Payment (APP) fraud, defined as "when someone is tricked into sending money to a fraudster posing as a genuine payee" [232, 372, 386]. This replaced a previous voluntary regulatory scheme. Reimbursement is limited to transactions between two UK-based bank accounts that were made via Faster Payments, a real-time payment service administered by Pay.UK, or the Clearing House Automated Payment System, a same-day bank transfer system administered by the Bank of England. Covered scams include many common forms of fraud, such as government and family impersonation scams. The financial institutions involved in sending and receiving the fraudulent payment are required to split the cost 50:50, placing a direct incentive on sending banks to identify potential scam victims and receiving banks to identify scammers.

Notably, the regulation includes provisions for "vulnerable consumers." Under UK law, a vulnerable consumer is defined as someone who "is especially susceptible to harm, particularly when a firm is not acting with appropriate levels of care" as a result of "their personal circumstances" [179]. Consumers deemed vulnerable are not subject to the £100 out-of-pocket minimum set by law or to a provision that excludes reimbursement for grossly negligent consumers (e.g., those who ignore warnings). This approach is strongly aligned with the thesis's analysis of how contextual factors beyond one's control may increase one's risk of harm.

[1] The bank plaintiffs co-own Early Warning Services.

A 2026 review commissioned by the PSR and conducted by Frontier Economics suggests this policy has been moderately successful [189]. Specifically, the report attributes a 21% reduction in APP fraud to the policy and an increase in consumer reimbursement rates for APP fraud from 54% before the policy to 65% after. The study highlights increased costs to payment service providers due to reimbursement, anti-fraud technology, and administration costs, but concludes that the policy is a net economic benefit. The report also highlights that service providers vary in their application of the regulations' requirements, and many scams fall outside of the category of APP fraud. Still, this policy seems to represent a successful shifting of economic incentives onto banks to prevent scams. US policymakers should consider extending existing regulations, especially 12 CFR Part 1005 (Regulation E) [492], to cover scam fund transfers, following the UK model for distributing liability and protecting vulnerable consumers.

10.2.3 Other anti-scam policies

There are other policy-making efforts in the US that primarily focus on helping users recognize or remediate scams. The proposed Romance Scam Prevention Act (S.841) would require dating services to notify users who have previously spoken with a romance scammer who is subsequently banned, with the notification to include a warning to not send money or financial information to the scammer [516]. Failure to comply would be punished as an unfair or deceptive trade practice under the FTC Act. The ReportScams.gov Act (S.4782) seeks to establish a uniform federal policy for combating scams by establishing a Scams Steering Committee that would be responsible for setting a four-year target for reducing scam incidents and costs [515]. The law would also establish a single centralized digital hub for scam awareness education, prevention advice, and reporting tools.

Both bills could address aspects of the scam crisis. The Romance Scam Prevention Act explicitly requires warning users about known romance scammers, which may be helpful to some potential victims. Its scope is narrow, however, and it seems less likely to help victims who are already deeply deceived, such as those described in Chapter 7, whose family members sought help on Reddit. The ReportScams.gov Act has the potential to develop a stronger anti-scam policy by establishing a committee representing various stakeholders from different parts of the federal government. The bill's explicit provisions focus on providing educational resources and centralizing reporting, which places the responsibility on victims. Both of these bills fail to address the systems that enable scams to succeed and are less promising than regulations directed at service providers.

10.3 Directions for Future Research

I close this thesis with a discussion of future research directions following from its findings and limitations. Specifically, I discuss the investigation of anti-scam interventions for at-risk users in ecologically valid settings, further investigation of demographic targeting, and the potential to leverage social media data to conduct large-scale studies across diverse populations.

10.3.1 Assessing anti-scam interventions

As observed in Chapter 3, there is a gap between the extent to which interventions are recommended and the extent to which they are evaluated. Chapter 9 contributes towards filling this gap, but also speaks to the inherent challenges in conducting ecologically valid evaluations with special populations. The study drew on general-population samples, but the participation requirements yielded a sample that was more technically adept. While steps can certainly be taken to improve the accessibility of our methods, for example, by making the installation process easier, deployments of this kind will exclude those who are uncomfortable installing and adapting to a new interface within a short period of time. Reliance on screen readers, for example, may introduce significant barriers to accessing unfamiliar interfaces, making participation without a longer onboarding procedure impossible [523]. Testing interventions at a macro scale using data from service providers is preferable, but infeasible for most researchers. Moreover, it limits the type of data that may be collected. This leaves smaller-scale laboratory work as the most reasonable option for testing interventions within many populations.

These types of studies will never achieve the same level of realism as scam simulations, but we can build the methodological infrastructure to support greater validity. For example, interactive mockups of email clients have become common in email phishing research (e.g., [274, 569]), and Reithmaier et al.'s PhishyMailbox provides a framework for constructing them [399]. Work on instant message phishing, by contrast, has generally relied on screenshots of individual messages [470, 481]. Chapter 8 suggests this matters, as presenting tasks within a mockup of a messaging app surfaced how users begin forming judgments from the inbox, before a message is ever opened. This dynamic was confirmed by Chapter 9, as roughly a third of simulated messages were never opened at all. Building these mockups required substantial manual effort from a researcher with UX experience, since each view had to be constructed by hand. The lack of more general-purpose tooling in SMS either requires researchers to replicate this labor or forces reliance on less realistic designs.

The community should invest in developing platforms to enable more realistic role-play for SMS interfaces and other contexts where scams occur, such as social media or ecommerce websites. Crucially, these must be tested for fidelity, not just visually but also by users via accessibility tools like screen readers.

10.3.2 Investigating scammer targeting

Another gap identified in Chapter 3 is the lack of research on how scammers target specific populations. Chapter 6 illustrates both the phenomenon and the limits of current evidence. Teenage participants were more likely to have encountered the scam search terms before the study, indicating that these scams reached teenagers at higher rates than adults. This result cannot tell us whether scammers deliberately selected content appealing to younger users or whether this is an incidental correlation. The study by Logie and Das provides compelling evidence that fraudsters deliberately seek out contact information of people perceived as vulnerable, especially older adults [300]. The extent of these practices and their broader impact remains underappreciated in research focused on at-risk users. More work is necessary to understand how victims are targeted.

In addition to approaches focused on understanding scammers' perspectives, discussed in Section 3.5, it may be possible to assess targeting through more direct measurement of data flows. For example, a recent study from Vekaria et al. developed 210 synthetic profiles with standardized demographic information [521]. They fed these profiles into 105 health-insurance lead generation websites[2] in two rounds and measured both how this data was transmitted and the nature of resulting communications. The results highlight the aggressive data-sharing practices of these websites and their apparent role in causing spam. The researchers received over 8,000 telemarketing phone calls, 600 text messages, and 200 emails across the profiles. Notably, they do not investigate the content of the resulting spam, nor do they consider the potential role of demographic variables in shaping it. A similar design, which intentionally manipulates demographic variables, could provide strong evidence on the nature of demographic targeting facilitated via lead generation or other participants in the data economy.

From the other side of the pipeline, testing the effect of privacy-protecting practices on the frequency and content of scams received could help determine whether these tools provide meaningful protection. As discussed in Section 3.5, privacy-enhancing technologies, such as commercial data-removal services, market themselves as capable of suppressing scams. An evaluation by He et al., which recruited 71 participants to use four different services, found that these services are both inaccurate in correctly identifying customers' PII and ineffective at removing data they do correctly identify [218]. The more complex claim, that these opt-outs protect users from harm, has not been evaluated. Vekaria et al. tested the effects of phone-based opt-outs (adding numbers to the Do Not Call registry and calling provider-designated numbers) and email-based opt-outs (sending opt-out requests to listed email addresses and clicking unsubscribe links). While both approaches reduced communications over their respective channels, neither cut it off entirely, illustrating pervasive non-compliance. Future work should extend this same analysis to broader parts of the ecosystem and examine the specific impact on fraud [521].

Lead generation services, data brokers, and other components of the under-regulated[3] data economy are only one possible supply of targeting information. Credentials and other personal data leaked through breaches circulate widely on underground forums and Telegram groups [413, 478]. While these are most straightforwardly used for identity theft or account compromise, they may also be leveraged to identify potential victims. Chapter 5 discusses an example where a breached password was used as "proof" that a scammer had access to intimate imagery featuring the user. This channel is less easily regulated, but may also be measured. For example, Peng et al. submitted "honey credentials" to live phishing sites and monitored subsequent access [377]. Analogous designs could be adopted, although the criminal context presents more logistical and ethical challenges than straightforwardly submitting synthetic profiles to lead-generation websites.

[2]Lead generation websites collect users' personal information and sell it on to marketers seeking to sell various services.

[3]As illustrated by the results of Vekaria et al. and He et al., many of these players act without regard to privacy regulations.

10.3.3 Collecting large-scale scams data on social media

A final important gap highlighted in Chapter 3 is the lack of large-scale data on scam experiences across many groups. Chapters 5 and 7 demonstrate one possible approach to solving this issue by using publicly available, qualitative data to provide insight into invisible safety experiences. The scale of these studies was limited by the nature of manual analysis. Subsequent methodological advances in applying LLMs to Reddit data, especially by Wei et al. [539] and Chanenson et al. [94], enable studies at a greater scale.

In the context of this thesis, this approach seems most useful to characterize the scams discussed within specific communities, using subreddits organized around an identity or circumstance as a proxy for the population of interest. This approach was used in Chapter 5 to analyze the experience of teenagers with cryptocurrency, and the results were later corroborated by Klucinec et al.'s survey of emerging adults who had used cryptocurrency as minors [261]. While keyword searches provided an acceptable level of accuracy for the topic of crypto assets, the broad usage of terms like "scam" makes keyword sampling infeasible [351]. LLM-based sampling could allow for a broader picture of the types of scams that teenagers and other specific groups discuss.

This approach is limited by the fact that many identity groups lack relevant communities, that posting within these groups is not inherently restricted to group members, and that posters may differ from the broader population. Saying that one scam is discussed more often than another is not a reliable indicator of prevalence, although it does show that at least some users in these groups experience this harm. Moreover, discussion of fraud is likely to be rare in identity-focused communities, limiting the availability of evidence.

A complementary approach would be to identify identity disclosure in scam-focused communities, such as `/r/Scams`. While not discussed explicitly in Chapter 7, we observed that some posters disclosed characteristics such as age, gender, or occupational status, as in the case of a member of the military who requested help with a threat-based scam. These disclosures are likely to be biased towards cases where the poster feels their identity is relevant to the situation, skewing the resulting samples, but such cases may be more prevalent than discussions of scams in identity-focused communities.

Both of these approaches are contingent on demonstrating LLMs' accuracy in screening posts for discussion of scams or identity disclosure. Each also comes with significant ethical issues, as explored in Section 10.1. Whereas scam-focused communities have privacy norms that are amenable to research based on their data (see Section 7.1.3), this may not be the case for smaller, identity-oriented communities [177]. There is also a dual-use risk: a classifier designed to detect identity groups could easily be used by adversaries to target vulnerable users. Researchers must carefully consider the context they are investigating and, if appropriate, abstain from conducting the research entirely where potential harms outweigh benefits.

Ultimately, these three directions represent ongoing methodological projects rather than questions that can be answered. Like many security threats, scams evolve in response to the defenses raised against them and with changes in technology. Our methods for studying them, and the interventions we build, must evolve as well.

Appendix A

A Systematic Review of Internet Scam Research with At-Risk Users

A.1 Codebooks

A.1.1 Population

This codebook has been trimmed to include only codes discussed in the chapter. *A priori*, we included all codes from Warford et al. [535].

1. **Demographics:** The paper focuses on analyzing age, gender, or other demographic factors without focusing on specific at-risk groups.
2. **Child:** Minors who are 12 or younger.
3. **Teenager:** Minors between the age of 13 and 17.
4. **Minors:** Under 18, including both teens and children.
5. **Emerging adults:** Young adults who are generally between 18 and 29 years old (inclusive).
6. **Older adults:** People who are generally 65 years of age or older.
7. **LGBTQ+ People:** People who identify as lesbian, gay, bisexual, transgender, queer, or another gender/sexual minority.
8. **Women:** Women.
9. **Men:** Men.
10. **Marginalized racial/ethnic group:** People who experience discrimination, exclusion, or systemic disadvantage based on their race or ethnicity.
11. **People with an illness:** People living with an ongoing physical or mental health condition that requires medical treatment or management.
12. **People with cognitive impairments:** People who have problems with their ability to think, learn, remember, use judgment, and make decisions.

13. **People with visual impairment:** People who are blind or low-vision, i.e., they have vision loss that is not correctable with glasses and impairs their ability to perform everyday tasks.
14. **Other disabilities:** People who experience a disability not specified elsewhere in this codebook or who experience multiple disabilities simultaneously (e.g., low vision and hearing impairment).
15. **Country:** Used as a modifier to indicate the country context relevant to the population being studied (e.g., *Demographics X US* corresponds to a paper that investigates the demographics of people in the US)
16. **Undocumented immigrants:** Non-citizens without valid immigration documentation (e.g., visa, residency permit, or green card).
17. **Refugee:** A person who is forced to leave their home country due to war, persecution, natural disaster, or another comparable cause.
18. **Migrant:** Migrants generally, including but not limited to refugees or undocumented immigrants.
19. **Low-SES:** People with low socioeconomic status.
20. **Other at-risk group:** Another group that is at elevated risk but is not captured by the categories above.

A.1.2 Scam Type

These codes are a subset of categories selected *a priori* from the Stanford Center on Longevity and FINRA Investor Education Foundation taxonomy of fraud [39].

1. **Not focused on particular scams:** The paper discusses scams, fraud, or scam prevention broadly, but the methods and results are not focused on a specific scam type.
2. **Consumer Investment Fraud:** Fraud schemes in which someone knowingly misleads an investor using false information to obtain money.
3. **Worthless or non-existent products:** Schemes involving products, goods, housing, experiences, or other items that are falsely advertised, worth substantially less than claimed, or do not exist.
4. **Worthless, unnecessary, or non-existent services:** Schemes in which victims agree to purchase a service that is misrepresented, unnecessary, overpriced, or never provided.
5. **Phishing:** Fraudulent attempts to obtain sensitive information by impersonating a trustworthy entity.
6. **Tech support scam:** Scams in which fraudsters pose as technical support representatives, often claiming to be affiliated with well-known technology companies. They falsely claim that the victim's device has malware, security problems, or other technical issues in order to obtain payment, install malicious software, or gain remote access to the victim's device.
7. **Employment Fraud:** Fraud schemes where the expected benefit is employment, freelance work, business opportunities, or training for profitable ventures. Fraudsters promise

attractive employment opportunities but require upfront fees, purchases of materials, or other payments.

8. **Prize and Grant Fraud:** Schemes in which victims are falsely led to believe they have won or are entitled to receive a prize, lottery payout, inheritance, grant, sweepstakes winnings, or other financial windfall.
9. **Phantom Debt Collection Fraud:** Scams involving fake debt collectors who attempt to convince or threaten victims into paying debts they do not owe.
10. **Charity Fraud:** Schemes in which fraudsters collect money while pretending to represent a legitimate charitable organization or cause. Victims expect their money to support charitable activities, but the organization is fraudulent or the funds are misappropriated.
11. **Relationship and Trust Fraud:** Scams in which the fraudster exploits an actual or fabricated personal relationship with the victim. The victim's expected benefit is social connection, friendship, family assistance, or romantic involvement rather than a product or service.
12. **Threat-based scams:** Scams in which fraudsters falsely threaten harm, legal consequences, social consequences, or physical violence unless the victim provides money or other resources.
13. **Other scams:** Scam types not adequately captured by any of the categories above.
14. **No results focused on scams:** The article does not include results related to internet-mediated scams or scam victimization and should be considered for removal from the dataset.

A.1.3 Risk Factors

This codebook has been trimmed to include only codes discussed in the chapter. *A priori*, we included all codes from Warford et al. [535]. We also added themes developed throughout the analysis.

1. **Societal Factors**:
   (a) **Legal or political:** Government policies, political conditions, legal systems, or regulatory environments contribute to heightened scam victimization risk for a population.
   (b) **Marginalization:** Pervasive negative treatment, discrimination, exclusion, or structural disadvantage due to a person's identity attributes or life experiences contributes to elevated scam victimization risk.
   (c) **Social norms:** Informal societal rules, expectations, or cultural practices contribute to scam victimization risk by restricting options, shaping behavior, or creating mismatches between individuals' lived experiences and the assumptions embedded in technologies, institutions, or services.
2. **Relationship Factors**:

(a) **Relationship with the attacker:** A personal, familial, romantic, caregiving, professional, or otherwise close relationship with the attacker contributes to scam victimization risk. Such relationships may provide attackers with greater trust, authority, access to information, or opportunities for manipulation.

(b) **Reliance on a third party:** Dependence on another person or organization for care, assistance, decision-making, financial management, communication, or essential daily tasks contributes to scam victimization risk.

(c) **Access to other at-risk users:** A person's access to another at-risk individual or population increases their likelihood of being targeted as an intermediary, stepping stone, or pathway to the primary target.

3. **Personal Circumstances**:

(a) **Resource or time constrained:** Limited access to financial resources, technology, connectivity, devices, support services, education, or available time contributes to elevated scam victimization risk and reduces the ability to respond to threats.

(b) **Underserved accessibility needs:** Accessibility needs that are inadequately supported by current technologies, services, or institutions contribute to scam victimization risk. This includes needs related to disability, neurodiversity, language barriers, literacy, or developmental maturity.

(c) **Loneliness or isolation:** A person's subjective sense of loneliness, social isolation, lack of social support, or limited social connections contributes to scam victimization risk. This may include seeking companionship, emotional support, or social relationships online.

4. **Other Risk Factors**:

(a) **Self-reported or measured behavior:** The paper reports that specific behaviors, habits, routines, technology-use practices, or decision-making patterns are associated with scam victimization risk.

(b) **Personality traits or dispositional traits:** The paper reports that personality characteristics, psychological traits, cognitive styles, attitudes, beliefs, or other dispositional factors are associated with scam victimization risk.

(c) **Other:** Another factor that augments or amplifies scam victimization risk and is not adequately captured by the categories above.

5. **Targeting:** Scam victimization risk arises because fraudsters deliberately target specific demographic, social, cultural, or identity-based groups.

6. **N/A:** The paper does not discuss any contextual, behavioral, dispositional, or structural factors that augment or amplify the risk of scam victimization.

A.1.4 Methodological Characteristics

The codes in this section are derived from Distler et al.'s review of methods in usable privacy and security [148].

STUDY METHODS:

1. **Survey:** a written questionnaire form without a conversation-style interaction between the researcher and the research participant.
2. **Interview:** Use of a conversation between researcher and research participant to obtain information. Does not include oral debriefs that may occur after an experiment.
3. **Experiment:** Procedures where experimental conditions were manipulated, and the effect of this manipulation was measured.
4. **Focus group, co-design, or other group activity:** An interview with a group of participants that may involve an activity or co-creation activity.
5. **Analysis of existing datasets:** Analysis of data (e.g., survey data) that existed prior to the study.
6. **Log analysis:** Analysis of data from website or other software logs to study user behavior.
7. **Diary study:** A study where participants record their experiences or thoughts on a regular basis over time.
8. **Observation study:** A study of participants' real-world behavior observed in public.
9. **Other user study method:** Research uses another type of user study method not specified above.
10. **Unclear:** Method is not clearly specified.
11. **Not a user study:** Paper does not have a user study component.

STUDY OBJECTIVE:

1. **Descriptive**: "Descriptive research provides a snapshot or summary of participants and their opinions or behavior with respect to a particular context or setting."
2. **Relational**: "Relational research is designed to discover relationships among variables... for example the impact of certain demographics or past experiences on behavior."
3. **Experimental**: "Experimental research partitions participants into equivalent groups and measures the influence of different experimental manipulations applied to each group."

PROTOTYPE USED: *Yes / No / Not a user study:* Is the study a user study that involves a prototype of any fidelity? A prototype is a technical solution, such as a warning message, an icon, or an interface.

EDUCATIONAL INTERVENTION: *Yes / No / Not a user study:* Did the researchers attempt to educate research participants on privacy and security-related topics?

OTHER INTERVENTION: *Yes / No / NA*: Is the paper **not a user study** but presents a prototype, educational intervention, or other intervention?

SAMPLE SIZE: *# / Not a user study:* Specify the number of participants or data log entries that were analyzed

Appendix B

Exploring the Privacy Experiences of Closeted Users of Online Dating Services in the US

B.1 Recruitment Materials



(a) The first Grindr advertisement.

(b) The second Grindr advertisement.

(c) The physical poster/digital flyer sent to LGBTQ+ organizations.

Figure B.1: Advertisements used to recruit study participants.

B.2 Survey Instrument

The section headings were not visible to participants. Italicized text is used to indicate survey flow and response type. Answer choices are shown in bullets below each question. Answer

responses with the text "please specify" or "please describe" included a free response box for participants' to explain their answer.

B.2.1 Demographics

Q1: How do you describe your gender identity?

- Male
- Female
- Non-binary
- Agender
- Genderqueer
- Prefer to self-describe (please specify)
- Prefer to not respond

Q2: How do you describe your sexual orientation?

- Homosexual
- Asexual
- Bisexual
- Pansexual
- Queer
- Prefer to self-describe (please specify)
- Prefer to not respond

Q3: Please select the age group you are in:

- 18-25
- 26-33
- 34-41
- 42-49
- 50-57
- Above 57
- Prefer to not respond

Q4: How do you describe your race or ethnic identity? (You may select more than one option.) *(participants can select multiple options)*

- White
- Black
- American Indian or Alaska Native
- Asian
- Native Hawaiian or Pacific Islander
- Prefer to self-describe (please specify)

- Prefer to not respond

Q5: What is the highest level of education you have completed?

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

Q6: What is your annual household income?

- Less than $25,000
- $25,000 - $50,000
- $50,000 - $100,000
- $100,000 - $200,000
- More than $200,000
- Prefer not to respond

B.2.2 Outness Inventory Questionnaire

Q7: For each of the following statements, please select an option corresponding to how "out" you are with the people specified below. If you are more out with some people in a particular group than others, please provide the answer that applies to how out you are with the majority of people in that group.

- Mother
- Father
- Sibling(s)
- Extended family and relatives
- My old straight friends
- My new straight friends
- My work peers
- My work supervisors

For each statement, participants had the following options:

1. *Person definitely does not know about my sexual orientation status*
2. *Person might know about my sexual orientation status, but it is never talked about*
3. *Person probably knows about my sexual orientation status, but it is never talked about*

4. *Person probably knows about my sexual orientation status, but it is rarely talked about*
5. *Person definitely knows about my sexual orientation status, but it is rarely talked about*
6. *Person definitely knows about my sexual orientation status, and it is sometimes talked about*
7. *Person definitely knows about my sexual orientation status, and it is openly talked about*
8. *N/A*

B.2.3 Characterizing Use

Q8: Which of the following dating applications or online dating services have you used within the last 3 months (select all which apply)? *(participants can select multiple options)*

- Tinder
- Grindr
- Lex
- Bumble
- Hinge
- OkCupid
- Match.com
- eHarmony
- Other dating app/service not listed (please specify)

Q9: When was the first time you used a dating app or online dating service?

- In the last week
- In the last month
- In the last 3 months
- In the last 6 months
- In the last year
- In the last 2 years
- In the last 3 years
- More than 3 years ago

Q10: When was the last time you used a dating app or online dating service?

- Today
- In the last week
- In the two weeks
- In the last month
- In the last two months
- In the last three months

Q11: How often have you used dating apps or online dating services in the past 3 months?

- A few times a day
- Once a day
- A few times a week
- Once a week
- A few times a month
- Once a month
- Less than once a month

Q12: What is your primary purpose for using a dating app or online dating service?

- To meet new friends or acquaintances
- To have casual sex
- To find a romantic partner
- Other (please specify)
- No specific purpose/Unsure

B.2.4 Privacy Attitudes and Behaviors

Q13: What information do you include on your dating profile(s). Please select all that apply *(participants can select multiple options)*

- Full name
- First name only
- Nickname/Initials
- Age
- Sexual orientation
- Photos of self
- Gender
- Physical characteristics (e.g. height, weight, etc.)
- General location (e.g. the state or city you live in)
- Exact location (i.e. GPS or street address)
- Other (please specify)

Q14: Do any of the dating app(s)/service(s) you use require location permissions?

- Yes
- No
- Other (please specify)

Q15: Do you do anything to protect your privacy when using dating apps/online dating services?

- Yes
- No

Q16: How do you protect your privacy when using dating apps/online dating services? *(Free response field) Question displayed if answer to Q16 is "Yes"*

Q17: For each of the following statements, please select the option corresponding to how much you agree or disagree with the statement. *(Options for each statement: Strongly disagree, Disagree, Neither agree nor disagree, Agree, and Strongly agree)*

1. I am truthful about my sexual orientation on my dating app profile.
2. I trust that the information I share about myself on dating apps/online dating services will not be shared outside the app or service.
3. I don't want people who I know in real life to see my profile on dating apps/online dating services.
4. I am concerned that people who I haven't been "out" to in real life will see my sexual orientation on dating apps/online dating services.
5. I find the paid features on dating apps/online dating services helpful.
6. Paid features on dating apps/online dating services allow me to better protect my privacy.

Q18: For each of the following statements, please select the option corresponding to how much you agree or disagree with the statement. *(Options for each statement: Strongly disagree, Disagree, Neither agree nor disagree, Agree, and Strongly agree)*

1. I feel comfortable sharing photos of myself with others in direct messages on dating apps/online dating services.
2. I feel comfortable with someone sharing images of my dating profile to their friends or family.
3. I feel comfortable with someone sharing images of conversations I had on dating apps/online dating services.
4. I am fine with non-dating websites purchasing anonymous information from the dating apps/online dating services I use.
5. I am fine with my dating profile being used to personalize advertisements.

Q19: For each of the following statements, please select the option corresponding to how much you agree or disagree with the statement. *(Options for each statement: Strongly disagree, Disagree, Neither agree nor disagree, Agree, and Strongly agree)*

1. I limit the amount of information I put on my dating app profiles.
2. I have displayed/shared my social media accounts to my profile on dating apps/online dating services.
3. I lie about information on my dating profile(s) to protect my privacy.
4. I have taken screenshots of images I saw or messages people sent me on dating apps/online dating services.
5. I have shown other people's dating profiles to my friends or family.

Q20: For each of the following statements, please select the option corresponding to how much you agree or disagree with the statement. *(Options for each statement: Strongly disagree, Disagree, Neither agree nor disagree, Agree, and Strongly agree)*

- I have shown conversations I had on dating apps/online dating services to other people.
- Before I go on a date with someone I met on dating apps/online dating services, I usually look them up on the internet.
- I have paid to access additional features on dating apps/online dating services.
- People have sent me unsolicited intimate photos on dating apps/online dating services.
- I have found profiles of people who I know in real life on dating apps/online dating services.

Q21: Now that you have thought more about privacy issues on dating apps/online dating services, does anything come to mind that you have done to protect your privacy? *(Free response field)*

B.2.5 Improving User Experience

Q22: Has your privacy been violated while using a dating app/online dating service? If so, what happened? *(Free response field)*

Q23: Has someone threatened you on a dating app/service? If so, please describe what they threatened and what happened as a result. *(Free response field)*

Q24: Which features(s) do you think could be changed or added to better protect your privacy on dating apps/online dating services? Please reference the dating app or the online dating service you use in your answer. *(Free response field)*

B.2.6 Conclusion

Q25:Thank you for completing our survey. You will have the opportunity to enter the giveaway for an $125 Amazon gift card on the final screen, once you click "Submit." If you are interested in participating in a followup interview, for which you will be compensated $15 for 30 minutes of time, please select the appropriate checkbox below and enter your email address. Otherwise, select "No, I am not interested in participating in a follow-up interview" and proceed to the final screen.

- Yes, I am interested in participating in a follow-up interview, either by email exchange or Zoom, for a $15 Amazon gift card *(Free response field)*
- No, I am not interested in participating in a follow-up interview

B.3 Code Books

The following subsections list the codes derived from an inductive analysis of participants' free-response answers. Each code is accompanied by a definition, the frequency of use, and an example quote. For the privacy-protecting behavior codes, the frequency is the sum of unprimed and

primed instances of the code. If a participant gave the same answer both primed and unprimed, this was only counted as a single instance of the code.

B.3.1 Privacy Protecting Behaviors (Q16, Q21)

- **No response/irrelevant**: Participant does not have an answer or provides an answer that is irrelevant to the question. Almost exclusively applies to Q22 49 instances. (*"Nothing more than I initially stated."*)
- **No full name**: Participant indicates that they do not include their full name in their profile. This includes use of first name or initials only. 45 instances. (*"Not using my real name..."*)
- **No exact location** : Participant indicates that they do not share their exact location or address on the dating platform. 27 instances. (*"I do not give my exact location"*)
- **General information restriction**: Participant indicates that they limit the information they share on dating profile, but do not provide specifics. 25 instances. (*"...don't disclose too much personal info."*)
- **Other**: Participant protects their privacy in a way which is not specified by another code. 23 instances. (*"only meet people once we verify each other on social media."*)
- **No face photos**: Participant indicates that they do not include their face in pictures they publicly share. 17 instances. (*"Don't show full face photos"*)
- **Screen people**: Participant indicates that they interact with individuals, ask for information or otherwise "screen" people prior to sharing information. 16 instances. (*"...Asking for pictures before feeling comfortable to share..."*)
- **Restrict photos**: Participant indicates that they restrict or edit the type of photos they share on their profile in a non-specified way. Also includes specific photo restrictions that are not encompassed by the other codes. 14 instances. (*"No face photos with genitals..."*)
- **No social media**: Participant indicates that they do not link non-dating social media services to their dating profile. 12 instances. (*"...I don't link any of my social media platforms to the dating app"*)
- **No photos**: Participant indicates that they do not share any photos on their dating profile. 8 instances. (*"I not post my pics..."*)
- **No occupation/educational institution**: Participant indicates that they dod not share their occupation or educational institution on their dating profile. 7 instances. (*"Don't reveal my occupation"*)
- **Throwaway accounts**: Participant indicates that they use "throwaway" email addresses, phone numbers, social media accounts, etc. In other words, they create accounts disconnected from their identity to interact on dating applications. 6 instances. (*"...use a basic email that doesn't link directly to me"*)
- **No contact information** : Participant indicates that they do not share contact information such as phone number or email address. 4 instances. (*"By not giving out my.... phone number"*)

- **Purchase premium features**: Participant states that they paid for the application or purchased additional features in order to protect their privacy. 4 instances. (*"the upgraded tinder version allows me to put my location to anywhere so people in my hometown cannot see me"*)
- **VPN**: Participant indicates that they use a VPN when using a dating application. 3 instances. (*"VPN"*)
- **Avoid using public photos**: Participant indicates that they do not share pictures that they have publicly shared elsewhere online. 2 instances. (*"On one app I do not have a publicly available photo"*)
- **No tattoos**: Participant indicates that they avoid sharing photos that show their tattoos. 2 instances. (*"...I have tattoos and have photoshopped all of them out of my pictures..."*)

B.3.2 Privacy Violations (Q22)

- **No**: Participant indicates that their privacy has not been violated while using a dating service or app. 74 instances. (*"n/a"*)
- **Unauthorized sharing of use of dating application/service**: Participant indicates that someone shared that they use a dating application with others. This is specifically when sharing is done without criminal intent. 7 instances. (*"Other than talking about me to their friends, no"*)
- **Unauthorized sharing of photos**: Participant indicates that someone shared photos of the participant without their consent. May be paired with "unatuthorized screenshot/downloading of photos" but also includes the sharing of photos which were voluntarily sent. 7 instances. (*"My photos have been shared with others..."*)
- **Unauthorized screenshot/downloading of photos**: Participant indicates that someone took screenshots or saved their photos/videos/profile/messages without permission. 5 instances. (*"Yes. Someone sharing photos and conversations to a person who then Messaged me telling me what had happened"*)
- **Extortion**: Participant indicates that someone attempted to extort them. They may specify that they were extorted with the threat of outing them or sharing explicit photos. 5 instances. (*"Attempted blackmail with photos of me, threatening to send to friends/family. I don't care very much though I just block them."*)
- **Other**: Participant describes a situation not covered by other codes. 4 instances. (*"It wasn't anyone's fault but my own but I screenshot my profile to send to my friend and the screenshot got put into a shared Google photos album I have with my mom because of facial recognition"*)
- **Harassment**: Participant describe digital or physical harassment. 4 instances. (*"stalker tried to message me across many accounts and somehow figured out the general area i live in, continually objectifying me and pleading me for favors until i threatened to contact law enforcement"*)

- **Context collapse**: Participant encountered someone who does not know about their LGBTQ+ identity while using a dating application. 4 instances. (*"A coworker approached me after finding one of my dating profiles and publicly asked questions about it."*)
- **Threatened with outing**: Participant incates that someone threatened to out them, but does not mention extortion or a financial motive. 3 instances. (*"Threatened to be outed/doxxed at work"*)
- **Unspecified**: Participant states that their privacy has been violated but does not explain what happened. 2 instances. (*"Yes and it caused significant stress in my personal life"*)
- **Impersonation**: Participant indicates that someone reused their photos or otherwise tried to impersonate the participant. 2 instances. (*"People represented my intimate pics as their own"*)

B.3.3 Features to Improve Privacy (Q24)

- **No answer**: Participant does not have a suggestion for another feature or says something incomprehensible. 33 instances. (*"Nothing I can think of right now."*)
- **Profile visibility controls**: Participant would like the ability to more precisely control who can view their profile. 14 instances. (*"For all services, only users with an active account can see your profile, so they can't be seen by the public"*)
- **Disable/limit screenshots or saving photos**: Participant would like the ability to prevent or limit the ability of others to take screenshots or save photos. 13 instances. (*"\"No screenshots.\""*)
- **Other relevant**: Participant makes suggestion that does not fit with other codes, but does actually answer the question. 11 instances. (*"A safety feature would be nice."*)
- **Make paid features free**: Participant would like some or all of the paid features to be available for free. This code is sometimes accompanied by another code specifying the feature they would like to be made free. 11 instances. (*"Non-paid privacy options"*)
- **Other irrelevant**: Participant makes suggestion that does not fit with other codes and does not actually answer the question. 8 instances. (*"Tinder could add better features related to connecting people better"*)
- **Disable/limit exact location**: Participant would like the ability to disable or limit the ability to view exact location. 8 instances. (*"Maybe it can support the feature of disabling GPS"*)
- **User verification**: participant would like the dating service/application to verify the identity of users in someway. 7 instances. (*"Tinder requires a verification process, whereas Grindr does not. Grindr should take note"*)
- **Describes feature they like**: Participant describes a feature they like or used to protect their privacy. Always paired with another code corresponding to the feature type. 6 instances. (*"Actually, I turned off the tinder search when I was back at my college, so people wouldn't find me."*)

- **Improve blocking**: Participant specifies that they would like the platform to improve the ability to block people. 5 instances. (*"A block feature that actually works"*)
- **Disable/limit the sales of information**: Participant would like the dating service to limit or entirely stop sharing data with other companies. 4 instances. (*"Not have info sold to companies as a means to show personal ads"*)
- **Improve other existing feature**: Participant specifies an existing feature they would like to improve, other than blocking. 3 instances. (*"Better password locking for Scruff.'*)
- **Improved privacy education**: Participant would like the dating service/application to improve the app's privacy education or otherwise make privacy features more visible to users. 3 instances. (*"...maybe a quick tutorial before you add photos to remind you to be careful of adding photos with identifiable info or ones that could be easily imaged searched on google"*)
- **Unsend messages**: Participant would like the ability to delete messages after they have been sent (*"Allow me to unsend messages..."*)
- **Better moderation**: Participant would like to have better moderation of users on the platform. 2 instances. (*"They need to definitely research when profiles are reported for various reasons"*)

B.4 Outness Inventory Correlations

Table B.1: Spearman's rank order correlation (ρ) between the average OI score and each of the Likert scale questions. * indicates a weak correlation (absolute values between 0.2 and 0.4), ** indicates a moderate correlation (absolute values between 0.4 and 0.7)

	ρ	% Agree	% Disagree	p-value
Q17.1	.260*	82.5%	10.5%	.005
Q17.2	.029	40.4%	30.7%	.760
Q17.3	−.449**	71.1%	10.5%	< .001
Q17.4	−.385*	55.3%	28.9%	< .001
Q17.5	−.023	19.3%	53.5%	.811
Q17.6	−.018	18.4%	41.2%	.849
Q18.1	.229*	67.5%	16.7%	.0144
Q18.2	.260*	18.4%	71.9%	.005
Q18.3	.278*	11.4%	79.8%	.003
Q18.4	.029	7.0%	79.8%	.759
Q18.5	.169	8.8%	79.8%	.073
Q19.1	−.137	84.2%	5.3%	.145
Q19.2	.307*	27.2%	69.3%	< .001
Q19.3	−.272*	26.3%	54.4%	.003
Q19.4	.195	54.4%	36.0%	.038
Q19.5	.272*	43.0%	50.9%	.003
Q20.1	.316*	50%	42.1%	< .001
Q20.2	.190	56.1%	28.1%	.043
Q20.3	−.160	27.2%	68.4%	.089
Q20.4	−.082	73.7%	22.8%	.387
Q20.5	.187	87.7%	7.9%	.047

B.5 Interview Participants

Table B.2: List of interview participants and their gender and sexual orientation.

Participant ID	Age	Gender	Sexual Orientation
S11	26 to 33	Non-binary	Queer
S25	18 to 25	Man	Homosexual
S33	18 to 25	Woman	Pansexual
S66	34 to 41	Man	Homosexual
S73	18 to 25	Woman	Homosexual
S74	18 to 25	Woman	Pansexual
S102	18 to 25	Non-binary	Queer
S106	18 to 25	Woman	Bisexual
S111	18 to 25	Non-binary	Pansexual

B.6 Interview Guide

The following text was used by interviewers to guide the interview procedure. The example questions were not in the original interview guide, but are included to illustrate our interview approach.

B.6.1 Introduction

Introduce yourself and study: "Good afternoon! Thank you for coming and participating in our interview study. I am... from Carnegie Mellon University. Today we're going to be asking you some questions about your experiences on dating services." Remind the participant of consent, confidentiality, and opt-out: "When taking our survey you signed a consent form. As stated in the consent form, what you say to us during the interview will be kept confidential. You can decline to answer any questions or you can stop participating at any time, while still receiving compensation." Ask "Do you consent to this interview being recorded for our internal use?" Once the participant confirms, begin the recording and ask them again to confirm that they have seen the consent form and consent to being recorded.

B.6.2 Interview

Ask questions based on the participant's survey responses. Prepare questions prior to the interview. Feel empowered to add new follow up questions as relevant. Examples:

- "Is there any particular reason you use hinge over other dating apps?" (from interview with S33)
- "In the survey you say that you use dating apps in order to connect... with people in general, but not looking for another relationship in particular. Could you explain more about why you use dating apps for this purpose rather than other social media applications?" (from interview with S25)
- "On the survey... you mentioned that people have taken screenshot photos without your permission. Can you describe an incident where this happened?" (from interview with S66)
- "On the survey you stated... that you're cautious about who you give pictures of yourself to. Could you explain how you decide who you want to share photos with?" (from interview with S74)

B.6.3 Conclusion

When all the questions are complete or 30 minutes have elapsed, conclude the interview. End the recording of the interview. Thank the participants for their time. Check if the participant has any questions before ending the Zoom meeting. Remind them: "We will send you an Amazon gift card code through email over the next week, if you have trouble or do not receive payment please contact us and we will solve the issue. You can use the email of the principal investigator included in the consent form."

Appendix C

Investigating the Experiences of Teenage Cryptocurrency Users on Reddit

C.1 Code Book

The following lists the codes derived from the inductive analysis of Reddit posts. Each definition includes an example of one of the posts assigned this code, the number of posts assigned this code, and the number of deleted or removed posts that were assigned this code.

Irrelevant: Not actually discussing cryptocurrency, but containing a keyword. e.g., ***Title:*** *We ZOINKED the kraken! Haha!* ***Body:****The kraken is now dead, ZOINKED by our fine crew.But the JINKIES nation lies in the distance. Prepare for D-Day, and prepare to ***ZOINK****
Count: 181 (10.8% of sample), 16 removed/deleted

Other (relevant): Post that mentions cryptocurrency but does not fit into other categories and/or does not really have content of interest. e.g., ***Title:*** *Fun fact:* ***Body:*** *In germany it's legal to scam nfts, because they don't count as reallife possessions, same with cs:go skins etc.*
Count: 19 (1.1% of sample), 3 removed/deleted

Irrelevant, cryptocurrency mentioned in passing: Cryptocurrency is mentioned in passing but is not actually the subject of the post or discussion in the comments. e.g., ***Title:*** *After you guys complete school what will you do?* ***Body:*** *everyone: become a gamer and and nft bro [newline] no like fr what will you do?*
Count: 86 (5.1% of sample), 3 removed/deleted

Deleted/Removed (ambiguous content): Post content is removed or deleted. The title mentions cryptocurrency, but it's impossible to determine exactly what the content was about. There are either no comments or the existing comments do not permit us to infer the content. e.g., ***Title:*** *Cryptocurrency* ***Body:*** *[removed]*
Count: 67 (4.0% of sample), all removed/deleted

Joke or meme: Post is either explicitly labeled as or is clearly interpreted by users as a joke; it also applies to meme images or videos. e.g., ***Title:*** *You think its FUNNY to take screenshots of people NFT huh* ***Body:*** *Property theft is a joke to you*
Count: 254 (15.2% of sample), 29 removed/deleted

Sarcasm: The post is not clearly a joke, but the content is so outrageous that it seems likely that it is sarcastic. e.g., ***Title:*** *GIVEAWAY ALERT Will send 1 Bitcoin to everyone who sends me 2 bitcoins HURRY!* ***Body:*** *Edit:this is no scam trust me bro*
Count: 52 (3.1% of sample), 16 removed/deleted

Reddit NFT: Post mentions the Reddit NFTs. e.g., ***Title:*** *Nft update, accepted it and it looks terrible, thank you Reddit* **Body**: [Post linked to image of Reddit NFT]
Count: 156 (9.3%), 28 removed/deleted

Criticism of crypto: Post is either a critique of crypto in general or specific projects. e.g., ***Title:*** *NFTS are scams and not worth any sort of money* ***Body:*** *You are buying the rights to some gettyimages monkey for $200k lol*
Count: 110 (6.6% of sample), 24 removed/deleted

Discussion of obtaining/holding crypto: Post includes a discussion of purchasing or otherwise obtaining crypto, including via giveaways; also includes situations where users just describe that they own crypto. e.g., ***Title:*** *I put my entire allowence into crypto* ***Body:*** *[deleted]*
Count: 226 (13.5% of sample), 57 removed/deleted

Discussion of mining crypto: Posts discusses mining crypto currency. e.g., ***Title:*** *[other] Late night Dogecoin miner thread, go!* ***Body:*** *[Post linked to an image of a terminal running Dogecoin mining software]*
Count: 59 (3.5% of sample), 21 removed/deleted

Discussion of trading crypto: Post discusses trading crypto currency or NFTs. Includes instances where users describe the act of trading without calling it trading. e.g., ***Title:*** *i'm trading 10$ worth of crypto* ***Body:*** *so quirky amirite [newline] no*
Count: 25 (1.5% of sample), 6 removed/deleted

Discussion of transferring crypto: Comments or post include discussion of users selling or otherwise transferring crypto. e.g., ***Title:*** *found some emails from 2017 showing Bitcoin i sent/received. This is now worth $333* ***Body:*** *[Post linked to an image showing a Coinbase confirmation email for sending 0.008931 Bitcoin]*
Count: 30 (1.8% of sample) , 15 removed/deleted

Cryptocurrency expected to increase in value: Post suggests an expected boost in a cryptocurrency or other digital asset. e.g., ***Title:*** *DOGECOIN TO THE MOON* ***Body:*** *[removed]*
Count: 29 (1.7% of sample), 18 removed/deleted

Profit: Post includes discussion of user making money in some way off of crypto assets. e.g., ***Title:*** *Bought more of a crypto coin then it went up 10% within a few hours lul.* ***Body:****[removed]*
Count: 51 (3.0%), 13 removed/deleted

Loss: Post includes discussion of a user losing money in some way off of NFTs or another crypto. e.g., ***Title:*** *Lost almost 5K in my crypto since Elon Musk was a Douche on twitter and value plummeted by over 30%. I feel so dispondant. I need hugs.* ***Body****: [Post linked to an image of a portfolio of crypto assets worth around $7,700.]*
Count: 17 (1.0% of sample), 6 removed/deleted

Explicit promotion of specific project: Posts encourage users to use a particular project or service. e.g., ***Title:*** *Pi Network Cryptocurrency the most promising Cryptocurrency.* ***Body:*** *[removed]*
Count: 305 (18.2% of sample) , 228 removed/deleted

General promotion of crypto: The post generally promotes crpyto assets without specifying a specific project; includes posts that push back against criticism. e.g., ***Title:*** *Cryptocurrency is the only valuable thing in the world that can be set up in a way where nobody can take it from you.* ***Body:*** *[no text]*
Count: 35 (2.1% of sample) , 8 removed/deleted

Informational question about crypto: Post or comments that asks for information about crypto (e.g., what it is, how it works, etc.) e.g., ***Title:*** *Can someone explain to me what is dogecoin is it a new form of crypto currency or sumthin.* ***Body:*** *[deleted]*
Count: 68 (4.1% of sample), 11 removed/deleted

Subjective question about crypto: Post that asks a subjective question about users' perspective on cryptocurrency or personal experience with cryptocurrency. e.g., ***Title:*** *What is your opinion on NFTs?* ***Body:*** *I'm curious to hear what everyone thinks about these things. edit: Interesting. It sounds like most people don't like them, but also dont know very much about them.*
Count: 105 (6.3% of sample), 34 removed/deleted

Using crypto for payment: Post mentions or is about using crypto for payment. e.g., ***Title:*** *SKIDS ACCOUNTS SHOP QUALITY SUPPORT + LIFETIME Warranties - ONLY $4+ Netflix, Hulu, Spotify, etc. [W] PayPal, BTC, Cash App.* ***Body:****[removed]*
Count: 25 (1.5% of sample), 10 removed/deleted

Crime: Post or comments suggest that the user has fallen victim to a scam, virus, or other crime; also applies to posts that describe a crime, even if the user is not the victim. e.g., ***Title:*** *Crypto scammers are the scum of the earth.* ***Body:*** *So, i went onto my YouTube subscriptions inbox today, only to notice a Livestream promoting Ethereum. Turns out one of the YT channels I'm subbed to was hacked by crypto scammers. Seriously, fuck these people. I know it's fun to laugh at NFT collectors, spending millions on Bored Apes, but screw these people.*

Count: 27 (1.6% of sample) , 4 removed/deleted

Working on a crypto project: Post or comments discuss a teenager creating or working for crypto asset project. e.g., ***Title:*** *I made some new nfts* ***Body:*** *So far I have giga chad, peter Griffin, moist critical, and American syco. Who else should I make?*
Count: 36 (2.1% of sample), 3 removed/deleted

Discussion of news in crypto: Post or comments about news related to a crypto project. e.g., ***Title:*** *Binance Acquired regulated Crypto exchange.* ***Body:*** *[Post linked to article about acquisition]*
Count: 69 (4.1% of sample), 13 removed/deleted

Giveaway: Post or comment purports to be giving away/airdropping some sort of cryptocurrency; includes "faucets." e.g., ***Title:*** *Free Bitcoin for Teens Only.* ***Body:*** *Make any comment below for your complimentary 100 bits of Bitcoin and it will be delivered to you via the ChangeTip bot! What is Bitcoin? What is ChangeTip? Edit: I am going to bed, but leave your comments below if you haven't been tipped and I will get to you in the morning. Night.*
Count: 145 (8.7% of sample) , 118 removed/deleted

Gambling: Post discusses gambling with crypto. e.g., ***Title:*** *I just made $250 gambling bitcoin.* ***Body:*** *[Post links to a referral code for Bitsler.com]*
Count: 13 (0.8% of sample), 3 removed/deleted

Begging: Post requests that users send cryptocurrency to the poster in exchange for no consideration. e.g., ***Title:*** *Anyone have some bitcoin to spare.* ***Body:*** *[removed]*
Count: 8 (0.5% of sample), 5 removed/deleted

C.2 Filtering Keywords

The following tables show the regular expression strings used to identify posts that were potentially about crypto in the teenage subreddits. Matches were identified using Python 3.9.7 re package. The `\b` at the beginning and end of each string ensures that matches are only returned if the match is a free-standing word. For example, `\bterra\b` matches "terra" or "(terra)" but not "terrarium." Complete documentation can be found at `https://docs.python.org/3.9/library/re.html`.

Table C.1: Keywords derived from the top ten coins by market cap, according to `coinmarketcap.com` as of June 1st, 2023. For most currencies, regex strings are included for both the currency's name and abbreviation(s). Some names and abbreviations are excluded, as we anticipated that their inclusion cause too many false positives (i.e., polygon, ripple, and doge). A regex string for "binance coin" was also excluded, since there is a string for "binance" in another category of keywords

Keyword	# in Sample	% Irrelevant (#)
`\bbit( |-)?coin(s)?\b`	280	7.5% (21)
`\bbtc\b`	33	3.0% (1)
`\bdoge( |-)?coin(s)?\b`	83	6.0% (5)
`\bethereum(s)?\b`	56	3.6% (2)
`\beth\b`	35	60.0% (21)
`\bether(s)?\b`	removed	removed
`\btether(s)?\b`	14	85.7% (12)
`\busdt\b`	3	33.3% (1)
`\bmatic\b`	5	100.0% (5)
`\bsolana\b`	3	0.0% (0)
`\bsol\b`	removed	removed
`\bcardano\b`	1	0.0% (0)
`\bada\b`	removed	removed
`\bxrp\b`	1	0.0% (0)
`\busd( |-)?coin(s)?\b`	2	50.0% (1)
`\busdc\b`	1	100.0% (1)
`\bbnb\b`	6	83.3% (5)

Table C.2: Keywords derived from three high-profile failed crypto projects: the Terra USD stable coin [67], SafeMoon [513], and Squid Game Token [461]. For TerraUSD, we include both the project name (Terra) and the abbreviations for the coin pair that was central to the algorithmic stablecoin system (TerraUSD and Luna). For Squid Game Token, we include two possible variations (Coin or Token)

Keyword	# in Sample	% Irrelevant (#)
\blun(c)?\b	4	100.0% (4)
\bust(c)?\b	33	93.9% (31)
\bterra\b	34	97.1% (33)
\bsafe(|-)?moon\b	8	0.0% (0)
\bsquid(|-)?game(|-)?coin\b	0	—
\bsquid(|-)?game(|-)?token\b	0	—

Table C.3: Keywords derived from the names of centralized crypto exchanges, decentralized crypto exchanges, and interest bearing services.

Keyword	# in Sample	% Irrelevant (#)
\bbinance\b	7	0.0% (0)
\bftx\b	3	0.0% (0)
\bcoin(|-)?base\b	10	10.0% (1)
\bkraken\b	35	100.0% (35)
\bsushi(|-)?swap\b	0	—
\buni(|-)?swap\b	2	0.0% (0)
\bpancake(|-)?swap\b	2	0.0% (0)
\bcelsius\b	removed	removed
\bblock(|-)?fi\b	0	—
\bnexo\b	3	100.0% (3)

Table C.4: Keywords derived from the names of cryptocurrency gambling services

Keyword	# in Sample	% Irrelevant (#)
\bstake(\.com|\.us)\b	1	0.0% (0)
\bcloud(|-)?bet\b	0	—
\bmeta(|-)?spin(s)?(|-)?casino\b	0	—
\b7(|-)?bit(|-)?casino\b	0	—
\bbets\.io\b	0	—
\bbit(-|)?starz\b	0	—
\bm(-|)?bit(|-)?casino\b	0	—
\broll(|-)?bit\b	0	—

Table C.5: Keywords derived from words and concepts related to crypto assets.

Keyword	# in Sample	% Irrelevant (#)
\bcrypto(s)?\b	281	11.0% (31)
\bblock(|-)?chain(s)?\b	35	25.7% (9)
\bnft(s)?\b	680	3.5% (24)
\bdefi\b	5	20.0% (1)
\bcrypto(|-)?currenc(y|ies)\b	173	12.1% (21)
\bdao(s)?\b	1	100.0% (1)
\bdecentralize(d|s)?\b	12	66.7% (8)
\bstable(|-)?coin(s)?\b	1	0.0% (0)
\bmoon(|-)?shot(s)?\b	8	100.0% (8)
\bdegen(s)?\b	removed	removed
\baping\b	removed	removed
\bde(|-)?peg(s|ging)?\b	0	—
\bshill(s|ing|ed)?\b	removed	removed
\bmarket(|-)?cap(s)?\b	8	12.5% (1)

Appendix D

Investigating the Susceptibility of Teens and Adults to YouTube Giveaway Scams

D.1 Artifact Information

While selected elements of replication material are provided in this thesis, we provide an external artifact with the materials necessary to replicate the behavioral experiment. It also provides the data we collected in the experiment and the analysis code used to obtain statistical results. The artifact is available on Zenodo[1] and Github[2], and is composed of the following parts:

1. **Survey instruments**: The survey instruments used to conduct the study, including the informed consent forms, main experiment survey, and teen compensation survey.
2. **Media**: the pictures and videos presented to participants during the study.
3. **Data**: the data obtained from the experiment and metadata explaining the meaning of each field.
4. **Recruitment materials**: the flier sent to parents of teenage participants and the advertisement posted on Prolific to recruit adult participants.
5. **Analysis code**: the analysis script used to perform statistical testing for the experiment. The analysis script is distributed in the form of an R notebook, with explanations provided alongside the code.
6. **README**: a file explaining the file structure of the artifact.

Analysis was performed using R version 4.3.3. However, it is not expected that the version of R will affect the results. The analysis script is distributed in the form of an R Markdown (.Rmd) file, requiring RStudio[3] to view and run. Data is stored in Excel format (.xlsx), which requires Microsoft Excel or compatible spreadsheet software to view. The survey is in Microsoft Word format (.docx), which requires Microsoft Word or compatible document editing software to view. Media is distributed as PNG files for images and MP4 files for videos. Finally, the follow-

[1] `https://doi.org/10.5281/zenodo.13910629`
[2] `https://github.com/elijahbs/The-Kids-Are-All-Right`
[3] `https://posit.co/download/rstudio-desktop/`

ing readily available R packages are used in the analysis script: `RVAideMemoire`,[4] `dplyr`,[5] `rstatix`,[6] `readxl`,[7] `rcompanion`,[8] `DescTools`[9] and `stringr`.[10]

D.2 Survey Instrument

The following subsections include the survey instrument used in our study. Participants were randomly sorted into one of two conditions corresponding to the topic of the stimuli they were shown (i.e., Spotify or Roblox). We use (Spotify text)/(Roblox text) to indicate how questions vary for each condition. Italicized text is used to indicate survey flow or other annotations. This text was not shown to participants. If a question option ends in the text "(please specify)," a text box was included for participants to explain their answer. Finally, section headings are used to organize the text but were not shown to participants.

D.2.1 Internet Usage Questionnaire

We will now ask you some questions about your technology habits.

Note: As a reminder, you should not share this survey with anyone else. Multiple submissions from the same survey link will result in your submission being discarded without receiving compensation.

1. About how much time do you spend each day on digital entertainment? This includes things like watching television, playing video games, using social media, and other activities on devices with screens that you do for fun.
   (a) None
   (b) Less than 2 hours
   (c) 2 to 4 hours
   (d) 4 to 8 hours
   (e) More than 8 hours
2. What devices have you used for digital activities for fun in the last week? Please select all that apply.
   (a) Smartphone
   (b) Tablet (e.g., iPad, Samsung Galaxy Tab, etc.)
   (c) Laptop or desktop computer
   (d) Game console (e.g., Nintendo Switch, PlayStation 5, etc.)

[4] https://cran.r-project.org/web/packages/RVAideMemoire/index.html
[5] https://cran.r-project.org/web/packages/dplyr/index.html
[6] https://cran.r-project.org/web/packages/rstatix/index.html
[7] https://cran.r-project.org/web/packages/readxl/index.html
[8] https://cran.r-project.org/web/packages/rcompanion/index.html
[9] https://cran.r-project.org/web/packages/DescTools/index.html
[10] https://cran.r-project.org/web/packages/stringr/index.html

(e) Smartwatch (e.g., Apple Watch, Samsung Galaxy Watch, etc.)
(f) Television
(g) Virtual Reality Devices (e.g, Meta Quest 3, Oculus Quest 2, etc.)
(h) Other (please specify)

3. What social media services have you used in the last week? Please select all that apply. *Participants who did not select "YouTube" in this question were screened out*
   (a) Discord
   (b) Facebook
   (c) Instagram
   (d) Pinterest
   (e) Reddit
   (f) Snapchat
   (g) TikTok
   (h) Twitter
   (i) YouTube
   (j) BeReal
   (k) Other (please specify)
   (l) I do not use social media *(exclusive answer choice)*
   (m) Prefer not to share *(exclusive answer choice)*
4. Approximately how many hours per week do you spend watching online videos for fun?
   (a) None
   (b) Less than an hour
   (c) 1 to 5 hours
   (d) 5 to 10 hours
   (e) 10 to 15 hours
   (f) 15 to 20 hours
   (g) More than 20 hours
5. Approximately how many hours per week do you spend playing mobile games for fun?
   (a) None
   (b) Less than an hour
   (c) 1 to 5 hours
   (d) 5 to 10 hours
   (e) 10 to 15 hours
   (f) 15 to 20 hours

(g) More than 20 hours

6. Approximately how many hours per week do you spend playing games on a computer or game console for fun?
   (a) None
   (b) Less than an hour
   (c) 1 to 5 hours
   (d) 5 to 10 hours
   (e) 10 to 15 hours
   (f) 15 to 20 hours
   (g) More than 20 hours
7. Approximately how many hours per week do you spend using social media for fun?
   (a) None
   (b) Less than an hour
   (c) 1 to 5 hours
   (d) 5 to 10 hours
   (e) 10 to 15 hours
   (f) 15 to 20 hours
   (g) More than 20 hours
8. Approximately how many hours per week do you spend browsing non-social media websites for fun?
   (a) None
   (b) Less than an hour
   (c) 1 to 5 hours
   (d) 5 to 10 hours
   (e) 10 to 15 hours
   (f) 15 to 20 hours
   (g) More than 20 hours

D.2.2 Search result selection task

For the following questions, you should imagine you are visiting a friend. While you are together, they tell you that they are looking online for ways (to get free access to Spotify Premium, which normally costs $10.99 per month.)/(to earn Robux, which are the in game currency for the multiplayer game Roblox. Robux typically cost $5 for 400 Robux.) They are searching on YouTube to try to learn more, and they want your advice on what they should do.

9. The following image shows the first 9 results your friend found after searching for ("free spotify premium")/("free roblox robux") on YouTube. Select the videos that you recommend your friend click on, if any.

   Participants were shown nine YouTube search results corresponding to their condition in random order: three legitimate search results, three clickbait search results, and three scam search results. Participants could also chose "None of the videos"

10. What did you consider in selecting which videos to click on? *(paragraph text box)*
11. Which part(s) of the search results helped you make your decision?
    (a) The thumbnail image
    (b) The title
    (c) The name of the channel
    (d) The number of views
    (e) The length of the video
    (f) The profile picture of the channel
    (g) Other (please specify)

D.2.3 Video task

Participants were now randomly assigned to view two search results corresponding to their condition: one of the three scam videos and one of the three legitimate videos. Questions 12 to 17 and the associated instructional text were shown twice for each video.

Your friend clicked on one of the videos they found while searching for ("free spotify premium")/("free roblox robux.") *An image of the search result for the video was embedded here, followed by a redacted version of the YouTube video.*

12. Please view the video shown above with the sound and/or captions on and select the option(s) corresponding to the actions you think your friend should take. Please note: Any links to or instructions to access external sites will be covered with a grey rectangle. You should assume that you would be able to access the websites if you viewed the video in real life.
    (a) Visit the website(s) shown in the video
    (b) Search online to learn more about what the video describes
    (c) Look at the comments on the video
    (d) Leave a comment on the video (please specify)
    (e) Report the video to YouTube
    (f) Exit the video without doing anything
    (g) Other (please specify)
13. Why would you recommend that they take these action(s)? *(paragraph text box)*

14. What specific part(s) of the video most influenced your recommendation? *(paragraph text box)*

Participants were unable to advance to the next questions until at least half the length of the video had elapsed

Your friend went to the website linked in the video you just saw. Here is a video of your friend visiting the website. We have only recorded the video and not the audio. *A video of a person on a mobile phone browsing the legitimate or scam website was embedded here.*

15. Please view the video above and select the action(s) you recommend that your friend takes.
    (a) Follow all of the instructions in the video to complete the offer
    (b) Register for the website
    (c) Ask a parent for help
    (d) Search online to learn more about the website
    (e) Exit from the website without doing anything
    (f) Other (please specify)
16. Why would you recommend that they take these action(s)? *(paragraph text box)*
17. What specific part(s) of the website most influenced your recommendation? *(paragraph text box)*

Participants were unable to advance to the next questions until at least half the length of the video had elapsed

D.2.4 Post-task Questions

18. How often have you (used Spotify?)/(played the game Roblox?)
    (a) Never
    (b) Once or twice
    (c) Three or five times
    (d) More than five times
19. Have you ever (paid for Spotify Premium?)/(purchased Robux?)
    (a) Yes
    (b) No
20. Have you ever received (a free subscription to Spotify Premium)/(free Robux) through a coupon, free trial offer, or giveaway?
    (a) Yes
    (b) No
21. Have you ever searched online for ("free Spotify premium")/("free Roblox robux") or something similar before?
    (a) Yes
    (b) No

22. *If participant says "Yes" to question 21* What happened as a result of searching ("free Spotify premium")/("free Roblox robux") or something similar? *(paragraph text box)*
23. *If participant says "No" to question 21* How likely are you to search online for "free Roblox robux" or something similar?
    (a) Extremely unlikely
    (b) Somewhat unlikely
    (c) Neither likely nor unlikely
    (d) Somewhat likely
    (e) Extremely likely
24. Thinking back to this video that you saw previously, what were your impressions of the video? *An image of the search result for one of the videos the participant saw was embedded here. This question was repeated two times, once for each video that the participant saw.*
    (a) Definitely a scam
    (b) Probably a scam
    (c) I'm not sure
    (d) Probably legitimate
    (e) Definitely legitimate

We will now ask a few more questions about you.

25. Approximately how much money in dollars do you have to spend on yourself in an average week? *(number entry box)*
26. Approximately how much money in dollars do you spend online in an average week? *(number entry box)*
27. How often have you paid for something online but never received the item or a refund?
    (a) Never
    (b) Once or twice
    (c) Three or five times
    (d) More than five times
28. How often have you received cash back or a rebate for buying an item?
    (a) Never
    (b) Once or twice
    (c) Three or five times
    (d) More than five times
29. How often have you completed a task online, such as a survey, in exchange for money or gift cards?
    (a) Never

(b) Once or twice

(c) Three or five times

(d) More than five times

30. How often have you purchased traditional investments, such as stocks or bonds?

(a) Never

(b) Once or twice

(c) Three or five times

(d) More than five times

31. How often have you purchased digital assets, such as cryptocurrencies or non-fungible tokens (NFTs)?

(a) Never

(b) Once or twice

(c) Three or five times

(d) More than five times

32. How often have you used coupons or discount codes to purchase items online?

(a) Never

(b) Once or twice

(c) Three or five times

(d) More than five times

33. How often have you completed a task online, such as a survey or downloading an app, expecting to be paid but never received the money?

(a) Never

(b) Once or twice

(c) Three or five times

(d) More than five times

D.2.5 Demographics

This section of questions was only included in the adult version of the survey. Parents were asked to answer these questions on behalf of their teenage children in modified form (e.g., rather than asking for their race, parents were asked for their child's race). Parents were also asked if their child contributed to their household income.

34. What is your approximate annual household income?

(a) Less than $20,000

(b) $20,000 to $39,999

(c) $40,000 to $59,999

(d) $60,000 to $79,999
(e) $80,000 to $99,999
(f) $100,000 to $149,999
(g) Over $150,000
(h) Prefer not to say

35. What is your gender?
    (a) Male
    (b) Female
    (c) Non-binary
    (d) Prefer to self-describe *(text box)*
    (e) Prefer not to say
36. What is your age? *(number entry box)*
37. What is your race? (select all that apply)
    (a) White
    (b) Black or African American
    (c) American Indian or Alaska Native
    (d) Asian
    (e) Native Hawaiian or Pacific Islander
    (f) Hispanic and/or Latino/a/x
    (g) Prefer to self-describe *(text entry box)*
    (h) Prefer not to say
38. In which state do you currently reside? *(drop-down box with all 50 US states, D.C., and Puerto Rico)*
39. What type of community do you live in?
    (a) Rural
    (b) Suburban
    (c) Urban

Debrief

Thank you for finishing the survey. Some of the videos you were shown were examples of scams. If you follow their advice you will not get anything for free (but you might waste your time or give your personal information to scammers). Here are the thumbnails of the scam videos: *The three scam search results shown in the thumbnail selection task were embedded here.*

Here are some tips to avoid these and other scams on social media:

- If a post seems too good to be true, it probably is. People offering expensive items for free could be trying to scam you out of your personal information or your money.

- Always go to a service's official website to start a free trial.
- When games charge money for in-game currency, there is no way to get currencies for free outside the game.
- Be careful when giveaways offer items for an unlimited amount of time. All real giveaways must have clear rules and end dates.
- If you aren't sure if a post is real, ask a trusted adult for help, such as a parent, guardian, or teacher.

You can visit YouTube's Community Guidelines on scams for more information. *This text included a link to Google's YouTube Community Guidelines page on scams:* `https://support.google.com/youtube/answer/2801973`

D.3 Codebook

Below is the list of codes used to analyze participants' explanations for choosing certain actions on both the scam and legitimate stimuli they were shown. The codes are listed with their definitions, example quotes from participant responses, and the number of times the code was used for the different types of stimuli.

D.3.1 Video and Website Codes

- **Reasoning Codes:**
  - **Identifies video/website as suspicious**: The response indicates that the user finds this website or video suspicious in some way. *Example: "It's a scam"*. **Counts**: *Scam Video Stimuli*: 137, *Scam Web Stimuli*: 160, *Legit Video Stimuli*: 37, *Legit Website Stimuli*: 32
  - **Privacy/security risk**: Participant views the video/website as a risk to their privacy or security (e.g., malware, hacking, etc.). *Example: "The website...wants you to download something that will most likely have a virus, malware, or something mischievous program embedded in it."* **Counts**: *Scam Video Stimuli*: 16, *Scam Web Stimuli*: 22, *Legit Video Stimuli*: 6, *Legit Website Stimuli*: 15
  - **Too good to be true**: Participant indicates that the value of the offer is too good to be true. *Example: "just the idea in general of trying to get something for nothing. it seems illegal even"* **Counts**: *Scam Video Stimuli*: 15, *Scam Web Stimuli*: 28, *Legit Video Stimuli*: 11, *Legit Website Stimuli*: 6
  - **Too much work**: The response indicates that completing the task involves too much work. *Example: "Again it is too much hassle for what little bit you get. I'm sure they're getting plenty from you for signing up. Not worth it"* **Counts**: *Scam Video Stimuli*: 28, *Scam Web Stimuli*: 18, *Legit Video Stimuli*: 9, *Legit Website Stimuli*: 10
  - **Prior knowledge (doesn't work)**: Participant references their personal experience in explaining that they think a method doesn't work. *Example: "I'm familiar with this model where apps must be downloaded/played in order for a reward... 90% of the time, they are a scam and don't end up giving you the reward..."* **Counts**: *Scam Video Stimuli*: 15, *Scam Web Stimuli*: 10, *Legit Video Stimuli*: 2, *Legit Website Stimuli*: 1
  - **Wants to protect others**: Participant indicates that they want to protect others from this type of video. *Example: "So people can stop watching these videos"* **Counts**: *Scam Video Stimuli*: 5, *Scam Web Stimuli*: 0, *Legit Video Stimuli*: 1, *Legit Website Stimuli*: 0
  - **Caution**: Participant indicates that their friend should be cautious/careful without explaining more. *Example: "Be cautious"* **Counts**: *Scam Video Stimuli*: 4, *Scam Web Stimuli*: 4, *Legit Video Stimuli*: 2, *Legit Website Stimuli*: 7
  - **Identifies video/website as reasonable**: The response indicates that the participant thinks the video/website presents a genuine offer. *Example"... it is probably real."*

Counts: *Scam Video Stimuli*: 23, *Scam Web Stimuli*: 22, *Legit Video Stimuli*: 95, *Legit Website Stimuli*: 123

- **Prior knowledge (works)**: Participant references their personal experience or prior knowledge in explaining that they think a method probably works. *Example: "I have seen sites like this where you can do offers to earn money, so I wouldn't immediately call it a scam but I would want to do more research before going to the site."* **Counts**: *Scam Video Stimuli*: 9, *Scam Web Stimuli*: 2, *Legit Video Stimuli*: 14, *Legit Website Stimuli*: 9
- **Uninterested in offer (non-security related)**: Answer indicates that the participant is not interested in the offer in the video/on the website, but their reasoning does not seem to be based on whether or not the method works. *Example: "No interest."* **Counts**: *Scam Video Stimuli*: 0, *Scam Web Stimuli*: 4, *Legit Video Stimuli*: 1, *Legit Website Stimuli*: 3
- **Confusing**: Participant indicates that they find the video or website to be confusing. *Example: "...It was very confusing."* **Counts**: *Scam Video Stimuli*: 2, *Scam Web Stimuli*: 2, *Legit Video Stimuli*: 0, *Legit Website Stimuli*: 0
- **Simple/easy**: Participant indicates that they think the process described in the video is simple or easy to do. *Example: "the simplicity of it"* **Counts**: *Scam Video Stimuli*: 1, *Scam Web Stimuli*: 6, *Legit Video Stimuli*: 7, *Legit Website Stimuli*: 21
- **Existing account needed**: Participant references the need to be part of a loyalty program or have an existing account (e.g., OnePLus, Paypal, Walmart). *Example: "Because A walmart plus account is needed."* **Counts**: *Scam Video Stimuli*: 0, *Scam Web Stimuli*: 0, *Legit Video Stimuli*: 3, *Legit Website Stimuli*: 0
- **Video is promotional**: Participant says that the video is a promotion for the website and influences their decision. *Example: "It seems like an ad or promotion for a website..."* **Counts**: *Scam Video Stimuli*: 4, *Scam Web Stimuli*: 3, *Legit Video Stimuli*: 1, *Legit Website Stimuli*: 2
- **Wants to learn more**: The response indicates that the participant wants to gather more information using external resources or feedback from others. *Example: "To make sure that the process is legal and that it's real."* **Counts**: *Scam Video Stimuli*: 80, *Scam Web Stimuli*: 43, *Legit Video Stimuli*: 113, *Legit Website Stimuli*: 65
- **Reiterates action**: Answer reiterates the action that the user selected without explaining their motivation. *Example: "See what the commenters say first"* **Counts**: *Scam Video Stimuli*: 6, *Scam Web Stimuli*: 9, *Legit Video Stimuli*: 3, *Legit Website Stimuli*: 22
- **Prior knowledge (unclear)**: Participant discusses a personal experience with similar or the same website, but does not explain whether that experience shows that they do or don't work. *Example: "Like I said in the last one some are okay to use..."* **Counts**: *Scam Video Stimuli*: 0, *Scam Web Stimuli*: 1, *Legit Video Stimuli*: 1, *Legit Website Stimuli*: 0

- **Unclear/Irrelevant**: Answer is unclear or fails to answer the questions asked. *Example: "No"* **Counts**: *Scam Video Stimuli*: 12, *Scam Web Stimuli*: 11, *Legit Video Stimuli*: 15, *Legit Website Stimuli*: 12
- **Unsure**: Participant expresses doubt or lack of confidence in answer. *Example: "video seems like a generic voiceover thing that's not super trustworthy but maybe not a scam"* **Counts**: *Scam Video Stimuli*: 16, *Scam Web Stimuli*: 21, *Legit Video Stimuli*: 17, *Legit Website Stimuli*: 24
- **Misunderstood task**: Participant's response suggests that they did not understand the survey task. *Example: "Not explaining anything and blocking informed on screen"* **Counts**: *Scam Video Stimuli*: 1, *Scam Web Stimuli*: 8, *Legit Video Stimuli*: 0, *Legit Website Stimuli*: 8
- **Other reasoning**: Participant expresses reasoning that does not fall into one of the above categories. *Example: "I don't support stealing, and accessing platforms in this way qualifies as theft..."* **Counts**: *Scam Video Stimuli*: 3, *Scam Web Stimuli*: 14, *Legit Video Stimuli*: 6, *Legit Website Stimuli*: 9

- **Feature Codes**
  - **All of the video**: Participant states that the entire video influenced their decision without particularity. *Example: "The whole video"* **Counts**: *Scam Video Stimuli*: 25, *Scam Web Stimuli*: 13, *Legit Video Stimuli*: 12, *Legit Website Stimuli*: 5
  - **All of the website**: Participant states that the website influenced their decision without particularity. *Examples: "Asking me to go to a third party site.," "When the website is mentioned"* **Counts**: *Scam Video Stimuli*: 26, *Scam Web Stimuli*: 17, *Legit Video Stimuli*: 7, *Legit Website Stimuli*: 6
  - **Region of offer**: Participant notes the region of the offer shown (e.g., currency) and indicates that it affected their answer. *Example: "The amount to be paid is in a foreign currency and it seems like searching for a different link would be appropriate."* **Counts**: *Scam Video Stimuli*: 0, *Scam Web Stimuli*: 0, *Legit Video Stimuli*: 3, *Legit Website Stimuli*: 6
  - **Length of video (too long)**: Participant indicates that the video is too long and that this affected their decision. *Example: "The length, I like shorter videos..."* **Counts**: *Scam Video Website*: 2, *Scam Web Stimuli*: 0, *Legit Video Stimuli*: 3, *Legit Website Stimuli*: 0
  - **Length of video**: Participant mentions the length of the video but does not elaborate about it being too long or too short. *Example: "The length..."* **Counts**: *Scam Video Stimuli*: 1, *Scam Web Stimuli*: 0, *Legit Video Stimuli*: 0, *Legit Website Stimuli*: 0
  - **Length of video (too short)**: Participant indicates that the video is too short, and that this affected their decision. *Example: "the fact that it is very short..."* **Counts**: *Scam Video Stimuli*: 1, *Scam Web Stimuli*: 0, *Legit Video Stimuli*: 3, *Legit Website Stimuli*: 0

- **Personal information required**: Participant is suspicious of the requirement to enter personal information or believes that tasks will require their personal information. *Example: "You have to enter personal details and credit card information and have no idea if this website is legit."* **Counts**: *Scam Video Stimuli*: 29, *Scam Web Stimuli*: 34, *Legit Video Stimuli*: 19, *Legit Website Stimuli*: 26
- **Presenter**: Participant references an aspect of the presenter (i.e., their voice or appearance). *Example: "the artificial voice"* **Counts**: *Scam Video Stimuli*: 52, *Scam Website Stimuli*: 3, *Legit Video Stimuli*: 37, *Legit Website Stimuli*: 2
- **Chronological part of video**: Participant references a chronological part of the video without explaining what specifically they mean. *Example: "the middle part"* **Counts**: *Scam Video Stimuli*: 9, *Scam Web Stimuli*: 4, *Legit Video Stimuli*: 5, *Legit Website Stimuli*: 5
- **Urgency**: Participant references the limited time nature or urgency of the offer. *Example: "The fact that he kept saying there's a limited amount so act fast"* **Counts**: *Scam Video Stimuli*: 4, *Scam Web Stimuli*: 4, *Legit Video Stimuli*: 0, *Legit Website Stimuli*: 0
- **Instructional aspect of video/website**: Participant references parts of the video related to instructions. *Example: "they show how to do it"* **Counts**: *Scam Video Stimuli*: 33, *Scam Web Stimuli*: 11, *Legit Video Stimuli*: 46, *Legit Website Stimuli*: 3
- **Request to like and subscribe**: Participant mentions the request to like and subscribe, and that this influenced their answer. *Example: "They want you to like and subscribe to their channel"* **Counts**: *Scam Video Stimuli*: 1, *Scam Web Stimuli*: 0, *Legit Video Stimuli*: 5, *Legit Website Stimuli*: 0
- **Thumbnail**: Participant mentions the thumbnail of the video. *Example: "Thumbnail"* **Counts**: *Scam Video Stimuli*: 3, *Scam Web Stimuli*: 1, *Legit Video Stimuli*: 2, *Legit Website Stimuli*: 0
- **Links**: Participant mentions the external links in the video, and this influenced their response. *"All of the links, especially that the link was in the description (not something you could search yourself)"* **Counts**: *Scam Video Stimuli*: 3, *Scam Web Stimuli*: 3, *Legit Video Stimuli*: 8, *Legit Website Stimuli*: 11
- **Non-official website**: Participant mentions that offer directs users to a non-official website. *Example: "Going to a site that wasn't connected to the Spotify brand."* **Counts**: *Scam Video Stimuli*: 10, *Scam Web Stimuli*: 9, *Legit Video Stimuli*: 5, *Legit Website Stimuli*: 9
- **Personal information not required**: The participant mentioned that a Roblox or Spotify sign-in, or other personal information, is not needed for the process in the video. *Example: "They weren't asking you to sign into some dodgy site that's probably phishing for information."* **Counts**: *Scam Video Stimuli*: 4, *Scam Web Stimuli*: 3, *Legit Video Stimuli*: 1, *Legit Website Stimuli*: 0
- **Referral/promo codes**: Participant mentions that the video showed promo codes or referral codes for a legitimate or scam website. *Example: "...they give promo codes*

too" **Counts**: *Scam Video Stimuli*: 1, *Scam Web Stimuli*: 1, *Legit Video Stimuli*: 2, *Legit Website Stimuli*: 0

- **Trial**: Participant points out that video shows a free trial. *Example: "...is just showing him how to sign up for a free trial, rather than showing ways to get Spotify premium for free"* **Counts**: *Scam Video Stimuli*: 0, *Scam Web Stimuli*: 0, *Legit Video Stimuli*: 18, *Legit Website Stimuli*: 12
- **Aspects of production quality**: Participant mentions aspects of the production quality of the video in general or in particular (e.g., video editing, background music, sound effects, visual style, etc.). *Examples: "The animation and music just seemed a little sketchy.," "The screen at the side," and "The quickness of it and the annoying money sounds."* **Counts**: *Scam Video Stimuli*: 25, *Scam Web Stimuli*: 1, *Legit Video Stimuli*: 17, *Legit Website Stimuli*: 2
- **Tasks**: Participant references the need to do surveys/download apps/other tasks without explaining how this influenced their decision. *Example: "When it got to the part where you have to install 3 apps in order to unlock the spotify code."* **Counts**: *Scam Video Stimuli*: 78, *Scam Web Stimuli*: 71, *Legit Video Stimuli*: 10, *Legit Website Stimuli*: 13
- **Video appears on website**: Participant notices that the video they are viewing, or one by the same creator, appears on the website they are looking at. *Example: "...the fact that the man in the video appears to be in another video advertising the coupon site..."* **Counts**: *Scam Video Stimuli*: 3, *Scam Web Stimuli*: 5, *Legit Video Stimuli*: 0, *Legit Website Stimuli*: 0
- **Proof**: Participant feels that the video shows proof of the offer working. *Example: "the fact that he showed proof..."* **Counts**: *Scam Video Stimuli*: 0,*Scam Web Stimuli*: 0, *Legit Video Stimuli*: 4, *Legit Website Stimuli*: 0
- **Lack of Proof**: Participant feels that the video does not show proof of the offer working. *Example: "the fact that it... didn't actually show the robux in the account"* **Counts**: *Scam Video Stimuli*: 5, *Scam Web Stimuli*: 1, *Legit Video Stimuli*: 2, *Legit Website Stimuli*: 0
- **Reputable companies/websites**: Participant references the company running the website or involved in the tasks/offers (i.e., Microsoft, Spotify, OnePlus). *Example: "When he mentions paypal and walmart as they are reputable trusted companies and websites."* **Counts**: *Scam Video Stimuli*: 0, *Scam Web Stimuli*: 0, *Legit Video Stimuli*: 65, *Legit Website Stimuli*: 84
- **Code locker**: Participant notes the"human verification" or "unlocking code" portion of the website. *Example: "The "unlocking code" part, scam websites often do that to trick people into thinking that they're legit."* **Counts**: *Scam Video Stimuli*: 1, *Scam Web Stimuli*: 5, *Legit Video Stimuli*: 0, *Legit Website Stimuli*: 0
- **Other feature of video or website**: Participant references other features of the video not already specified by another code. This includes features of the channel icon or

the website in general. *Example: "free sign up"* **Counts**: *Scam Video Stimuli*: 19, *Scam Web Stimuli*: 17, *Legit Video Stimuli*: 27, *Legit Website Stimuli*: 35

- **Website-only Feature Codes**
  - **Refers back to video**: Participant clearly refers back to the video to make their decision. *Example: "It followed what the video said would happen..."* **Counts**: *Scam Web Stimuli*: 16, *Legit Website Stimuli*: 26
  - **Website moving**: Participant notes the website says it is moving/changing names. *"the part where it said they were changing name"* **Counts**: *Scam Web Stimuli*: 13, *Legit Website Stimuli*: 0
  - **Aesthetic quality of website**: Participant mentions aspects of the aesthetic quality of the website in general or in particular (e.g.,. design, layout, etc.). *Example: "The design of the site"* **Counts**: *Scam Web Stimuli*: 65, *Legit Website Stimuli*: 33

D.3.2 Search Result Codes

- **Reasoning Codes:**
  - **Reasonable**: Participant states that they considered plausibility/implausibility. *Example: "Well a big factor was which videos sounded truthful and didn't seem like scams or clickbait..."* **Count**: 101
  - **Too good to be true (search result)**: Participant states that the concept of free Spotify Premium/Roblox Robux in general is unrealistic, without necessarily discussing the thumbnails. *Example: "...also I don't believe you can access Spotify Premium for free legally."* **Count**: 20
  - **Other reasoning**: Participant relies on other reasoning not covered by the other codes. *Example: "I considered if this was something I was interested in and unfortunately I'm not"* **Count**: 16
  - **Unsure**: Participant expresses that they are unsure or don't know what their selection should be. *"I don't know"* **Count**: 1
  - **Unclear/irrelevant (thumbnail)**: Participant response does not relate to the question asked or is unclear. *Example: "The truth"* **Count**: 15
- **Feature Codes**
  - **Text**: Participant states that they considered the title(s) of videos OR describes text they were looking for from the titles/thumbnail images. *Example: "...offered a free trial or said it was free."* **Count**: 67
  - **Channel**: Participant states that they considered some aspect of the channel when selecting thumbnails. *Example: "...how legit the channel looked."* **Count**: 29
  - **Views**: Participant states that they considered the number of views on the video. *Example: "How many views a video had..."* **Count**: 67
  - **Dates**: Participants states that they considered when the videos were uploaded. *Example: "...How recent was this video uploaded..."* **Count**: 22

- **Length of video**: Participant states that they considered the length of the video, including whether it was a YouTube short or not. *Example: "I chose the ones that were not 'shorts' because I am not sure you could give the information needed in under a minute."* **Count**: 21
- **Mentions specific thumbnail**: Participant seems to describe the features shown in a specific thumbnail. *"He shows that he got Spotify for himself for free"* **Count**: 33
- **Other aesthetic qualities of thumbnail images**: Participant references some feature of thumbnail images in particular or general (e.g., emojis, people in images, etc.). *Example: "The people on the video promo looked trustworthy is the main thing, otherwise I would have my reservations on picking any of them."* **Count**: 94

D.4 Recruitment material

D.4.1 Parent/Child Recruitment Poster

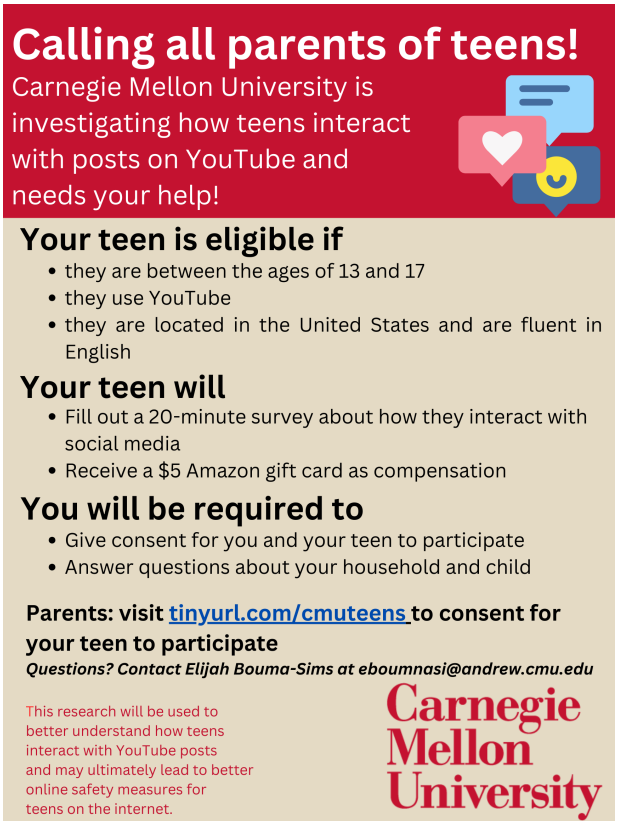


Figure D.1: Flier used to recruit parents of teen participants.

D.4.2 Prolific requirement advertisement

Study Title: Study of User Interaction with Common Social Media Posts.
Body: We are inviting you to participate in a voluntary research study evaluating how social media users interact with certain types of social media posts. Participants must be over the age of 18, reside in the United States, be fluent in English, and use YouTube.

This survey should take approximately 20 minutes to complete, and participants will be compensated $5.00. During the study, you will be asked to view social media posts and indicate how you would react to them.

D.5 Stimuli

In this section, we provide the thumbnail stimuli and describe the video stimuli represented in the behvioral experiment. The videos can be found the external artifact.

D.5.1 Spotify Stimuli

Spotify Legitimate 1 Stimuli

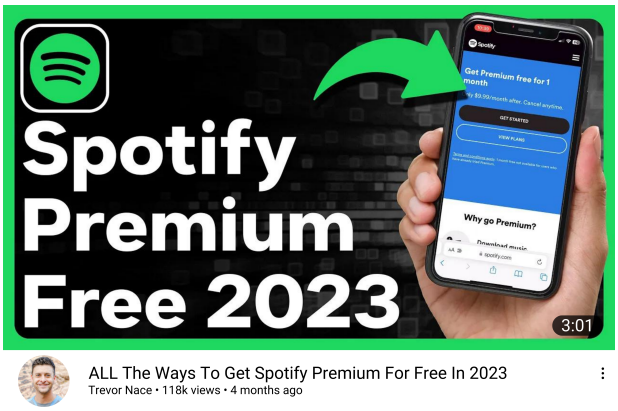


Figure D.2: Spotify Legitimate 1 Search Result

YouTube Video Description: A man discusses a variety of ways to get Spotify Premium for free for one, three, or six months, and all of these options are valid as of April 2023. Spotify itself will provide one month free for anyone or two months free for verified students. A three month free trial is available for those who connect a PayPal account as a payment method on Spotify. Finally, a six month free trial is available for those who connect their Walmart+ account. The man discusses that after the trials are over, people will have to pay to continue using Spotify Premium; the video ends with the speaker indicating that they will answer questions in the comments.
Website Video Description: The Spotify website is shown on the Spotify Premium page. There are free trial offers for all account types, which include: Individual, Duo, Family, and Student. The "Get Started" button is selected, which goes to the Spotify sign-in page.

Spotify Legitimate 2 Stimuli



Figure D.3: Spotify Legitimate 2 Search Result

YouTube Video Description: No speaker is shown, but a phone screen is displayed and there is a voiceover. The narrator directs people to the Spotify app which shows options for Premium plans, along with one month for free for verified students. Other offers include three months for free with a connected PayPal account for payment and six months for free with a connected Walmart+ account. Additionally, Spotify will give Premium for free for one month with each of its offered plans. Finally, the video closes with the speaker scrolling through the Spotify Premium plans, stating that they are affordable and can be purchased.
Website Video Description: The Spotify website is shown on the Spotify Premium page. There are free trial offers for all account types, which include: Individual, Duo, Family, and Student. The "Get Started" button is selected, which goes to the Spotify sign-in page.

Spotify Legitimate 3 Stimuli

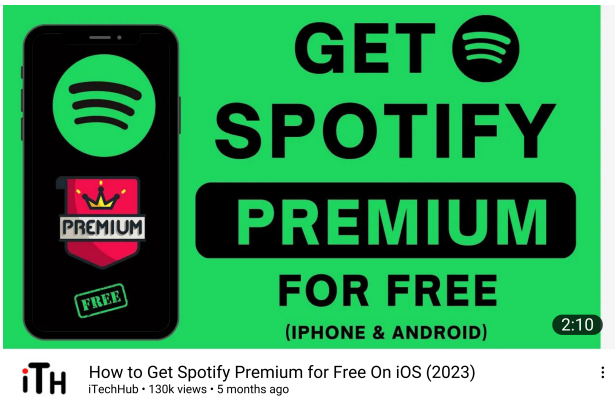


Figure D.4: Spotify Legitimate 3 Search Result

Youtubee Video Description: The video shows a phone screen and no person, but there is a voiceover. Initially, the speaker asks for people to watch the full video to not miss any steps and indicates that this is being done on an iPhone. They describe the perks of Spotify Premium, such as the ability to download music or not hear any advertisements. The speaker mentions that they are using the link they put in their description, which goes to a page showing an offer for six months of Spotify Premium through OnePlus. After logging in, they select the free trial and enter their payment details. The video ends with the speaker indicating that Spotify Premium will cost money after the trial ends, and that they are available to answer questions in the comments section.
Website Video Description: The video shows the Spotify page with the OnePlus offer at the top, as shown in the associated YouTube video. The "Got Questions" section on the website is shown, which has questions about Spotify Premium. The option to start a free trial is selected, which goes to the Spotify log-in page.

Spotify Scam 1 Stimuli



Figure D.5: Spotify Scam 1 Search Result

YouTube Video Description: A younger male in a mask in hoodie with a British accent (and possibly a disguished voice) discusses a website called Hot Shot Codes and claims that Spotify will hate him for sharing this website with the viewers. On the website, he scrolls past several offers for other platforms and selects the Spotify option, which reveals options for six months of Spotify Premium or a one hundred dollar credit for Spotify, both for free. He indicates that human verification is needed to unlock the codes, which requires the completion of surveys or downloading apps. He completes these tasks off-camera and states that the viewers must enter their real information so the website does not identify them as robots. At the end of the video, the code is shown but not entered into Spotify; he also offers to answer questions or help people in the comments.
Website Video Description: The video shows the Hot Shot codes website, where the screen is scrolled past many offers before finding the Spotify option. The same two offers for Spotify Premium are shown as were shown in the associated YouTube video. The offer for six free months of Spotify Premium is selected, and the human verification screen is shown. On the human verification screen, the following tasks are shown: entering an email and completing other steps to obtain either a $1000 Visa gift card or $750 in a Venmo account, downloading iHeartRadio and listening to a radio station on the app after signing up, or getting a $750 Starbucks gift card after completing steps which are not visible in the video.

Spotify Scam 2 Stimuli

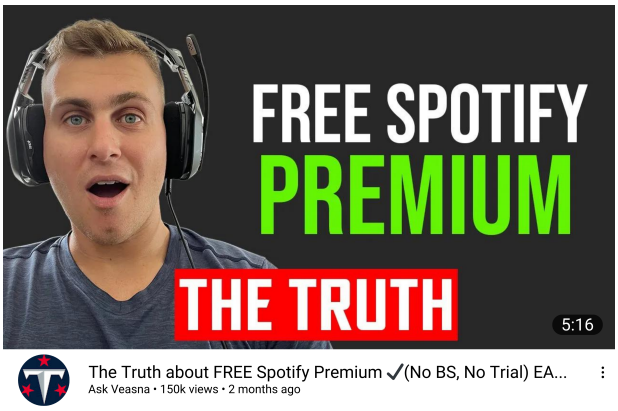


Figure D.6: Spotify Scam 2 Search Result

YouTube Video Description: A younger man with a British accent discusses obtaining an unlimited number of codes for one month free of Spotify Premium on Boujee Coupons, although he also mentions there are only 2,000 codes left. He shows that he is currently on the Spotify Premium Family plan, and indicates that the following steps must be completed on a mobile device. After going to Boujee Coupons and scrolling through other options, he searches for Spotify and shows that the code is locked. To unlock the code, one must download two to three apps they have not downloaded previously and reach certain requirements within the apps. The completion of these tasks is done off-screen, and the unlocked code is shown at the end of the video but not entered into Spotify; the speaker also offers to help people in the comments.
Website Video Description: The Boujee Coupons website is shown, and Spotify is entered into the search bar to find the offers available for Spotify. There are several free offers available, including: a one hundred dollar credit for Spotify, a 50 discount on Spotify Premium, a free one month trial of Spotify Premium, and a 70 discount on Spotify Premium Family plans. The one hundred dollar credit offer is selected, and a progress bar for unlocking the code almost goes to one hundred percent before the final unlock step is shown. There is a tutorial video on how to unlock the codes along with a list of apps to download to complete the unlock process. It is notable that the man in the tutorial video is the same man shown in the associated YouTube video.

Spotify Scam 3 Stimuli

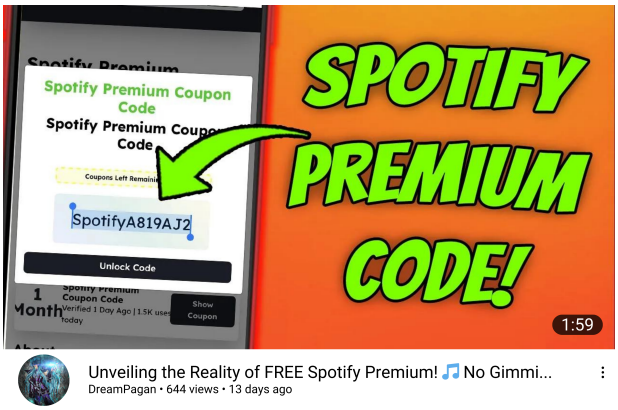
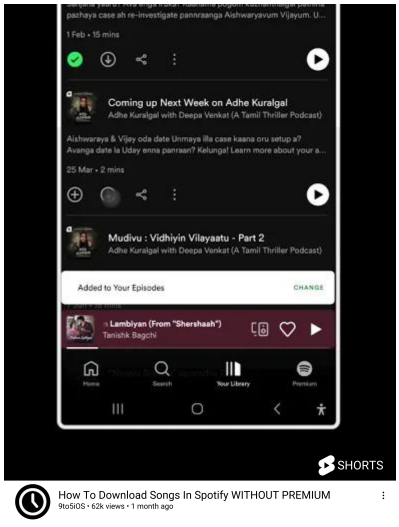


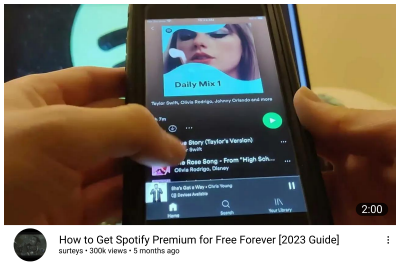
Figure D.7: Spotify Scam 3 Search Result

YouTube Video Description: No person's face is shown, but the camera shows a phone screen that proves the presenter has Spotify Premium. An AI-generated voiceover narrates the steps required to obtain free Spotify Premium, which starts by telling the viewer that this can only be done on a mobile device. Then, viewers are directed to Lucid Coupons, which shows multiple offers for other Platforms before going to the Spotify section to obtain a code for a free month of Spotify Premium. The code must be unlocked by completing two tasks that could differ based on location or time of day accessed. Additionally, these tasks cannot be associated with apps that one has already downloaded. The video closes with the completion of these tasks being done off-screen and an unlocked code shown but not entered into Spotify.
Website Video Description: The video shows the Lucid Coupons website, where the screen is scrolled until Spotify is found. An offer for one month free of Spotify Premium is shown, and the "Unlock Code" button is clicked. A progress bar gets to one hundred percent and then shows the page with tasks, which has a tutorial video that has a similar setup as the associated YouTube video. There are tasks for several apps shown, each with different requirements for completion.

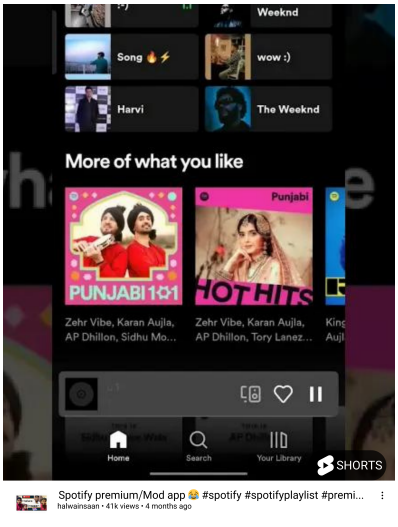
Spotify Clickbait Stimuli



(a) Spotify Clickbait 1



(b) Spotify Clickbait 2



(c) Spotify Clickbait 3

Figure D.8: Spotify Clickbait Search Results

D.5.2 Roblox Stimuli

Roblox Legitimate 1 Stimuli



Figure D.9: Roblox Legitimate 1 Search Result

YouTube Video Description: A younger man is shown initially showing off Robux gift cards, and then mentions how Roblox and Microsoft Rewards partnered up to allow people to redeem free Robux. To earn points on Microsoft Rewards, viewers are instructed to use Microsoft Bing to search. The correct offer to redeem is shown on Microsoft Rewards and the partnership is shown on the official Bing website. The video closes out with the presenter giving free Robux codes.
Website Video Description: The video shows the Microsoft Rewards page, which indicates that every search on Bing will provide five points and every dollar spent at the Microsoft Store will earn one point. To earn points, one must sign in with a Microsoft account. Other options for using the rewards points include donating them to charity. Additional ways to earn points include daily or weekly quizzes and daily offers. The video closes by starting the process to earn points, which requires a Microsoft account sign-in.

Roblox Legitimate 2 Stimuli

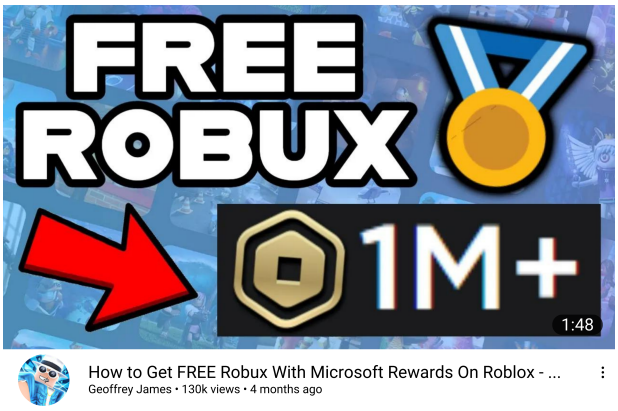


Figure D.10: Roblox Legitimate 2 Search Result

YouTube Video Description: There is no person shown in this video but there is a voiceover and a Roblox avatar. They state that the link in their description goes to Microsoft Rewards. To join, viewers are instructed to sign up or log-in using a Microsoft account and the various methods of earning points through daily tasks and other offers are shown. In the middle of the video, a message stating "Jesus Loves You <3" appears. Additionally, it is mentioned that viewers can obtain a streak by completing tasks every day, which also gives more points. To redeem points, viewers must go to the "Redeem" portion of the Microsoft Rewards website, look up "Roblox", and select the amount of points they would like to spend. The video closes with the presenter telling viewers to watch a playlist to learn how to "get rich" on Roblox.
Website Video Description: *same as in Roblox Legit 1*

Roblox Legitimate 3 Stimuli



Figure D.11: Roblox Legitimate 3 Search Result

YouTube Video Description: There is no presenter in the video, but there is a voiceover and Roblox gameplay shown on the screen. The narrator claims that there is a website "giving out millions of free Robux" and that it "only takes one second" to obtain the Robux. They attempt to persuade viewers to like and subscribe by betting their Robux that viewers cannot like and subscribe within three seconds with their eyes closed. Then, the video tells viewers about Microsoft Rewards and instructs them to redeem a digital code for Robux by selecting the desired amount and hitting the "Redeem" button.
Website Video Description: *same as in Roblox Legit 1*

Roblox Scam 1 Stimuli



Figure D.12: Roblox Scam 1 Search Result

YouTube Video Description: There is no person shown in the video, but there is a voiceover, which begins the video by claiming that thousands of free Robux can be earned in minutes. To earn Robux, the speaker states that viewers can play games, download apps, or complete other offers. In order for the Robux to be transferred, a Roblox account username must be entered onto the website and tasks must be completed. To cash out, one must enter the amount they would like to transfer. The video ends with a code given for viewers to earn Robux for free.
Website Video Description: The video shows the website mentioned in the associated YouTube video, which begins by showing the number of people who have joined. There is a section of the website explaining how the process works, which shows the same steps discussed in the YouTube video. Then, the process is started, and a Roblox username is entered. After this step is completed, a page comes up explaining that the website is moving to a new name, from CollectBux to LootBux. The video ends by scrolling down and showing the FAQs about this new website.

Roblox Scam 2 Stimuli



Figure D.13: Roblox Scam 2 Search Result

YouTube Video Description: There is no person in the video, but a potentially AI-generated voiceover initially narrates over videos of a Roblox avatar and memes. Viewers are directed to RBXHeist and the speaker tells viewers to read the home page to learn more about how the process works. Only a Roblox username needs to be entered, and the speaker creates a new Roblox account to be "safer" when interacting with the website. Methods of earning Robux include following the social media pages of the website, playing games, and doing surveys. The speaker does tasks off-screen and states that they received one hundred Robux; they redeem fifty Robux and show their Roblox transaction history as proof, although their account has a one hundred Robux credit pending. The video closes with the speaker telling people to visit the comments or the description for the website and also gives a code on the website to earn two Robux; they also ask for viewers to comment on the video and subscribe to the channel.
Website Video Description: The video begins on RBXHeist by scrolling through the section that describes how the process works. The website claims that the process is legitimate because no information is required and that the company is paid by their partners. Additionally, visitors to the website can join the Discord to receive help. The video then shows someone entering a Roblox username to create an account, and the website then shows a page with tasks to complete to earn Robux. These tasks include sections such as following social media accounts, referring others, redeeming codes, entering giveaways, and completing other tasks. An offer listed as giving the most amount of Robux is selected and goes to a page with more tasks to complete.

Roblox Scam 3 Stimuli

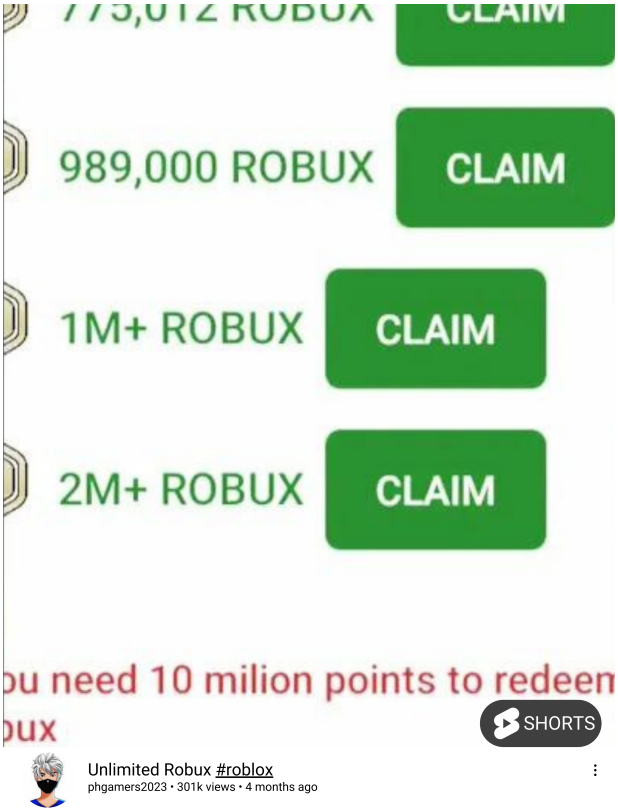
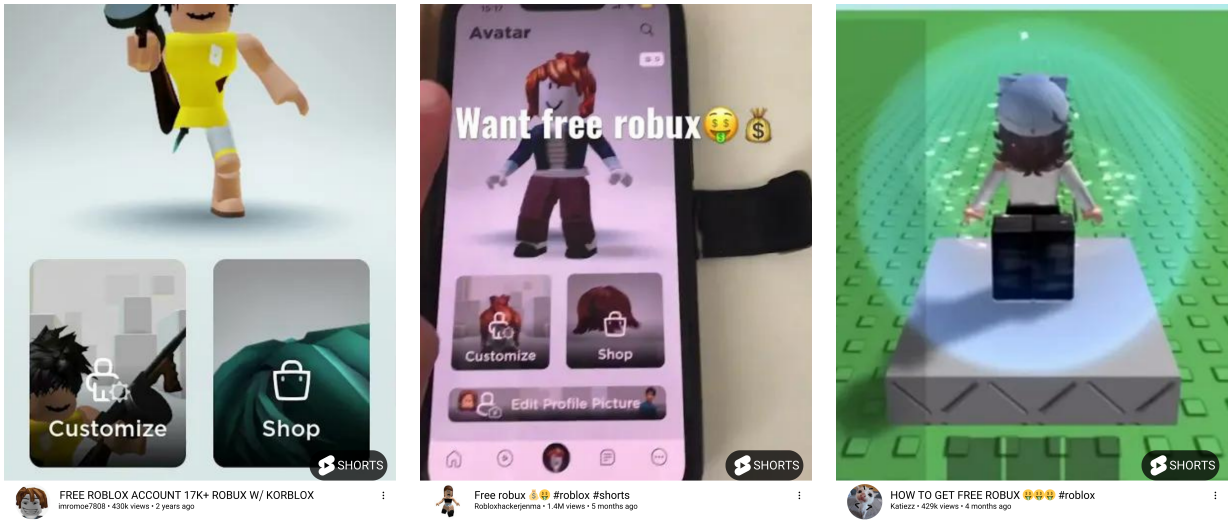


Figure D.14: Roblox Scam 3 Search Result

YouTube Video Description: There is no person or voiceover, only music playing in the background. A screen that says "Give Bux Loading" is shown with the Robux symbol. The next screen shows large amounts of Robux with "Claim" buttons shown next to them.
Website Video Description: A website called PHGamers is shown, claiming to provide free Robux. The video shows the screen scrolling past the website publisher's social media accounts. Throughout the video, various "Download" buttons are shown, and there is an area of the screen called "Click Bux" where a visitor can enter their account information. There are more "Download" buttons on the page, and one is selected, going to another page with various tasks to complete to unlock offers.

Roblox Clickbait Stimuli



(a) Roblox Clickbait 1 (b) Roblox Clickbait 2 (c) Roblox Clickbait 3

Figure D.15: Roblox Clickbait Search Results

Appendix E

The Nature and Quality of Reddit Discussion about Scams

E.1 Codebook

E.1.1 Post Type

This category describes the apparent purpose of a post. Only a single one of these codes is assigned

Irrelevant: Post is off-topic or incomprehensible. Includes spam.

Media Missing: Post links to images, videos, news articles, etc. that are not available. Does not include cases where the user links to the scam website and it has since been taken down. Does not include irrelevant posts.

Scambaiting: Post describes interacting with a scammer to waste their time or enact vigilante justice. Includes examples where users ask questions about how to scambait, posts where users request that others spam or report a scammer.

Report of Victimization: User shares their experience of being victimized by a scam without asking for advice or emotional support. If the post includes a question about whether something is a scam, code it as "Scam identification." If the post includes a question, code it as "Request for support."

Scam Identification: User describes a situation/shows media with the goal of determining whether something is a scam.

Request for Support: User asks for advice, information, or emotional support related to scam victimization, scam communications, or scam prevention. If the post also includes a question about whether something is a scam, code it as "Scam identification."

Discussion of Scam Communications: Post contains text, media, or other descriptions of communication from a scammer. The user seems to think the communication they encountered was fraudulent. They have not fallen victim to the scam (i.e., they have not been deceived into providing sensitive personal information or money).

Scam Related Media: User shares media related to scams that does not convey something they or a friend/relative experienced (e.g., YouTube videos, news reporting, documentaries, etc.).

Other: Post seems to be about scams, but does not fall into one of the above categories. Often, discussion of scams in general.

E.1.2 Technologies & Communication Methods

This category describes what technologies are used to facilitate social engineering or otherwise conduct the scam. Include all technologies that are used or mentioned by a scammer, unless it is obvious that the technology is not actually being used (e.g., if they spoof a PayPal email, non-crypto payment processing is not involved). Where a service has multiple components, the relevant component should be selected. For example, if someone found an item on Facebook marketplace, the initial communication method should be described as "Ecommerce website/service."

Due to space constraints, the paper only discusses initial communication methods; however, all posts were also assigned separate codes to describe other technologies used.

- **Unclear:** Post does not mention what technologies were used to initiate/facilitate a scam.
- **Multiple Stories:** Post involves multiple fraud victimization experiences. Do not code other categories for this column.
- **Websites & Online Services:**
  - **Ecommerce Website/Service:** Scam involves websites used to sell goods and services online, such as Amazon, eBay, Etsy, Facebook Marketplace, Uber, etc. Includes any messaging that occurs on these platforms. Includes potentially fraudulent ecommerce websites.
  - **Social Media:** Social media services such as YouTube, Facebook, Instagram, etc. are used as part of the scam. Includes cases where the scammer uses social media to obtain information about the victim (e.g., photos of their friends). Messaging on social media platforms and dating-based social media services are coded separately. If someone says something like "they contacted me on social media," assume this code applies.
  - **Dating Service:** A social media service dedicated to dating services such as Tinder, Grindr, Bumble, etc. Includes potentially fraudulent dating services.
  - **Job Board:** Websites or services like Indeed, CareerBuilder, etc. that are dedicated to finding or sharing employment opportunities.
  - **Classified Website:** Websites like Craigslist that are used to post general classified advertisements.
  - **Other Website or Online Service:** Another website or online service that doesn't fit into the categories above. Most often, this will be a website designed for a scam (e.g., a fake bank login page for phishing, a fake business webpage, etc.).
- **Communications Technologies:**
  - **Phone Call:** Calling via traditional land-line telephones, mobile phones, or VoIP services. If a user states that they were called or called someone without specifying the platform/medium, assume it falls into this category.

    - **Video Call:** Calling via FaceTime, Zoom, Skype, etc., where video is transmitted alongside audio.
    - **Email:** Messages via email, including the mention of specific email providers (e.g., Gmail, Hotmail, etc.).
    - **Instant Messaging:** Messaging via SMS or other comparable internet technologies (e.g., iMessage, RCS, etc.). Does NOT include messaging on social media or messaging platforms with social media features (see Social media messaging). If a user states that they "texted" or "messaged" someone without specifying the platform, assume it falls under this category.
    - **Social Media Messaging:** Messaging via social media platforms (E.g., Facebook Messenger, Twitter Direct Messages, etc.). Also applies to messaging platforms like Discord, Telegram, WhatsApp, and Snapchat that have social media features like a directory and stories.
- **Payment Methods:**
    - **Crypto Assets:** Scam involves the use of cryptocurrencies (e.g., Bitcoin, Ethereum, USDT, etc.) or other cryptocurrency-based technologies (e.g., NFTs, smart contracts, etc.).
    - **Non-Crypto Digital Payment Processing:** Scam involves PayPal, Venmo, Zelle, or other similar online payment processors that are not cryptocurrency. Doesn't include checks, credit/debit cards, or ACH transfers.
    - **Gift Cards:** Any gift card (e.g., Apple gift cards, Amazon gift cards, Walmart gift cards, etc.).
- **Analogue Communication Methods:**
    - **Mail:** Scam involves material sent through the mail, including commercial parcel services (e.g., UPS, FedEx, DHL, etc.).
    - **In-Person Interaction:** Scam involves communicating with a person face-to-face.
- **Other Technologies:**
    - **Remote Access:** Scam involves software that is used to remotely control or view a device. Typically employed in customer service scams.
    - **File Download:** Scam involves downloading a file, such as an email attachment or software package. May or may not be malware.
    - **Other technologies:** Scam involves other technologies or communication methods not listed above.

E.1.3 Type of Scam

This category captures the distinctive features of the scam described by a post. On `/r/Sextortion`, we assumed posts met the definition of financial sextortion, unless other details were provided that contradicted this assumption.

- **Likely not a scam:** The post describes something that seems not to be an illegal scam. May be another type of crime or just business practices that the user finds distasteful.
- **Multiple stories:** The post involves multiple fraud victimization experiences or attempts at fraud. Do not assign other codes if this code applies.
- **Unclear:** The post lacks enough information to categorize the type of scam. Paired with other codes where it is likely to be a particular type of scam, but there isn't enough information to be sure.
- **Imposter Scams:**
  - **Phishing:** Post describes/shows communications that falsely purport to be from a trusted entity that directs people to provide sensitive information. Only applied where a user is explicitly directed to provide information.
  - **Government imposter:** Post describes/shows communications that falsely purport to be from a government entity (i.e., police, the Internal Revenue Service, USPS, etc.).
  - **Corporate imposter:** Post describes/shows communications that falsely purport to be from a trusted company or service (i.e., a bank, a social media service, etc.).
  - **Customer support scam:** Post describes a scam where the victim is led to believe that they are communicating with tech support or customer support for some business. They are convinced to share access to a computer, provide information, or pay for a service. Almost always paired with "Corporate imposter."
  - **Public figure imposter:** Post describes/shows communications that falsely claim to be from a prominent person, like a celebrity, politician, or social media influencer.
  - **Family/friend imposter:** Post describes/shows communications that falsely claim to be from a friend or family member. May come from a hacked account that belongs to that person.
  - **Other imposter:** Post describes/shows communications that impersonate a different type of person/entity than listed above (i.e., an employee or customer).
- **Commerce:**
  - **Ecommerce seller:** Post describes an apparent scam involving a scammer who is the seller of goods or services online. Often paired with "Non-delivery" or "Fake/defective product."
  - **Ecommerce buyer:** Post describes a scam involving a scammer who is the purchaser of goods/services. Often paired with "Non-payment" or "Fake payment."
  - **Non-delivery:** Post describes a scam where the victim pays for a product/service but never receives it. Almost always paired with "Ecommerce seller."
  - **Non-payment:** Post describes a scam where the victim sells a product/service but never receives payment. Almost always paired with "Ecommerce buyer."
  - **Unauthorized subscription:** Post describes a scam involving a subscription that is hidden, unrequested, or impossible to cancel.

- **Fake or defective product:** Post describes a scam where the victim receives or is offered a product that is defective, counterfeit, or otherwise not as advertised.
- **Non-internet commerce:** Post describes an apparent scam involving a scammer who is the seller of goods/services using phone, mail, or in-person interaction.

- **Opportunity Schemes:**
  - **Relationship scam:** A scam characterized by building a false relationship (typically romantic) with a victim in order to manipulate and/or steal from the victim. Does not apply to shorter-term scams that do not involve a long-term relationship (i.e., sextortion).
  - **Advance fee:** Post describes a scheme where a victim is asked to pay a fee before they can receive something of value (i.e., a prize, investment windfall, inheritance, housing, etc.). May be paired with codes like "Prize scam", "Investment scam", etc.
  - **Rental scam:** Post describes a scam where a victim is offered a short-term or long-term rental. It may be paired with "Advance fee."
  - **Investment scam:** Post describes a scam where a victim is offered an investment opportunity, usually with absurdly high returns. The investment is not real or otherwise fraudulent (like a Ponzi scheme).
  - **Prize scam:** Post describes a scam where the victim is led to believe that they have been selected to receive a prize, grant, inheritance, or other windfall. It may be paired with an "Imposter scheme" and/or involve an "Advance fee" or "Fake payment." Includes “sugar mama” scams.
  - **Employment scam:** Post describes a scheme where a user is offered a fake job. May involve performing real tasks as a ruse to obtain money or information. User may be an unknowing accessory to crime (i.e., sending/receiving stolen funds). Includes cases where someone is commissioned to produce a good, like art.
- **Threat-based Schemes:**
  - **Fake warning:** Post describes or shows a pop-up, notification, or other communication that is meant to mimic OS/software warnings.
  - **Fake debt:** Post describes a scam where the victim is led to believe that they owe a debt that they must pay. May be paired with "Government or Corporate imposter."
  - **Sextortion:** Post describes a scam where the victim is threatened with the releasing of explicit media featuring the victim. The material may actually exist or be fictitious.
  - **Sympathy scam:** Post describes a scam where the victim is led to believe that the scammer is facing some adverse personal circumstances (i.e., illness, imprisonment, etc.) and requires money. May be paired with "Family/friend imposter."
  - **Other threat:** Post describes another scam where the victim is threatened with some harm (i.e., arrest, physical harm, etc.) in order to convince them to pay. This code does not apply where another, more specific category applies.
- **Other ruses:**

- **Pin scam:** Post describes a scam where a scammer attempts to get a user to provide a verification code sent via instant messaging, phone call, or via mail. This is usually framed as wanting to verify a user's identity. Most often, this is a Google Voice authentication code, used to falsely verify possession of the victim's phone number. Does NOT include cases where the code is an MFA authentication code. Often paired with "Ecommerce buyer."
- **Fake payment:** Post describes a scheme where a victim is led to believe they have been paid (most often via a fake check). Many variants of the scheme involve a victim receiving an apparent overpayment and being asked to send part of the money back.
- **Wrong number:** Conversation initiated by a person posing as an individual who has mistakenly texted the wrong account. Often not enough information to tell if it's a genuine wrong number or a scam.

E.1.4 Type of Harm

This category was used to define the types of harm users reported.

- **None:** User does not state that they experienced financial or privacy-related harms.
- **Unclear:** User seems to describe harm, but it's not specified what type (e.g., they say they were scammed, but don't say what that means). May be paired with other harm codes if one or more types of harm seems likely, but is not clearly specified.
- **Exposure:** User describes intimate imagery being shared publicly without permission.
- **Account compromise:** Attacker gains access to one or more of a victim's online accounts.
- **Financial harm:** Victim is financially harmed. Includes loss of money, loss of goods, or uncompensated labor.
- **Financial charges reversed:** The user lost money but had the charge successfully reversed.
- **Personal information disclosed:** User describes that a scammer obtained their personal information (see E.1.7).
- **Intimate imagery shared:** It is implied by their post or explicitly stated that the scammer has intimate imagery of the victim.

E.1.5 Type of Support Requested

This category was used to define the types of support requested by users in posts.

- **Understanding of victimization experience:** Poster requests information or clarification about how a scam they were victimized by works.
- **Other users' experiences:** Post/comment requests for users to share their personal or similar experiences.
- **Scam prevention advice:** Poster requests advice on blocking scam communications or otherwise avoiding scams.

- **How to report scam:** Poster requests advice on reporting a scammer or scam communications.
- **Financial recovery advice:** Poster requests advice on recovering funds after being victimized by a scam.
- **Advice on collecting evidence:** Poster requests advice on collecting evidence about their scam experience (i.e., tracing IP, identifying a photo, finding posted nudes, etc.).
- **General scam remediation advice:** Poster requests advice on other aspects of remediation. May be specific or general (i.e., what should I do?).
- **Reassurance about scam outlook:** Poster seeks reassurance about whether or not they will be fine after being victimized by a scam. Often relates to wanting assurance that the steps they have taken are sufficient.
- **General emotional support:** Poster seeks emotional or therapeutic support connected to scam victimization.
- **Request to DM:** Poster requests to DM other users. This code only occurred in comments.

E.1.6 Remediation Steps and Other Advice

These codes describe the advice given to users by commenters. These codes were also applied to describe the steps taken by the posters of requests for support, but this was not described in the paper due to space constraints.

Remediation is defined as the actions taken to restore the status quo after scam victimization, including technical and non-technical actions. Security/scam avoidance advice is defined as actions meant to improve one's security, which may be employed after scam victimization but are more generally beneficial.

- **Scam Remediation Advice:**
  - **Contact financial institution/payment processor:** Recommendation or remediation step of reporting the crime to a financial institution/payment processor in order to lock a payment card, get a replacement card, reverse the charge, or otherwise report the crime.
  - **Contact service provider:** Recommendation or remediation step of contacting a non-financial service that was involved in perpetrating the scam (i.e., social media company, ecommerce platform, etc.) to report a scam. Includes moderators or admins of a community). When someone says ''Report" without explicitly specifying an entity, assume this code applies unless that contextually doesn't make sense.
  - **Contact law enforcement:** Recommendation or remediation step of contacting law enforcement (i.e., local police, FBI, etc.) to report the scam or otherwise aid in remediation.
  - **Contact recovery service:** Recommendation or remediation step of contacting a paid service that claims to help recover lost money. Seems like these are mostly scams.

- **Contact credit bureau:** Recommendation or remediation step of freezing credit, placing an identity theft warning, or otherwise contacting a credit bureau related to a scam.
- **Contact other government entity:** Recommendation or remediation step of contacting a non-law enforcement government entity like the Federal Trade Commission.
- **Contact NGO:** Recommendation or remediation step of contacting a non-governmental organization connected to scams, such as StopNCII.
- **Contact lawyer:** Recommendation or remediation step of contacting a lawyer to aid in remediation of a crime. Includes cases where someone suggests that they want to initiate legal proceedings.
- **Contact scammer for redress:** Recommendation or remediation step of contacting the alleged scammer to obtain redress, such as a refund. Typically in the ecommerce context.
- **Contact personal connection(s):** Recommendation or remediation step of discussing the crime with a friend, family member (i.e., parent), or other trusted individual (i.e., school counselor).
- **Contact (Unclear):** Recommendation or remediation step of contacting a type of entity that is unspecified or unclear.
- **Write review/report to crowdsourced anti-scam list:** Recommendation or remediation step of leaving a review for a scammer or otherwise reporting to a crowdsourced anti-scam resource.
- **Cut-off communication:** Recommendation or remediation step of blocking a scammer, deleting communications, or otherwise ignoring their attempts at communication.
- **Disable/delete social media account:** Recommendation or remediation step of deleting or otherwise disabling social media accounts.
- **Change identifier:** Recommendation or remediation step of changing phone number, username, email address, or other identifier.
- **Change password:** Recommendation or remediation step of changing password to account(s).
- **Change security/privacy setting:** Recommendation or remediation step of changing security or privacy settings for a social media account such as make the account private, make the follower list private, disable DMs, enable a PIN, etc.
- **Collect/preserve evidence:** Recommendation or remediation step of obtaining/preserving evidence to aid in remediation.
- **Other remediation step(s):** Other remediation step(s) that the poster takes or are recommended.

- **General Security or Scam Avoidance Practices:**

- **Enable MFA:** Recommendation or remediation step of enabling 2FA or other MFA technology.
- **Don't send nudes:** Recommendation to not take or share intimate pictures/videos/video calls with others online.
- **Don't click links:** Advice not to click on links from communications.
- **Don't answer communications from unknown sources:** Advice not to answer calls, texts, etc. from unrecognized sources.
- **Use anti-virus/security scan:** Recommendation or remediation step to use anti-virus/other security software on a device.
- **Use unique passwords:** Recommendation or remediation step to use different passwords on each account.
- **Watch out for recovery services/scams:** Recommendation to avoid any service/individual that claims to help recover lost money or otherwise remediate scams.
- **Don't pay scammer:** Recommendation to not pay/stop paying a scammer. Includes cases where the commenter advises the poster that paying was a bad idea after they already paid.
- **Be careful:** Non-specific advice to be careful or avoid scams.
- **Advice on helping friend/relative identify scam:** Recommendation on how to convince someone that something is a scam.
- **Advice on ecommerce payment:** Advice about how to accept/send payment in ecommerce. Includes recommendations on using a particular platform for sales.
- **Advice on where to shop**: Recommendations about how to avoid scams when online shopping, such as doing research, avoiding unknown sellers, looking at DNS, etc.
- **Other Security/Scam avoidance advice:** The post or comment includes other advice for technical security or avoiding scams not covered by the above codes.

E.1.7 Types of Personal Information

This category is used to describe the type of information disclosed to a scammer. Includes only that which is explicitly stated to have been given to the scammer.

- **Unclear:** Poster says PI was exposed, but it's unclear what it was.
- **Name:** Person's real legal name, in-part or whole.
- **Location:** Person's address or location, in-part or whole.
- **Account identifier:** Username or other method of identifying the account.
- **Card number:** Credit or debit card number.
- **Bank account number:** Account number for bank.
- **Bank identifier:** Bank name or routing number.
- **Password:** Password to one or more accounts.

- **OTP:** A code used for authentication, usually sent to email or text.
- **Date of birth:** Person's date of birth.
- **Phone number:** Person's phone number.
- **Email:** Person's email.
- **Non-explicit photo:** An identifying photo.
- **Tax ID:** Social security number or other similar tax ID.
- **Part of tax ID:** Part of SSN or other similar tax ID.
- **Identity documents:** ID, passport, social security card, birth certificate or other similar identity documents.
- **Other information:** Other information not covered above.

E.1.8 Other Comment Types

This category is used to characterize other aspects of comments that do not relate to security/remediation advice for the poster

- **Deleted:** The comment was deleted or removed before collection by PushShift/ArcticShift.
- **OP:** The commenter is the original poster.
- **Clarification question:** The commenter asks a clarification question about the someone's experience.
- **Additional detail:** Commenter is the original poster, who provides additional detail about their experience.
- **External resource:** User points to an external resource. Includes looking at other posts on the same subreddit. (i.e., website) to provide advice.
- **Explain scam:** The commenter attempts to explain the type of scam and/or its mechanics. Includes cases where they might be wrong. Also includes cases where user just states that situation is a scam without explaining the type. May be paired with "Reassurance. "
- **Similar experience:** Comment provides an example of a similar experience to the scam provided.
- **Non-OP request for support:** Commenter is a person other than original poster who is requesting support.
- **Reassurance:** The comment provides reassurance that poster will be OK or other emotional support.
- **Chastise:** The comment criticizes or makes fun of poster for falling for scam.
- **Money is gone:** The comment tells the poster that their money is gone and recovery is unlikely/impossible.
- **you can DM me:** The comment tells the user that they can/should DM the commenter, typically for emotional support.

- **Automod:** Comment is written by the Automoderator. Other codes are not applied to these comments.
- **Other advice:** Commenter provides other advice not related to scam remediation or security/scam avoidance.
- **Other/irrelevant comment:** The comment is not advice and doesn't fit in any of the other categories. Includes general discussion about scams unrelated to a victim's experience and off-topic content.

Appendix F

Investigating User Reasoning with Context-Specific SMS Scam Warnings

F.1 Demographic Survey

Thank you for agreeing to participate in our study. The following demographic questions are optional but will be used to ensure that we recruit a diverse group of participants.

1. How do you describe your gender identity?
   (a) Man
   (b) Woman
   (c) Non-binary
   (d) Prefer to self-describe *[text box]*
   (e) Prefer not to disclose
2. What is your age in years?
   (a) 18 to 24
   (b) 25 to 34
   (c) 35 to 44
   (d) 45 to 54
   (e) 55 to 64
   (f) 65 or older
   (g) Prefer not to disclose
3. What is your race? You may select multiple options
   (a) American Indian or Alaskan Native
   (b) Asian
   (c) Black or African American
   (d) Native Hawaiian or Other Pacific Islander

(e) White

(f) Prefer to self-describe *[text box]*

(g) Prefer not to disclose

4. Are you Hispanic or Latino?
   (a) Yes
   (b) No
   (c) Prefer not to disclose
5. What is your approximate household income?
   (a) Less than $25,000 per year
   (b) $25,000 to $50,000 per year
   (c) $50,000 to $100,000 per year
   (d) $100,000 to $150,000 per year
   (e) More than $150,000 per year
   (f) Prefer not to disclose
6. What is the highest level of school you have completed or the highest degree you have received?
   (a) Less than high school degree
   (b) High school graduate (high school diploma or equivalent including GED)
   (c) Some college but no degree
   (d) Associate degree in college (2-year)
   (e) Bachelor's degree in college (4-year)
   (f) Graduate certificate or degree (e.g., MS, PhD, JD, etc.)
   (g) Prefer not to respond
7. Do you have a formal education in a computer-related field, such as computer science or IT? ("Formal education" could mean a completed degree or certificate, or classes or trainings you took towards a degree or certificate.)
   (a) Yes
   (b) No
   (c) I'm not sure
   (d) Prefer not to disclose
8. Do you have work experience in a computer-related field, such as computer science or IT?
   (a) Yes
   (b) No
   (c) I'm not sure
   (d) Prefer not to disclose

We will now ask for your contact information in order to schedule an interview.

9. What is the best way to contact you to schedule an interview?
   (a) Email
   (b) SMS Message
10. *If the participant selects email:* Please enter the email address that you would like to be contacted at. You will be contacted by [redacted]
    [text box]
11. *If the participant selects cell phone:* Please enter the cell phone number you would like to be contacted at. You will be contacted by [redacted]. Normal messaging rates will apply
    [text box]

F.2 Interview Guide

The semi-structured interview format allowed the interviewer to ask follow-up questions as needed. Text in italics or square brackets was not read verbatim. Suggested follow-ups are included, but were not always asked. As the task progressed, participants would often begin answering questions without being explicitly prompted, so previously answered questions were not re-asked. The recording of the interview began after the introduction section.

F.2.1 Introduction

Thank you for taking the time to participate in this interview. We will begin by reviewing key areas of consent. This interview will be audio-recorded and transcribed for the purpose of our note-taking. Is it ok to audio-record you during this interview? *All participants said yes to this question*

Participation in the interview is voluntary, and you may choose to withdraw at any time. It will take around 60 minutes. The format will start with some background questions about your usage of mobile phones. We will then review various instant messages and discuss how you perceive each message. This task will require you to scan a QR code and then view images of instant messages on your phone. To help us protect your privacy, please do not reveal any private or personally-identifiable information about yourself or others, such as names, in response to our questions. Do you have any questions before we begin? *[Answer any questions]*
Feel free to ask any questions throughout the process.

F.2.2 Background Questions

1. What type of phone do you use on a daily basis?
2. When did you begin using this phone (approximately)?
3. Is this the phone you will be using during the interview today?
4. Who is your mobile phone carrier?

5. What types of things do you use instant messaging for? By instant messaging I mean SMS or RCS, but you may also consider things like WhatsApp or Facebook Messenger.
   (a) **Follow-up:** What types of instant messaging do you do?
   (b) **Follow-up:** Do you ever receive messages from your doctor or healthcare provider? What kinds of messages?
   (c) **Follow-up:** Do you ever receive messages from your bank or other financial institutions? What kinds of messages?
   (d) **Follow-up:** Do you ever receive codes that are used to login or sign-up to online services? What kinds of messages?
   (e) **Follow-up**: Do you ever receive messages from other companies or websites? What kinds of messages?

F.2.3 Task Introduction

During the next portion of the interview, we will review some instant messages and get your impression on how you would respond to them. We are going to show you some text messages that were received by a mobile phone user named Bob who lives in Pittsburgh and frequently receives text messages from delivery services like USPS or FedEx, his bank, which is Bank of America, and other companies. He also receives spam. You're trying to decide how to respond to messages. Once the task begins, you will scan a QR code with your phone and view a set of messages . For each message, you should explain what you think Bob should do in response to the message and why. I will share these instructions with you in the zoom chat so you may refer to them throughout the task. Do you have any questions before we get started? *[Answer any questions]*

Feel free to ask any other questions you may have during this task

F.2.4 Task

[Prompt participant to view Nth message, referring to some salient feature like the phone number or start of text.]

Please take a moment to read the message. Let me know when you are finished.

6. What do you think Bob should do after receiving this message? ***Potential follow-up prompting****: For example, you may recommend he reply, ignore, report, delete, click, etc.*
7. *If participant doesn't justify answer:* Why do you think that's the best response?
   (a) ***Follow-up:*** Were there specific parts of the message that influenced your thinking?
   (b) ***Follow-up (if applicable)*:**This message includes a warning/notice. Did that affect what you think Bob should do?
   (c) ***Follow-up (if applicable)***: Have you seen this warning/notice before?
   (d) ***Follow-up (if applicable)*:** What do you think of the text of the warning/notice?

(e) ***Follow-up (if applicable)***: This warning says that it was generated using AI. Does that impact your opinion of the warning or its content?

(f) ***Follow-up (if applicable)***: Are there any changes you would make to improve this warning/notice? Why/Why Not?

(g) ***Follow-up (if applicable)***: Do you think this warning/notice is better or worse than the others you have seen? Why or why not?

8. How confident are you in your decision/assessment? From 1 to 10.
9. Have you received a message like this before in real life?

Repeat questions for all 12 instant messages.

F.2.5 Post-task questions

Thanks for walking through those messages. Now I'd like to ask a few final questions to get your overall thoughts.

10. You saw several different warnings/notices while completing this task. How helpful were the warnings in making your decisions?
    (a) **Follow-up:** Are there any warnings that stood out as particularly helpful to you? Any that were particularly unhelpful?
    (b) **Follow-up:** Was there anything missing from the warnings or messages that would've helped you decide more easily?
    (c) **Follow-up:** Do you think warnings like these would be helpful in real life? Why or why not?
11. Are there any features (other than warnings) that you think your phone should have to help deal with instant messages from unknown senders?
12. Is there anything else you'd like to share about this experience or about instant messaging in general?

F.3 Stimuli

All formatting within warning text was generated by ChatGPT. Sender identifiers are described for privacy purposes.

F.3.1 Fake Amazon Account Alert

Sender: Email address that contains the word "Amazon" followed by random text and an unrelated domain name.
Text: Unusual Activity : We've suspend your Amazon account due to unusual activity. Login attempt blocked from : Mozilla Firefox 8.7 [168.254.81.174 - Turkey] Our system has suspend your Amazon account for security reasons. To unlock your Amazon account, please verify with

link below: http://www.xprswebsite.com/free/userf111/redirf111/?=5826865 You need to take action within 2 days before account will be suspended. Regards, Amazon Teams
Long Contextual Warning Text: **Poor grammar**, **a suspicious sender email**, and **a non-Amazon URL** suggest this message is a phishing attempt. **Do not click the link**—go directly to the official Amazon website if you're concerned about your account.
Short Contextual Warning Text: Do not click links; verify directly via official app or website.

F.3.2 Fake Bank of America Alert

Sender: US number with mid-western area code.
Text: BofA Alert: We 've placed a hold on your debit/atm card due to some unusual activity, To recover Visit https://profib04.com and verify your identity
Long Contextual Warning Text: The message includes a **suspicious URL that is not affiliated with Bank of America** and has **grammar errors** (e.g., "We 've"). **Do not click the link**—contact your bank through official channels to verify any issues.
Short Contextual Warning Text: Don't click links; contact your bank directly.

F.3.3 Fake Delivery Alert

Sender: US number with west-coast area code
Text: We have detected an issue with your recent shipment; invalid shipping information. Access http://uspsdeliveryrd.ga/ to update your shipping address.
Long Contextual Warning Text: The message uses a **fake USPS-related URL** (".ga" domain) and **creates urgency to trick you** into clicking. **Do not open the link**—check shipment status directly on the official USPS website.
Short Contextual Warning Text: Avoid clicking links; verify shipment status on official site.

F.3.4 Employment Scam

Sender: US number with Pittsburgh, PA area code
Text: Hello, are you looking for a part-time job now? We provide sales growth services for Amazon, and you can get a stable income of $30-200 in just one hour a day. You do not need to pay any deposit or membership fee, regardless of gender, 23-68 years old, and the Commission is paid every day. If you are interested, please add my telegram. t.me/JRE7708 or WhatsApp: [redacted UK phone number]
Long Contextual Warning Text: The message promises **unrealistically high pay for little work**, includes **unsolicited contact links (Telegram, WhatsApp)**, and uses **vague job details**—common signs of job scams. **Do not respond or click any links**; legitimate employers do not recruit this way.
Short Contextual Warning Text: Do not respond or share personal info; offers like this are often scams.
Note: The phone number was arbitrarily selected for this message, as the SmishTank dataset did not include a phone number for this message.

F.3.5 Fake T-Mobile Giveaway

Sender: US number with west-coast area code
Text: This is a gift to express our gratitude for your exceptional payment record. HAPPY 2025! T Mobile POSC9.XYZ/YgSpZZjZoh
Long Contextual Warning Text: The message contains a **suspicious, shortened-looking URL** and **unsolicited reward claims**, which are classic signs of phishing. **Do not click the link**—check directly with T-Mobile through official channels if you receive unexpected offers.
Short Contextual Warning Text: Ignore unexpected gifts and don't click links.

F.3.6 IRS Arrest Threat

Sender: US Phone number with Washington D.C. area code.
Text: WARNING: (Criminal Investigation Division) I.R.S is filing lawsuit against you, for more information call on [anonymous US number with south-eastern area code] on urgent basis, Otherwise your arrest warrant will be forwarded to your local police department and your property and bank accounts and social benifits will be frozen by government.
Long Contextual Warning Text: This message uses **threatening language, urgent legal claims**, and **spelling errors** (e.g., "benifits")—tactics commonly used in government impersonation scams. **Do not call the number**; the IRS does not contact people this way.
Short Contextual Warning Text: Do not call or respond; official agencies don't threaten by text.

F.3.7 FedEx Notification

Sender: 5-digit short code
Text: FedEx: A package from Aarke was delivered 05/13 at 12:08P. See photo and rate delivery: fedex.com/t/[tracking number]/en_US Reply HELP for help. STOP to cancel
False Positive Warning Text: This message is likely a scam because it urges you to click a suspicious link pretending to be FedEx; **do not click links or provide personal information** and verify delivery status directly through official FedEx channels.

F.3.8 Healthcare Reminder

Sender: 5-digit short code
Text: MyUPMC: You have an upcoming appt. at UPMC. Log in to view appt details with MyUPMC. https://myupmc.upmc.com/appointments Reply XRemind to opt out
False Positive Warning Text: This message pretends to be from UPMC but contains a link that could steal your login; **do not click links or enter your credentials**—always access your account directly through official websites.

F.3.9 Amazon OTP

Sender: 5-digit short code

Text: Amazon: Your code is 134091. Don't share it. If you didn't request it, deny here https://amazon.com/a/c/r/[25 character base 64 string]
False Positive Warning Text: This message may be phishing for your Amazon credentials—**do not click the link or enter any codes**, and go to Amazon's site directly if you didn't request a login.

F.3.10 Overstock.com Promotion

Sender: 5-digit short code
Text: Overstock Day is here with our best home deals of the year! TODAY ONLY get 20% off + 2x rewards + free shipping!* https://o.bttn.io/0nvkx2i4Cqp
False Positive Warning Text: This message uses urgent sales language and a suspicious link to lure clicks—**avoid clicking and shop only through official retailer websites**.

F.3.11 Bank of America PIN Alert

Sender: 5-digit short code
Text: BofA Security: Debit card ending in [4 digits] was declined because a PIN wasn't entered. Please try again with your PIN. Reply STOP to end texts.
False Positive Warning Text: This message impersonates Bank of America to create urgency—**do not reply or share your PIN**, and contact your bank directly through verified channels.

F.3.12 Political Solicitation

Sender: US phone number with area code located in the same state as Pittsburgh, PA
Text: Hey Bob, OnePA here! Mayor Ed Gainey is standing up to Trump & billionaires trying to buy Pittsburgh's election on May 20th! Read more here to see how Mayor Ed Gainey is fighting back! https://onepa.org/pittsburgh/?id=403V Stop to End
False Positive Warning Text: This message uses **emotional political language and a suspicious link** to manipulate you—avoid clicking links and verify campaign information through trusted sources.

F.4 GAI Prompts

Long contextual warning prompt: You are to act as a generator of explanations for an SMS scam warning system. You should first decide whether a message looks like a scam or not. If it does, you should generate a short warning. The warning should explain the most important reasons why the message looks like a scam (e.g., URL features, context, etc.) and advise users on what to do. You should bold the most important points from the explanation. It should be no longer than two sentences.
Note: *To generate false positive warnings, GPT was additionally told to assume the text was a scam*

Short contextual warning prompt: You are to act as a generator of text for an SMS scam warning system. You should first decide whether a message looks like a scam or not. If it does, you should generate a very short warning. The warning should advise users on what to do. It should be no longer than a few words or one sentence.

F.5 Codebook

F.5.1 Introductory questions

Type of phone used for interview (Q1, Q3)

- **[Manufacturer Name]**: Open coding for name of manufacturer of phone. In practice, participants used phones from Samsung (15), Google (3), Motorola (1), OnePlus (1).

Length of ownership of phone (Q2)

- **Less than a year**: Participant indicates that they have owned their phone for less than a year.
- **One to three years**: Participant indicates that they have owned their phone for between one and three years.
- **More than three years**: Participant indicates that they have owned their phone for more than three years.

Phone Carrier (Q2)

- **[Provider Name]**: Open coding for the name of the carrier of their phone service. In practice, participants reported using T-Mobile (11), Verizon (3), Xfinity (2), GoogleFi (1), Tello (1), Ultra (1), and Consumer Cellular (1).

Purposes of Texting (Q4)

- **Social**: Participant indicates they use instant messaging to talk to friends/family. May be phrased vaguely.
  E.g., Interviewer: "this will be a broad question, and... I'll follow up on this. But what types of things do you use instant messaging for?..." P1: "Communicating with friends and family."
- **Work**: Participant indicates they use instant messaging as part of their job or for educaiton, if the partiicpant is a student.
  E.g., P3: "text is mostly used when there is like some sort of group project that's involved. So you need to coordinate with a lot of people, and a lot of people don't use Whatsapp. So that's when I use text message."
- **Healthcare**: Participant indicates that they receive instant messages from their doctor or healthcare provider.

E.g., Interviewer: "So do you ever receive messages from a doctor or healthcare provider that are over instant messaging?" P2: "Yes, my primary care office has my number, and I get updates from them."

- **Finance**: Participant indicates that they receives messages related to finance, like credit card alerts or recipts
  E.g., Interviewer: "Do you ever receive messages from your bank or other financial institutions?" P1: "Yes." ... Interviewer: "would you mind sharing like what kind of messages like credit card alerts or things like that? Or?" "P1: Yeah, mainly security alerts if they think like, if I've charged something that's over a spending limit, or if they feel like a charge could be fraudulent."
- **OTP**: Participant indicates that they receive instant messages with OTP for login or registraiton, including for healthcare or banks.
  E.g., Interviewer: "Do you ever receive codes that are used to log in or sign up for online services?" P1: "Yes." Interviewer: "Okay, could you give an example of sort of like a context where that occurs." P1: "I mean, like some of our work applications, they do the like 2 factor. And we get sent a code email clients like Constant Contact or Mailchimp."
- **Marketing**: Participant indicates they receive marketing messages form services like ecommerce websites
  E.g., Interviewer: "and do you ever receive messages from other companies or websites like promotional type things? stuff like that." P1: "Yes, they're the worst."
- **Political**: Participant mentions political messages of some kind
  E.g., "P4: also political messages also often go in my junk folder... like fundraising"
- **Delivery**: Participant mentions package or food delivery notifications
  E.g., [after task introduction] P4: "No questions. But you did remind me that I do also get text messages telling me when packages are delivered."
- **Other**: Participant mentions another purpose for texting not specified by the categories above
  E.g.,P15: Like... I got a message this morning. I'm kind of not sure if it was spam or not... Yeah. Somebody said something. 'Good morning, beautiful souls. The words you speak today, Happy Friday.'... I replied, 'who is this?'... I know numbers are easy enough to spoof, but I figured I'd give them the benefit of that and reach out and see who it was."

F.5.2 Task

Message Actions (Q6)

- **Conditional**: Participant indicates that they may take one of multiple actions depending on conditions. Where this code is used, all stated actions should be included.
  E.g.,P1 @ Fedex Message: "I would probably just delete this message, because if I got a package and I saw the delivery time well, I might wait till I'm home to make sure it's really there, but I probably would not take any action on it, and then make sure I got the package and then delete it later... If I know I have a package coming, this looks valid, and then, if

something was wrong, if I was away from my house. It does say that there is a photo so I could double check if it wasn't there."

- **Interact**: Participant indicates they would click the link in the message, call the number provided, enter the provided MFA code, reenter their pin or otherwise interact directly with the message in a positive manner.
  E.g., P1 @ Fedex Notification: "...If I know I have a package coming, this looks valid, and then, if something was wrong, if I was away from my house. It does say that there is a photo so I could double check if it wasn't there."
- **Don't interact**: Participant explicity states they wouldn't click the included link, call an included number, or otherwise interact positively.
  *E.g., P1 @ Healthcare Reminder: "I would probably go to my My[organization] app **instead of clicking the link.**" (emphasis added to the relevant part of the quote)*
- **Contact service externally**: Participant recommends vising a website, viewing an app, or otherwise contacting a service through methods external to the instant message (e.g., the number on the back of the card, the app for their healthcare provider, searching the site on google, etc.)
  *E.g., P1 @ Healthcare Reminder: "**I would probably go to my My[organization] app** instead of clicking the link." (emphasis added to the relevant part of the quote)*
- **Research**: Participant indicates they would conduct research to verify information within the message
  E.g., P2 @ Political Solicitation: "Sorry. No, there's 1 more thing that I would do. I would look up [organization] on Google"
- **Report**: Participant indicates they would report the message as spam. Includes where they just say "Report" without clearly specifying how.
  *E.g., P1 @ Fake Delivery Alert: "I think Bob should **report this as spam** and block the number." (emphasis added to the relevant part of the quote)*
- **Block**: Participant indicates they would block the sender of the message.
  *E.g., P1 @ Fake Delivery Alert: "I think Bob should report this as spam and **block the number**." (emphasis added to the relevant part of the quote)*
- **Ignore**: Participant indicates that they would ignore the message. *Don't interact* applies when a participant negates an action (e.g., don't click)
  E.g., P12 @ Healthcare Reminder: "Well, I mean, if he knows about his appointment, I would just ignore it..."
- **Delete**: Participant indicates they would delete the message.
  E.g.,P1 @ FedEx Notification: "probably just delete this message, because if I got a package and I saw the delivery time—well, I might wait till I'm home to make sure it's really there, but I probably would not take any action on it, and then make sure I got the package and then delete it later."
- **Other action**: Participant specifies other action not covered by the codes above.
  *E.g., P3 @ Bank of America PIN Alert: "They could try again with their PIN... or **maybe use some other card**" (emphasis added to the relevant part of the quote)*

F.5.3 Message Reasoning (Q7)

Message Judgment

- **Suspicious**: General code for when partiicpant indicates that they are suspicious of the legitimacy of a message or are certain that it is a scam.
  E.g., P8 @ Employment Scam: "As as soon as I read this I'm aware that this is a scam..."
- **Legitimate**: General code for when participant indicates that they believe that a message is legitimate.
  E.g., P1 @ Fedex Notification: "...If I know I have a package coming, this looks valid..."
- **Unclear**: General code for when partiicpant does not indicate whether they think a message is real or when they are unsure whether its real or not. May include some sentences that could be coded as *legitimate* and others that could be coded as *suspicious*.

Message Features

- **Warning (unprompted)**: Participant mentions the warning on the message without being prompted when asked to justify their response.
  E.g., P1 @ Fake T-Mobile Giveaway: "because it's already flagged as spam.... [after follow-up] I mean the alert. And then that says it's a scam.."
- **Lack of warning**: Participant mentions that the lack of the warning effected their decision in some way.
  E.g., P9 @ FedEx Notification: "because... it doesn't come with the warning"
- **Urgency**: Participant mentions the urgency that the message attempts to instill.
  E.g., P4 @ IRS Arrest Threat: [Laughs] oh, oh, my gosh, yeah. So this one I definitely would also report and block... even without the additional information it's creating that sense of urgency.
- **Personal experience**: Participant reasons based on their personal experience with services or how they operate. Must include an explicit mention of personal experience
  E.g., P13 @ IRS Arrest Threat: "I haven't gotten a criminal investigation, but I have had a certified letter from the IRS years ago. So I know what those look like, too."
- **Absurd offer**: Participant indicates that the consideration offered in a message is absurd or "too good to be true"
  E.g., P18 @ Employment scam: "saying, the stable income, you know, making that in just 1 hour day. Now 30 would be a feasible salary, but not likely for a side gig. 200? I am not a CEO or a doctor."
- **Other absurd details**: Other details of the scenario are considered to be absurd or unlikely by the participant.
  E.g., P8 @ Fake T-Mobile Giveaway: "Also it says, 'Happy 2025,' and I was texted this in May."
- **Correct details**: Participant mentins factual information from the message that effects their reasoning.
  E.g., P7 @ Political Solicitation: "the information contained within makes sense. You

know, the election day being what it is, the fact that it's [local politician] against Trump. So it's valid in that, in the context of what they have in there."

- **Unfamiliar organization**: Participant indicates that they are unfamiliar with the sender or organization associated with the message. Distinct from unknown sender as relates participants intrinsic knowledge of the organization discussed, rather than the properties of the message.
  E.g., P1 @ Political Solicitaiton: "I'm not particularly familiar with [anonymous political organization]..."
- **Missing expected details**: Participant indicates that their reasoning is based on the lack of expected details (e..g, the provider name, a tracking ID, etc.)
  E.g., P2 @ Healthcare Reminder: "There are a couple of things that I've just noticed with doctor's appointments that this doesn't have... that detail is literally just, typically it would tell you what time and date your appointment is in the text message."
- **Personalization**: Person mentions that the message includes personal information that makes it feel more individually targetted (e.g., name)
  E.g., P2 @ Political Solicitation: "the... personally identifiable information. They have your name previously. They have your name. They got it right."
- **Linguistic features**: Participant mentions superficial features of the language used in the text message, like grammatical errors, abbreviations used, level of formality, akward phrasing, etc.
  E.g., P1 @ Healthcare Reminder: "just because it looks a little bit funny to me where it's reply XRemind to opt out. like the it looks a little bit like it. It could be fake... [after follow up] there's no space between the X and remind to opt out that part is just funny to me."
- **URL features**: Participant mentions the URL included in the message, including specific parts like the domain name, TLD, etc.
  E.g., P1 @ Fake Delivery Alert: "The URL is a USPS, but it is a.ga, and I think usps is a.gov address."
- **Sender information**: Participant mentions the **features** of the phone number/email in some way, e.g., that the number is/isn't a short code, the area code, etc.
  E.g., P1 @ Political Solicitation: "I know [area code] is not a [anonymous northeastern city] area code."
- **Unknown number**: Participant mentions that the number is not incldued in the contact books or lacks message history. Includes notion that the contact is unsolicited
  E.g., P2 @ Fake T-Mobile Giveaway: "gift for your payment record. I think that's suspicious.... It's an unknown number."
- **Request for action**: Partiipcant mentions that the mesasge is asking for personal information, requesting to click a link, or requesting some other action (e.g., reply/call).
  E.g., P2 @ Fake Delivery Alert: "It's asking for like personal information which typically like, they don't do that over that over this interface..."

- **No action requested**: Participant mentions that the text **does not** request an action of them
  E.g., P5 @ Bank of America PIN Alert: "Well, it's not trying to get me to click on a link or trying to get me to call something."
- **Communication channel**: Participant mentions the communication channels used as being inappropriate (e.g., recommending WhatsApp)
  E.g., P6 @ IRS Arrest "IRS doesn't send text messages. Period end of discussion."
- **Other message feature**: Participant mentions other feature(s) of the message not covered above.
  E.g., P11 @ IRS Arrest Trhreat: "Like at 12 0. 8 Pm? It just seems really off."
- **Other reasoning**: Participant describes other reasoning that is disconnected from the message features.
  E.g., P1 @ Political Solicitation: "Yeah, I feel like... going to a web browser and typing it in like, there's extra layers of security on your computer that would flag ... if it is a spam site more so than what's just on your phone, if you would tap the link."

F.5.4 Warning

Effect of warning on decision (Q7 and and follow-ups)

- **Warning as red flag**: Participant indicates that the warning reduced trust in the message or increased suspicion, sometimes in combination with other message features.
  E.g., P1 @ Fake Amazon Account Alert w/ contextual warning (long, confident framing): "I think the warning is like an initial red flag. But then, also looking at the the start of that email address, and then the URL that they want you to click."
- **Affirmed right decision**: Participant indicates that they would have come to the same decision without the warning, but the warning made them more confident or reassured in their judgment.
  E.g., P2 @ IRS Arrest Threat w/ contextual warning (short, uncertain frmaing): "Without the scam message I might be a little bit worried... I wouldn't actually call them, but it would be a little bit like concerning."
- **Highlighted particular action**: Participant indicates that the warning led them to select a particular option (e.g., because it made it easy).
  E.g., P2 @ Fake Bank of America Alert 2/ Save contact notice "Oh, I guess there is a 'report spam'. I would probably click that. I like it when they make it easy."
- **Highlighted evidence**: Participant indicates that the warning highlighted evidence that factored into their decision.
  E.g., P16 @ Employment Scam w/ contextual warning (long, confident framing): "The pay being ridiculously... generous, for like an hour of work, I might have missed that if I didn't have the information on the message."
- **Ignored**: Participant didn't see the warning or felt it had no effect
  E.g., P8 @ Fake Amazon Alert w/ Generic Scam Warning Interviewer: "This message includes a warning on it. Did that affect your reasoning at all? "P8: "Not really." Inter-

viewer: "Okay." P8: "It's actually kind of too small, so I read the message before I even saw it."

- **Unclear**: Participant expresses that a warning had an effect but its unclear what the exact effect was.

Warning Feedback (Q7 and follow-ups

- **No changes**: Participant indicates that hey wouldn't make any changes to the warning or suggests only extremely superficial changes (e.g., adding a period)
  E.g., P10 @ Spam Warning: Interviewer: "Alright. Are there any changes you think you would make to this warning to make it better—again, talking about the 'why this looks like spam' box?" P10: No.
- **Likes/dislikes color**: Participant likes or dislikes the color of the warning; dislike includes recommendation to change color.
  E.g., P5 @ Fake Bank of America Alert w/ contextual warning (short, uncertain framing) w/ "It stands out because it's red and it's like the 1st thing you see when you open up the message." (likes)
- **Likes/dislikes iconography**: Participant likes or dislikes the icons associated with the warning/notice; includes recommendation to change the icon
  E.g., P1 @ IRS Arrest Threat w/ Spam Warning:"So I noticed, like the little like Google Shield. And I read it because of that... [after many questions] it's also good that it has a little like red. Stop sign thing, too." (likes)
- **Likes/dislikes options**: Participant likes or dislikes the options associated with a warning/notice.
  E.g., P1 @ IRS Arrest Threat w/ Spam Warning: "But it's funny that it's like a not spam button, only not a report option." (dislikes)
- **Likes/dislikes tone of warning**: Participant likes or doesn't like the tone or level of confidence of warning; dislike includes recommendations to change (e..g, should be more confident or direct)
  E.g., P7 @ Fake T-Mobile Giveaway w/ Save cotnact notice:"maybe it influenced it negatively that... they're trying to get you to add [the number] rather than reporting it as spam." (dislike)
- **Likes/dislikes level of direction**: In the positive, participant indicates that they like that the warning provides specific direction; in the negative, this means that they feel that the warning should include more direction to the user.
  E.g., P4 @ Fake Bank of America Alert w/ contextual warning (short, uncertain framing): "I like that it gives a warning, and that it says to not click links and to contact your bank directly. That is very good advice." (likes)
- **Likes/dislikes level of evidence**: In the positive, participant indicates that they like that the warning provides evidence or an explanation of why something appears to be spam; in the negative, this means that they feel that the warning should include more evidence as to why the message is suspicious.

E.g., P2 @ Employment Scam w/ Spam warning: "I think the similar messages you receive to identify as spam. That doesn't feel like the strongest reason to me. Because it's it. It's not actually telling me why it's spam." (dislike)

- **Warning should / shouldn't include AI disclaimer**: Participator indicates that the warning should or should not include a disclaimer about hte use of AI to generate it
  E.g., P8 @ Fake T-Mobile Giveaway w/ contextual warning (short, uncertain framing): "I don't think it's useful to be told 'this was generated by AI. It may not be a hundred percent accurate' because I mean it's already saying 'this may be a scam'... it's adding more layers of maybes."
- **Likes/dislikes level of detail**: Participant indicates that the warning has the appropriate or inappropriate level of detail without specifying the type of detail they would like. Dislike includes where they say there should be more or less text.
  E.g., P20 @ Fake Amazon Account Alert w/ Generic Scam Warning:"Even though I said for the other ones that more information was useful. This is actually... a very effective message... what they're trying to convey to you is this is likely a scam. So pay attention..." (likes)
- **Text is unnecessary**: Participant indicates that they find the explanation text (i.e., text other than the headline or AI disclaimer) to be unnecessary or even harmful. Includes suggesting the dtext should be moved under the learn more button.
  E.g., P9 @ Spam Warning: "This could be something that's in the learn more button... And... like Google, keeps your information secure and private. That's like unnecessary for me to know."
- **Other specific wording change**: Participant recommends a textual change to the warning that is specific to the text they viewed and not easily generalizable.
  E.g., P1 @ Fake T-mobile Giveaway w/ Contextual warning (long, confident framing): "I mean, technically, the punctuation could be corrected. This is likely a scam period."
- **Reference to other warning**: Partiicpant references the text or content of another warning without being prompted.
  E.g., P9 @ Fake Delivery Alert w/ contextual warning (short, uncertain framing): "and of course and then there's a warning. Now the this one says that it this may be a scam, which is slightly better than this is likely a scam."
- **Learn more should...**: Participant indicates what they expect to be under the learn more button.
  E.g., P9 @ Spam Warning: "This could be something that's in the learn more button... And... like Google, keeps your information secure and private. That's like unnecessary for me to know."
- **General tust**: Participant indicates that they trust phone generated warnings in general.
  E.g., P1 @ Fake T-Mobile Giveaway w/ Contextual warning (long, confident framing): "I would probably trust it. I feel like, you know, phones have that built into their software now that they are trying to keep people safer."

- **General distrust**: Participant indicates that they distrust phone generated warnings to some degree
  E.g., P4 @ Fake Delivery Alert w/ Spam warning: "Recent shipment invalid shipping information—access to update your shipping address. Okay, so this is one I would take an extra minute, because at first glance it says, 'We think this is spam,' and it's that more generic version. It just looks similar to other messages that are spam, and I've even received a couple like this that I knew were not spam."
- **Warning might be wrong**: Participant thinks the particular warning may be wrong about the classification of the message.
  E.g., P16 @ Fedex Notification w/ false positive warning: Interviewer: "Okay. So you you think the warning here might be wrong?" P16: "Yeah."
- **Privacy concern**: Participant mentions that they are concerned about privacy related to the warning, especially the generative AI contextual warnings.
  E.g., P11 @ contexual warning (short, uncertain framing): "I guess I'm kind of confused why t would be generated by AI. Because and I feel like that shouldn't be on there. The artificial, intelligent records. Yeah, that's yeah. That's the only thing I'm kind of confused about. Interviewer: "So when you say it shouldn't be on the AI record. Do you mean like they shouldn't be sending your text messages to AI." "P11: Yeah, that's kind of awfully like confidential, I guess."
- **Helpful for others**: Participant indicates that the warning would be helpful for others, especially those with less knowledge.
  E.g., P1 @ Fake Amazon Account Alert w/ contextual warning (short, uncertain framing): "This one does tell you not to click the link and to verify via the official app or website, which I think is a good reminder for people who might not be familiar with... text as often, like an older adult... I'm thinking about my dad."
- **Other feedback**: Participant describes other specific change not discussed above.
  E.g., P5 @ Fake USPS Delivery Alert w/ Spam warning: "like the not spam button. I would make that like more clear, so make it look like an actual button, so people would know to actually press it."

Confidence (Q8)

- **#**: Open coding with number that expresses confidence from one to ten
- **Unclear**: Participant doesn't express a number, states multiple numbers, or expresses conditional confidence.

Seen similar messages (Q9)

- **Yes**: Participant indicates they have received similar messages
  E.g., P1 @ Fedex Notification: Interviewer: "And finally, I think you already mentioned, you've received messages like this" before. P1: "Yes."

- **No**: Participant indicates they have not seen similar messages. Also includes where participant says yes, but the type of message they describe doesn't actually fit the description. *E.g., P10 @ Fake Delivery Alert: Interviewer: "And have you received messages like this before?" P10: "No."*
- **Unclear**: Participant isn't sure or states that the message seems like something they have probably received or gives another unclear answer.
  E.g., P7 @ Fake Delivery Alert: Interviewer: "And have you seen messages like this before? I think you said you have." P7: "Like this particular—yeah. Well, again, yes, I have seen them, not necessarily to me, but my partner has received warnings from USPS, from UPS, from FedEx. So yeah."
- **Different channel**: Participant has received similar messages through a different messaging channel, like email
  E.g., P5 @ Employment Scam: Interviewer: "And do you receive messages like this in real life?" P5: "I always get these kinds of messages on my [university] email. I think I do receive some text messages like this, but I usually get those through my email."

F.5.5 Post-task Questions

Overall Warning Feedback (Q10 and follow-ups)

- **Positive**: Participant expresses an overall positive view of scam warnings
  E.g., P1: "I think [warnings are] pretty helpful. I think I kind of said this earlier. Everyone's moving fast, and we look at our phones so often that I think it's easy to miss things."
- **Compliment of specific warning**: Participant expresses a positive view of particular warning or set of warnings shown in the interview
  E.g., P2: "I think especially the This may be a scam one for official government agencies. I thought that they were very helpful and kind of giving another opinion as to what I suspect."
- **Criticism of specific warning**: Participant expresses a negative view of a particular warning or set of warnings shown in the interview
  E.g.,"...I thought like the non helpful one was like the one that was still like in the like purple shade that is also like the phone number was like in that same shade like it just didn't really stand out."
- Codes from the Warning section of the task codebook were also applied here.

Other Useful Features (Q11)

- **None**: Participant can't think of other useful features or reiterates the utility of warnings
  E.g., Interviewer: "Are there any features other than warnings that you think your phone should have to help you deal with instant messages from unknown senders?" P1: "I don't know what those features would be."

- **Filtering**: Participant indicates that better filtering or blocking of scam content would be helpful (i.e., from carrier or phone apps). Basically, expression of not wanting to review the messages at all.
  E.g., P9: "I mean it should it should the number, if it's once. Like, let's say, if there's a number like, Oh, [phone number], right. And if it's a similar number, it should just automatically, maybe market as spam or something and not give me a notification. Maybe there's settings for that. I don't know."

F.5.6 Other Codes

The codes is this section were applied across coding sections, wherever they were relevant

- **Reference to older adults**: Participant mentions older adults as a gorup that the warning may be helpful for or who may be vulnerable to scams
  E.g., P1: "I mean, they're similar. This one does tell you not to click the link and to verify via official app or website, which I think is like a good reminder for people that might not be familiar with, or maybe don't use text as often like in an older adult like, I'm thinking about my dad."
- **Reference to information source**: Participant references training, news reporting or other forms of educational material
  E.g., P4: "I have to take the training at work so that I don't click on emails at work. This has all the red flags..."
- **Reference to inbox view**: Participant references viewing the message on the inbox in some way.
  E.g., P4: "I mean coming from an email address is a big red flag. To begin with. So even before I clicked into the message. I mean, it has the little spam stop sign as as the icon."
- **Reference to country differences**: Participant mentions differences in messages received based on their experiences in different countries
  E.g., P9: "Not here, not in the Us. I've never gotten [marketing messages] here."

Appendix G

A Real-World Evaluation of GAI-generated, Context-Specific Warnings

G.1 Main Study Materials

G.1.1 Recruitment Materials

Prolific ad

Title: Investigating the Impact of a Novel SMS App on User Behavior
Overview: We are a group of researchers at Carnegie Mellon University working on a project that investigates how a novel SMS application affects user behavior. This project will involve you installing an app and using it as your primary SMS application for two weeks.
Additional Information: We are a group of researchers at Carnegie Mellon University working on a project that investigates how a novel SMS application affects user behavior. If you:

- Are 18 years of age or older
- Are located in the United States
- Own and use an Android smartphone as your primary mobile device
- Use SMS/text messaging at least once daily

You may be eligible to participate.

If you participate in the full study, you will install an SMS Application and use it as your primary SMS application for two weeks. During this time, you will complete three surveys, which will take 1.5-2 hours in total. We will collect messages you receive from unknown senders (i.e., phone numbers not in your "Contacts") and analyze them with AI. As this is a research app, it is not on the Google Play Store. The installation process will require you to dismiss several security warnings.

This application does not support Rich Communication Services (RCS), so you will not be able to use features such as read receipts, typing indicators, or enhanced group chats during the study. This means the study is best suited for people who do not rely on RCS group chats.

You can receive up to $75.50 for completing all study requirements: $0.50 for screening, $15 for completing the study screening/onboarding survey and installing the SMS app, $25 for

completing a midpoint survey (after 7 days), and $35 for completing a final survey (after 14 days). You are required to keep the SMS app as your primary messaging app throughout the study to receive compensation.

We would greatly appreciate your help with this research. Please feel free to email the Principal Investigator, Elijah Bouma-Sims, at eboumasi@andrew.cmu.edu with any questions.

Recruitment Flyer



Figure G.1: Non-Prolific recruitment advertisement

G.1.2 Sample Participant Communications

The following messages are examples of messages sent to participants throughout the real-world experiment. There were variations in the message content based on how the participants were recruited and based on circumstances (e.g., contacting participants who had technical difficulties). Italics are used to indicate customized content, such as links.

Invitation to participate in the study

Hi from CMU! Thank you for completing the screening survey for the study "Investigating the Impact of a Novel SMS App on User Behavior" (STUDY2026_00000185). Based on your responses, you are eligible to participate in the full 14-day research study. Please visit the following link to complete the onboarding survey *[Link]*

In this survey, you will sign a consent form for the full study, answer demographic questions, and receive instructions on setting up our app. Please be prepared to spend 30 minutes to an hour on this process. Once you complete the onboarding survey and the app has been verified as installed for a day, you will receive $15.

If you no longer wish to participate in the study, you may ignore this message. If you do not complete the onboarding process within 7 days, we will assume you do not want to participate.

Invitation to the midpoint survey

Hi from CMU! Thank you for your first week of participation in the study "Investigating the Impact of a Novel SMS App on User Behavior" (STUDY2026_00000185). You have reached the halfway point! Please visit the following link to complete the midpoint survey: *[Link]*

In this survey, you will provide feedback on your experience with the QUIK app so far. This should take approximately 10 minutes to complete. Once your response is reviewed, you will be sent a $25 gift card.

If you do not wish to participate in the second week of the study, you will have the option to withdraw within the survey link above. Please complete this survey within 72 hours, or we will assume you wish to withdraw.

Reminder to complete the midpoint survey

This is a friendly reminder to complete the midpoint survey: *[link]* This survey should take approximately 10 minutes to complete. Once you complete the midpoint survey and your response is reviewed, you will be sent your week one compensation of $25.00 in a gift card via SMS. If you do not wish to participate in the second week of the study, you will have the option to withdraw within the survey link above. Please complete this survey within *N* hours, or we will assume you wish to withdraw.

Invitation to the conclusion survey

Hi from CMU! Thank you for completing the full 14-day study "Investigating the Impact of a Novel SMS App on User Behavior" (STUDY2026_00000185). Please visit the following link to complete the final survey: *[Link]*

In this survey, you will provide your final feedback on the QUIK app and review the messages we collected during the study. You will also receive instructions on restoring your phone's original settings. Please allow 30 minutes to complete this survey.

Once you complete this final survey and your response is reviewed, you will receive your final payment of $35 in a gift card sent via SMS. Please complete this survey within 72 hours, or we will assume you wish to withdraw.

Reminder to complete the conclusion survey

This is a friendly reminder to complete the conclusion survey: *[link]* This survey should take approximately 30 minutes to complete. Once you complete the conclusion survey and your

response is reviewed, you will be sent your week two compensation of $35.00 in a gift card via SMS. Please complete this survey within *N* hours, or we will assume you wish to withdraw.

Withdrawal/debrief message

You have withdrawn from the study "Investigating the Impact of a Novel SMS App on User Behavior" (STUDY2026_00000185) due to *[reason, e.g., uninstalling the SMS app, failing to complete a survey, etc.]*

We make the following disclosure for transparency. We did not fully disclose the study's purpose. While we stated that our focus is generally on a novel SMS design, our research focuses more specifically on how warnings affect users' interactions with scam SMS messages. Our hypothesis is that AI-generated warnings will lead users to make safer decisions than generic warnings without AI-generated explanations. To test this hypothesis, we may have sent you several scam text messages. The text of these messages can be found in this document *[link to document with SMS messages]*.

This incomplete disclosure was necessary to test our hypothesis, as you may have responded unnaturally had you been aware of our full purpose. To help ensure that other participants also have unbiased responses, we ask that you do not share information about this research project with others.

Although you have already completed your participation in the study, your involvement in the research is still voluntary. Now that you are aware of the true nature of our study, you may choose to have your data removed without penalty. Your choice to remove your data will not negatively affect your relationship with the researchers or Carnegie Mellon University. Because we work quickly to de-identify the study data, you must make your removal request at this time. Please consider carefully. After this point, we will no longer be able to identify your data to remove it.

Questions? Please contact Elijah Bouma-Sims at eboumasi@andrew.cmu.edu. If you have questions or concerns about your rights as a research participant, you can contact the Carnegie Mellon IRB at (412) 268-4721, IRB-review@andrew.cmu.edu

To restore your phone to its original settings, please remember to:

1. Set your preferred messaging app back to 'Default' in your phone settings. For Google Messages, instructions on how to do this can be found here: `https://support.google.com/messages/answer/6089066?hl=en`
2. Uninstall the QUIK app.
3. Re-enable 'RCS' in your messaging settings if you previously disabled it. For Google Messages, instructions can be found here: `https://support.google.com/messages/answer/7189714?hl=en`

G.1.3 Screening Survey

Note: Section headings and text in italics were not shown to participants, but are used to indicate page breaks, survey flow, and text entry boxes.

Basic Screening Questions

1. What operating system does your primary smartphone currently use?
   (a) iOS (iPhone)
   (b) Android (e.g., Samsung, Pixel, Motorola, OnePlus, etc.)
   (c) HarmonyOS (Huawei)
   (d) I do not use a smartphone (e.g., Flip phone / Basic phone)
   (e) Other / I'm not sure
2. In the past 7 days, how often have you used the pre-installed "Messages" app (SMS/RCS) to send or read a message?
   (a) Hourly (or more)
   (b) A few times a day
   (c) Once a day
   (d) A few times a week
   (e) I haven't used it in the last week

Participants indicating that they did not use Android or who used the default messaging app less than once a day were redirected to the ineligible section.

Business Familiarity

3. Please select the option that best describes your familiarity with each of the following businesses. *For each business below, participants chose one of: (1) I am not familiar with this business, app, or service; (2) I'm aware of it, but I have not used the app/service/business in the last year; (3) I have used the app/service/business in the last year, but I* ***do not*** *receive text alerts for it; (4) I have used the app/service/business in the last year, and I* ***receive text alerts*** *for it.*
   (a) Amazon
   (b) Bank of America
   (c) EBay
   (d) T-Mobile
   (e) Nexis Pay
   (f) Discord
   (g) Venmo
   (h) PayPal
   (i) Discover
   (j) Capital One
   (k) Second Pacific National Bank
   (l) Indeed

(m) LinkedIn

(n) Zelle

Participants indicating that they did not use one of Amazon, Indeed, Bank of America, or Venmo were redirected to the ineligible section. Participants who indicated familiarity with Nexis Pay or Second Pacific National Bank were also redirected to the ineligible section.

Assent to Participation Requirements

4. I am willing to install an SMS Application **that is not on the Google Play Store** and use it as my primary SMS application for two weeks.
   (a) Yes
   (b) No
5. I am willing to have my messages from phone numbers not in my contacts collected by researchers and analyzed with AI.
   (a) Yes
   (b) No
6. I am willing to disable Rich Communication Services (RCS) if I use it. This will disable features like read receipts, typing indicators, enhanced group chats, and end-to-end encryption during the study.
   (a) Yes
   (b) No

Participants who answered no to any of the questions in this section were redirected to the ineligible section.

Ineligible

[For Prolific Participants]: Thank you for your interest in our study. Unfortunately, you are not eligible to participate in the full study. After your response is reviewed, you will be compensated $0.50. If you have any questions, contact the Principal Investigator, Elijah Bouma-Sims, at `eboumasi@andrew.cmu.edu`
[For non-prolific participants]: Thank you for your interest in our study. Unfortunately, you are not eligible to participate. If you have any questions, contact the Principal Investigator, Elijah Bouma-Sims, at `eboumasi@andrew.cmu.edu`

7. *[For non-prolific participants]*: If you are interested in entering the raffle for a $50 Amazon gift card, please enter your phone number below. If you win the raffle, the Amazon gift card will be sent to the phone number provided via SMS. *[text entry]*
8. If you win the raffle, in what manner would you prefer to receive compensation?
   (a) Amazon
   (b) Walmart
   (c) Target

(d) I don't want to the enter the raffle

Eligible

[For Prolific Participants] We have determined that you are eligible to participate in our study. Click the next button to proceed to the consent form for the full study.
[For Non-Prolific Participants] We have determined that you are eligible to participate in our study.

9. *[For non-prolific participants]*: What is your mobile phone number? [*text entry*]

   This number will be used for all study-related communications. You will receive a message with a link to the full study once your response is reviewed. Additionally, if you win the random drawing for participating in this screening survey, the gift card will be sent to the phone number provided via SMS.

10. If you win the random drawing, in what manner would you prefer to receive compensation?
    (a) An Amazon Giftcard
    (b) A Walmart Giftcard
    (c) A Target Giftcard

G.1.4 Onboarding Survey

Note: Section headings and text in italics were not shown to participants, but are used to indicate page breaks, survey flow, and text entry boxes.

Introduction

[For Non-Prolific participants] Thank you for agreeing to participate in our study. The purpose of this survey is to learn more about your background and guide you in installing the messaging app for this study.

You must install the app on a phone with the phone number you shared in the screening survey. We will not be able to identify you if you install the application on a different phone. Installing the app on a phone with a fake number or an international phone number will result in your withdrawal.

Please allow 30 minutes to complete this process. If needed, you may bookmark this link and come back later, when you have more time. We recommend completing this survey on a separate device than the phone you will use during the study.

1. *[For Prolific participants]*: What is the phone number for the phone you will use throughout the study?

   You must install the app on a phone with this phone number. We will not be able to identify you if you install the application on a different phone. Installing the app on a phone with a fake number or an international phone number will result in your withdrawal.

Phone Background

2. Approximately how old is your primary smartphone (i.e., the one you will use for the study)?
   (a) Less than 1 year old (purchased in the second half of 2025 or later)
   (b) 1–2 years old (purchased in 2024 or the first half of 2025)
   (c) 3 years old or older (purchased in 2023 or earlier)
   (d) I'm not sure
3. Who is the manufacturer of your primary smartphone?
   (a) Samsung (e.g., Galaxy S26, A-series)
   (b) Google (e.g., Pixel 10, Pixel 9a)
   (c) Motorola (e.g., Moto G, Razr)
   (d) OnePlus (e.g., OnePlus 15)
   (e) Xiaomi (e.g., Redmi, Poco)
   (f) Nothing (e.g., Nothing Phone 3 or 4)
   (g) Other *[text entry]*
   (h) I'm not sure
4. Who is your mobile service provider?
   (a) AT&T
   (b) T-Mobile
   (c) U.S. Cellular
   (d) Verizon
   (e) Other *[text entry]*
   (f) I'm not sure
5. Which of the following SMS applications do you use?
   (a) Google Messages
   (b) Something else / I'm not sure
6. Which, if any, of these other instant messaging apps do you use? *Select all that apply.*
   (a) Discord
   (b) Facebook Messenger
   (c) Google Chat
   (d) Signal
   (e) Telegram
   (f) WhatsApp
   (g) Other *[text entry]*

(h) I don't use other messaging apps. *[exclusive]*

7. Which of the following purposes do you use instant messaging for? *Select all that apply.*
   (a) Finance (e.g., bill reminders, bank fraud alerts, etc.)
   (b) Healthcare (e.g., appointment reminders, notifications about test results, etc.)
   (c) Marketing (e.g., receiving coupons, news, or updates from brands I follow)
   (d) Professional (e.g., communicating with co-workers or managing work tasks)
   (e) Security (e.g., login codes, security alerts, etc.)
   (f) Shopping/logistics (e.g., package tracking, delivery photos, etc.)
   (g) Social (texting friends and family)
   (h) Other *[text entry]*

Demographics

We will now ask you a few optional demographic questions before instructing you on installing the study application. Your answers to these questions will not impact your eligibility to participate in the study and will only be used to contextualize your experience.

8. How do you describe your gender identity?
   (a) Man
   (b) Woman
   (c) Non-binary
   (d) Prefer to self-describe *[text entry]*
   (e) Prefer not to disclose
9. What is your age in years?
   (a) 18 to 24
   (b) 25 to 34
   (c) 35 to 44
   (d) 45 to 54
   (e) 55 to 64
   (f) 65 to 74
   (g) 75 or older
   (h) Prefer not to disclose
10. What is your race? You may select multiple options
    (a) American Indian or Alaskan Native
    (b) Asian
    (c) Black or African American
    (d) Native Hawaiian or Other Pacific Islander

(e) White

(f) Prefer to self-describe *[text entry]*

(g) Prefer not to disclose

11. Are you Hispanic or Latino?

(a) Yes

(b) No

(c) Prefer not to disclose

12. What is your approximate household income?

(a) Less than $25,000 per year

(b) $25,000 to $49,999 per year

(c) $50,000 to $99,999 per year

(d) $100,000 to $199,999 per year

(e) $200,000 or more per year

(f) Prefer not to disclose

13. What is the highest level of school you have completed or the highest degree you have received?

(a) Less than high school degree

(b) High school graduate (high school diploma or equivalent including GED)

(c) Some college but no degree

(d) Associate degree in college (2-year)

(e) Bachelor's degree in college (4-year)

(f) Graduate degree (e.g., MS, PhD, JD, etc.)

(g) Prefer not to disclose

14. Do you have a formal education in a computer-related field, such as computer science or IT? ("Formal education" could mean a completed degree or certificate, or classes or trainings you took towards a degree or certificate.)

(a) Yes

(b) No

(c) I'm not sure *[text entry]*

(d) Prefer not to disclose

15. Do you have work experience in a computer-related field, such as computer science or IT?

(a) Yes

(b) No

(c) I'm not sure *[text entry]*

(d) Prefer not to disclose

Installation Instructions

This section included installation instructions for the QUIK app. Participants were randomly assigned an APK for their cohort. The text included screenshots and a video.

16. *[Shown to Prolific participants]* Were you able to install the app?
    (a) Yes
    (b) No

[Shown to Prolific participants] You have now completed the installation and set-up process for QUIK. Once we have reviewed your response and verified that the app has been installed for a day, we will send you compensation via Prolific. Be aware, you will only receive the full $15 if we can verify that you installed the app and have used it for a day. Otherwise, you will receive the amount for completing screening.

[Shown to non-Prolific Participants] You have now completed the installation and set-up process for QUIK. Once we have reviewed your response and verified that the app has been installed for a day, we will send you compensation via SMS. Please allow some time for the payment to be processed. We will contact you again after 7 days with a survey to assess the first half of your experience. If you have any questions or wish to withdraw from participation, please reach out to Elijah Bouma-Sims via email (eboumasi@andrew.cmu.edu) or text (*[Redacted # with PA Area Code]*)

17. *[Shown to non-Prolific participants]* In what manner would you prefer to receive compensation throughout the study?
    (a) An Amazon Giftcard
    (b) A Walmart Giftcard
    (c) A Target Giftcard

G.1.5 Midpoint Survey

Note: Section headings and text in italics were not shown to participants, but are used to indicate page breaks, survey flow, and text entry boxes.

Main Feedback

Thank you for participating in the first week of the study "Investigating the Impact of a Novel SMS App on User Behavior" (STUDY2026_00000185). This midpoint check-in is designed to gather feedback on the app. We hope to hear your honest opinion, whether positive or negative. You will be paid $25 once your response is reviewed. You will also have the opportunity to withdraw from the study if you do not want to participate in the second week of the study.

1. For the following questions, the term "the system" refers to the QUIK app you have been using during the study. Please provide your first impression for each statement without spending too much time on any one item. *Each statement was rated on a five-point scale from (1) Strongly Disagree to (5) Strongly Agree.*
    (a) I think that I would like to use this system frequently.

(b) I found the system unnecessarily complex.

(c) I thought the system was easy to use.

(d) I think that I would need the support of a technical person to be able to use this system.

(e) I found the various functions in this system were well integrated.

(f) I thought there was too much inconsistency in this system.

(g) I would imagine that most people would learn to use this system very quickly.

(h) I found the system very cumbersome to use.

(i) I felt very confident using the system.

(j) I needed to learn a lot of things before I could get going with this system.

2. Did you experience any technical issues, bugs, or unexpected behavior while using the system during the last week? This could include crashes, frozen screens, delayed notifications, or features that do not seem to work as intended.
   (a) Yes
   (b) No
   (c) I'm not sure
3. *[If answer yes to the previous question]* Please describe the issue you experienced in as much detail as possible. *[text entry]*
4. Were there any specific features or aspects of the app that you particularly liked or found helpful?
   (a) Yes
   (b) No
   (c) I'm not sure
5. *[If answer yes to the previous question]* Please tell us what you liked and why. *[text entry]*
6. If you have any other feedback about the QUIK app, please share it here. *[text entry]*

Withdrawal Check

7. Thank you for your feedback on your experience so far. To conclude the survey, please indicate whether you wish to continue participating in the research study for the second week. You will still receive your $25 midpoint payment, regardless of your response. You may receive an additional $35 for participating in the second week of the study.
   (a) Yes, I wish to continue with the second week of the study.
   (b) No, I wish to withdraw from the study now.

Once you click the next button, your choice will be confirmed. If you choose to withdraw, you will not be able to change your decision past this point.

Withdrawal Survey

This subsection was only shown if participants opted to withdraw.

8. Please briefly describe why you are choosing not to continue with the study. *[text entry]*

You have withdrawn from the study. We make the following disclosure for transparency. We did not fully disclose the study's purpose. While we stated that our focus is generally on a novel SMS design, our research focuses more specifically on how warnings affect users' interactions with scam SMS messages. Our hypothesis is that AI-generated warnings will lead users to make safer decisions than generic warnings without AI-generated explanations. To test this hypothesis, we may have sent you several scam text messages. The text of these messages can be found in this document [link].

This incomplete disclosure was necessary to test our hypothesis, as you may have responded unnaturally had you been aware of our full purpose. To help ensure that other participants also have unbiased responses, we ask that you do not share information about this research project with others.

Questions? Please contact Elijah Bouma-Sims at `eboumasi@andrew.cmu.edu`. If you have questions or concerns about your rights as a research participant, you can contact the Carnegie Mellon IRB at (412) 268-4721, `IRB-review@andrew.cmu.edu`

9. By default, we delete your identifiers (e.g., phone number) and text messages upon withdrawal. Would you like us to delete[1] all other research data we have collected? This includes your prior survey responses and data regarding how you navigated the app.
   (a) Yes, you may keep my anonymized data for research.
   (b) No, please delete all data collected during the research.

To restore your phone to its original settings, please remember to:

1. Set your preferred messaging app back to 'Default' in your phone settings. For Google Messages, instructions on how to do this can be found here: `https://support.google.com/messages/answer/6089066?hl=en`
2. Uninstall the QUIK app.
3. Re-enable 'RCS' in your messaging settings if you previously disabled it. For Google Messages, instructions can be found here: `https://support.google.com/messages/answer/7189714?hl=en`

G.1.6 Conclusion Survey

Note: Section headings and text in italics were not shown to participants, but are used to indicate page breaks, survey flow, and text entry boxes.

[1]This wording was present in the real study and is retained for transparency. We acknowledge this wording may have been confusing, but it was only viewed by a single participant.

General Feedback

Thank you for participating in the second week of the study "Investigating the Impact of a Novel SMS App on User Behavior" (STUDY2026_00000185). This final survey is designed to gather your final feedback on the app and instruct you on uninstalling the study app. You will be paid $35 once your response is reviewed. This survey will take approximately 30 minutes to complete.

1. For the following questions, the term "the system" refers to the QUIK app you have been using during the study. Please provide your first impression for each statement without spending too much time on any one item. *Each statement was rated on a five-point scale from (1) Strongly Disagree to (5) Strongly Agree.*
   (a) I think that I would like to use this system frequently.
   (b) I found the system unnecessarily complex.
   (c) I thought the system was easy to use.
   (d) I think that I would need the support of a technical person to be able to use this system.
   (e) I found the various functions in this system were well integrated.
   (f) I thought there was too much inconsistency in this system.
   (g) I would imagine that most people would learn to use this system very quickly.
   (h) I found the system very cumbersome to use.
   (i) I felt very confident using the system.
   (j) I needed to learn a lot of things before I could get going with this system.
2. Did you experience any technical issues, bugs, or unexpected behavior while using the system during the last week? This could include crashes, frozen screens, delayed notifications, or features that do not seem to work as intended.
   (a) Yes
   (b) No
   (c) I'm not sure
3. *[If answer yes to the previous question]* Please describe the issue you experienced in as much detail as possible. *[text entry]*
4. Were there any specific features or aspects of the app that you particularly liked or found helpful?
   (a) Yes
   (b) No
   (c) I'm not sure
5. *[If answer yes to the previous question]* Please tell us what you liked and why. *[text entry]*
6. If you have any other feedback about the QUIK app, please share it here. *[text entry]*

Message Review

In this section, we want to see how you evaluate different types of messages. Below are several screenshots of messages as they might appear in an SMS app. For each one, please tell us how you would categorize it. You will review eight messages total.

A legitimate message is one that is genuinely sent by the person or organization it claims to be from and has a real, non-deceptive purpose (for example, an authentic delivery update or account notification).

A fraudulent message is one that is deceptive or misleading, such as a scam, phishing attempt, or impersonation, and is intended to trick the recipient into taking an action (for example, clicking a link, sharing personal information, or sending money).

As described in the body of the paper, this section was repeated eight times with the stimuli presented in Appendix G.1.9.

7. Imagine you *[text based on message, e.g., regularly shop at Kohl's]* and receive the text message below. Select the option corresponding to your evaluation of the message.
   The same mockup text message interface from the pre-studies was displayed here. See Appendix G.5.
   (a) Definitely legitimate
   (b) Probably legitimate
   (c) I'm not sure whether it is legitimate or fraudulent.
   (d) Probably fraudulent
   (e) Definitely fraudulent

Warning Recall

Please view the screenshot from the QUIK app below. *Participants were shown a screenshot of the warning design corresponding to their assigned condition.*

8. Do you recall seeing red warning boxes like this while using QUIK?
   (a) Yes
   (b) No
   (c) I'm not sure

Sentiment Likert Scale

This subsection and the following subsection were only shown if participants answered "Yes" or "I'm not sure" to the previous question.

Please view the screenshot from the QUIK app below and reflect on your experience during the study. *Participants were shown a screenshot of the warning design from their assigned condition, with the Bank of America message*

9. How useful do you think these warnings were for you?
   (a) 1: Not at all useful

(b) 2: Slightly useful
(c) 3: Moderately useful
(d) 4: Very useful
(e) 5: Extremely useful

10. How effective were the warnings at grabbing your attention?
    (a) 1: Not at all effective
    (b) 2: Slightly effective
    (c) 3: Moderately effective
    (d) 4: Very effective
    (e) 5: Extremely effective
11. How easy or difficult was it to understand the text of the warnings?
    (a) 1: Very difficult
    (b) 2: Somewhat difficult
    (c) 3: Neither easy nor difficult
    (d) 4: Somewhat easy
    (e) 5: Very easy
12. To what extent did you find the warnings annoying?
    (a) 1: Not at all annoying
    (b) 2: Slightly annoying
    (c) 3: Moderately annoying
    (d) 4: Very annoying
    (e) 5: Extremely annoying
13. How would you rate the length of the warning text?
    (a) 1: Much too short
    (b) 2: Somewhat too short
    (c) 3: Just the right length
    (d) 4: Somewhat too long
    (e) 5: Much too long
14. How much do you trust AI to generate accurate warnings?
    (a) 1: Do not trust at all
    (b) 2: Trust slightly
    (c) 3: Trust moderately
    (d) 4: Trust a lot
    (e) 5: Trust completely

15. How useful do you think these warnings would be for the average person?
    (a) 1: Not at all useful
    (b) 2: Slightly useful
    (c) 3: Moderately useful
    (d) 4: Very useful
    (e) 5: Extremely useful
16. Please explain your ratings by sharing specific feedback about your experience with the warnings during the study. *[text entry]*

Warning Changes

Please view the screenshot from the QUIK app below. *Participants were shown a screenshot of the warning design from their assigned condition, with the Bank of America message*

17. Are there any changes you would make to the warning design?
    (a) Yes
    (b) No
    (c) I'm not sure
18. *[If answer yes to the previous question]* What changes would you make to the warning design? *[text entry]*

Consider the alternative warning design below. *Participants were shown a screenshot of the warning design from the condition they were not assigned to, with the Bank of America message*

19. Compared to the design you saw during the study, how would you rate this new design?
    (a) Much worse
    (b) Somewhat worse
    (c) Neither better nor worse
    (d) Somewhat better
    (e) Much better
20. Please briefly explain your answer. *[text entry]*

Debrief

Thank you for participating in our study.

To obtain your unbiased, natural reaction, we did not fully disclose the study's purpose. While we stated that our focus is generally on a novel SMS design, our research focuses more specifically on how warnings affect users' interactions with scam SMS messages. Our hypothesis is that AI-generated warnings will lead users to make safer decisions than generic warnings without AI-generated explanations. To test this hypothesis, we sent you several simulated scam text messages. The text of these messages can be found in this document: [link]

None of these messages were truly dangerous, as the links/phone numbers within the messages are owned by the research team. Your personal information was not exposed if you interacted with the messages.

This incomplete disclosure was necessary to test our hypothesis, as you may have responded unnaturally had you been aware of our full purpose. To help ensure that other participants also have unbiased responses, we ask that you do not share information about this research project with others.

Although you have already completed your participation in the study, your involvement in the research is still voluntary. Now that you are aware of the true nature of our study, you may choose to have your data removed without penalty or loss of compensation. Because we work quickly to de-identify the study data, you must submit your removal request now. Please consider carefully. After this point, we will no longer be able to identify your data to remove it.

Questions? Please contact Elijah Bouma-Sims at `eboumasi@andrew.cmu.edu`. If you have questions or concerns about your rights as a research participant, you can contact the Carnegie Mellon IRB at (412) 268-4721, `IRB-review@andrew.cmu.edu`

21. Now that you have read the disclosure above, which of the following statements best describes your perception of the simulated scam messages prior to starting this survey.
    (a) I noticed an increase in scam messages, and I thought some or all of the scam messages were sent as part of the study.
    (b) I noticed an increase in scam messages, but I thought it was due to poor spam filtering in the QUIK app.
    (c) I noticed an increase in scam messages, but I didn't think it had anything to do with the study.
    (d) I didn't notice an increase in scam messages.
    (e) Other *[text entry]*
    (f) I'm not sure
22. Please briefly explain your answer. *[text entry]*
23. Having received full disclosure of the study's purpose, would you like to remain in the study or withdraw and have your research data deleted? This will not affect your compensation.
    (a) I do not wish to withdraw.
    (b) I would like to withdraw my consent and have my research data deleted.

Message Redaction

This subsection was only shown to participants who did not withdraw.

We will now ask you to review the messages we collected during the study. You may choose to permanently delete any or all of the messages.

Please note that we will review messages and remove any personal information prior to moving the message to permanent storage. This includes names, addresses, phone numbers, personal website links, security codes, etc.

For example, the message "[#][TikTok] Hi Elijah. 123456 is your verification code" would be stored as "[#][TikTok] Hi [NAME]. XXXXXX is your verification code"

24. Review your messages below and enter your choice in the box: to delete specific messages, paste the message ID; to delete all messages from a sender, paste the sender ID; to delete everything, type "delete all messages". Because we work quickly to de-identify the study data, you must submit your removal request now. Please consider carefully. After this point, we will no longer be able to identify your data to remove it. *[text entry]*

    Participants were shown a table of the messages collected from their device during the study.

Conclusion

You are now done with the survey. To restore your phone to its original settings, please remember to:

1. Set your preferred messaging app back to 'Default' in your phone settings. For Google Messages, instructions on how to do this can be found here: `https://support.google.com/messages/answer/6089066?hl=en`
2. Uninstall the QUIK app.
3. Re-enable 'RCS' in your messaging settings if you previously disabled it. For Google Messages, instructions can be found here: `https://support.google.com/messages/answer/7189714?hl=en`

G.1.7 GPT 5.4 Prompts

Prompts are reproduced as deployed, including minor typographical errors. Refer to the original SmishX paper for more details on how these prompts are used.

Classification Prompt

```
I want you to act as a spam detector to determine whether a given SMS is phishing, spam,
or legitimate. Your analysis should be thorough and evidence-based. Analyze the SMS by
following these steps:

1. If the SMS is promoting any of the following categories: Online gambling, bets, spins,
 adult content, digital currency, lottery, it is either spam or phishing.

2. The SMS is legitimate if it is from known organizations, such as appointment reminders
,  OTP (One-Time Password) verification, delivery notifications, account updates,
tracking information, or other expected messages.

. The SMS is considered legitimate if it involves a conversation between friends, family
members, or colleagues.

4. Promotions and advertisements are considered spam. The SMS is spam if it is promotion
from legitimate companies and is not impersonating any brand, but it is advertisements,
app download promotions, sales promotions, donation requests, event promotions, online
loan services, or other irrelevant information.
```

5. The SMS is phishing if it is fraudulent and attempts to deceive recipients into providing sensitive information or clicking malicious links. Phishing SMS may exhibit the following characteristics:

Promotions or Rewards: Some phishing SMS offer fake prizes, rewards, or other incentives to lure recipients into clicking links or providing personal information.

Urgent or Alarming Language: Phishing messages often create a sense of urgency or fear, such as threats of account suspension, missed payments, or urgent security alerts.

Suspicious Links: Phishing messages may contain links to fake websites designed to steal personal information.

Suspicious Phone Numbers: Phishing messages may contain unofficial phone numbers to deceive users into providing personal information over the phone or through text messages .

Requests for Personal Information: Phishing SMS may ask for sensitive information like passwords, credit card numbers, social security numbers, or other personal details.

Grammatical and Spelling Errors: Many phishing messages contain grammatical mistakes or unusual wording, which can be a red flag for recipients.

\Expired Domain: Phishing websites often use domains that expire quickly or are already listed for sale.

Inconsistency: The URL may be irrelevant to the message content.

6. Please be aware that: It is common to see shortened URLs in SMS. You can get the expanded URL from the provided redirection chain. Both phishing and legitimate URLs can be shortened. And both phishing and legitimate websites may use a robot-human verification page (CAPTCHA-like mechanism) before granting access the content.

7. Please be aware that both phishing and legitimate messages may include phone numbers. Legitimate messages will include numbers officially known to be associted with a business. Oficial phoen numbers are often toll-free or short-code phone numbers. Smaller businesses or individuals may use normal, 10-digit phone numbers.

8. I will provide you with some external information if there is a URL in the SMS. The information includes:

- Redirect Chain: The URL may redirect through multiple intermediate links before reaching the final destination; if any of them is flagged as phishing, the original URL becomes suspicious.

- Brand Search Information: The top five results from a Google search of the brand name. You can compare if the URL’s domain matches the results from Google.

- Screenshot Description: A description of the website’s screenshot, highlighting any notable visual elements.

- HTML Content Summary: The title of HTML, and the summary of its content.

- Domain Information: The domain registration details, including registrar, creation date , and DNS records, which are analyzed to verify the domain’s legitimacy.

9. Please give your rationales before making a decision. And your output should be in json format and should not have any other output:

- brand_impersonated: brand name associated with the SMS, if applicable.

- URL: any URL appears in SMS, if no URL, answer "non".

```
- rationales: detailed rationales for the determination, up to 500 words. Directly give
sentences, do not categorize the rationales. Only tell the reasons why the SMS is
legitimate or not, do not include the reasons why the SMS is spam or phishing.

- brief_reason: brief reason for the determination.

- category: True or False. If the SMS is legitimate or spam, output False. If the SMS is
phishing, output True.

- advice: If the SMS is phishing, output potential risk and your advice for the
recipients, such as ''Do not respond to this message or access the link.''

Below is the information of the SMS:
```

Explanation Prompt

```
Based on the detailed analysis, If the SMS message is a scam, I want you to create two
versions of a simple and easy-to-understand explanation to appear within a warning. Use
plain language and avoid technical terms like URL or HTTP headers. The warning should
capture users' attention and present users with the evidence necessary to make an
informed decision. You should consider the visual context when writing the warning. You
should use bolding to help emphasize important words, but do not bold the entire
explanation.

The first explanation should be no more than a single, short sentence that gives clear,
context-specific direction on what the user should do with the message. This explanation
should prioritize brevity, but should not just assert that the message is a scam.

The second explanation should expand on the first. It should explain your conclusion in
no more than three bullet points, focusing on whether the message seems suspicious or
safe. Provide a simple reason to support your conclusion, including clear evidence such
as a suspicious website link or an urgent tone in the message. The explanation should be
reassuring and concise, easy for anyone to understand. One of the bullet points should
give clear guidance on what the user should do if the message is a scam.

The first explanation will appear within a warning in a mobile text messaging interface,
under a heading that reads, "This is likely a scam." The second explanation will appear
if the user presses the "Learn More" button. You should only provide the explanations,
not the headings. Format the output as JSON with the keys "short" and "long." The values
should be markdown code in a JSON string. If the detailed analysis indicates that the
message is not a scam, both output values should be false.
```

G.1.8 Message and Warning Stimuli

All formatting was visible to participants, except for things in italics. Italics are used to indicate redacted information.

Bank of America Message

Message: Bank of America Account Alert: A purchase at STARBUCKS STORE 6192 occurred outside the US. If this wasn't you, call *[Redacted # with MI Area Code]* immediately. Text STOP to quit.
Long Warning Text:

- This message looks **suspicious** because it claims to be from Bank of America but tells you to call a **regular phone number** that does not clearly look like an official Bank of America contact.
- It also uses an **urgent warning** about a purchase outside the US to push you to act quickly before checking whether the message is real.
- **Do not reply or call the number in the text.** If you want to verify the charge, contact Discover[2] using the **number on the back of your card** or the **official app**.

Chili's (OpenTable) Message

Message: You booked Chili's for 2 on July 31st Fri 5:15 PM. Click here to cancel: https://confirm-opentable.com/
Long Warning Text:

- This message is **suspicious** because the link does **not** use OpenTable's normal website name; it uses a lookalike name instead.
- It also pushes you to act on a reservation, but the link behavior doesn't clearly match a real reservation confirmation page for **Chili's**.
- **What to do: Do not click the link** or enter any information; check the booking in the OpenTable app, go to **opentable.com** yourself, or call **The Commoner**[3] using a trusted number.

Amazon Message

Message: Delivery attempted: Unfortunately, we couldn't complete your Amazon delivery. We'll try again. More info at https://www.amązon.com/track?packageId=582-9104736-2584109
Long Warning Text:

- This message looks **suspicious** because it pretends to be from Amazon but uses a lookalike website name: **amązon.com** instead of the normal **amazon.com**.
- The link leads to an Amazon-looking **sign-in page**, which is not normal for simply checking a delivery and may be trying to steal your login details.
- **Do not tap the link or enter any information**; check your order only in the Amazon app or by typing **amazon.com** yourself.

CVS Message

Message: CVS Pharmacy: Your ExtraCare account has a new 40% OFF reward for any prescription or health item. Log in to your dashboard to activate before it expires: https://get-apt.com/cvs
Long Warning Text:

[2]As discussed in the text, this issue was present in the real study, and is repeated here to faithfully reflect the study stimuli

[3]As discussed in the text, this issue was present in the real study, and is repeated here to faithfully reflect the study stimuli

- This message looks **suspicious** because it says it is from CVS, but the link goes to **get-apt.com**, which is **not** a CVS website.
- It tries to push you to act fast with a **limited-time reward** and asks you to log in, which is a common way scam texts try to get your sign-in details.
- **Do not tap the link** or enter any information. If you want to check your ExtraCare rewards, open the **CVS app** or type **cvs.com** into your browser yourself.

DocuSign Message

Message: Sarah Connolly sent you a DocuSign document "Invoice" to view and sign: https://docusign.net.signing
Long Warning Text:

- This message looks **suspicious** because it pretends to be from **DocuSign** but the link goes to a long, unusual **.xyz** address instead of DocuSign's normal site.
- It tries to get you to open and sign an **invoice**, which can pressure people to act quickly, and the page appears to ask for your email sign-in details.
- **Do not tap the link or reply.** If you were expecting a real document, open DocuSign directly from its official app or type the company's website yourself. If you already entered your details, **change your password right away**.

ID.me Message

Message: This message is from ID.me. Your identity is being used to log in to ID.me Gateway. Please click this link to upload a picture or to report unauthorized use of your identity: https://id-me.us/en/phones/c3f9f5a72e-b3424c9e73659a0a40b8df/authorize?vfp=r_zXElAp
Long Warning Text:

- This message looks **suspicious** because it says it is from ID.me, but the link goes to **id-me.us** instead of ID.me's official site.
- It also uses an **urgent warning** about someone using your identity to push you to act quickly, which is a common scam tactic.
- **Do not click the link or reply**; instead, go to the official **ID.me** website or app yourself to check your account, and if you already used the link, change your password right away.

Indeed Message

Message: Indeed: A recruiter from Apple has viewed your profile. Log in to view the update: https://indeed.com@hiring-portal@apply-now-jobs.com/#@indeed.com
Long Warning Text:

- This message is **suspicious** because it claims to be about **Indeed** and **Apple**, but the link actually goes to **apply-now-jobs.com**, not an official Indeed or Apple site.
- The message tries to get you to **log in right away**, which is a common trick used to steal usernames and passwords.

- **Do not tap the link or enter your login details**; instead, open Indeed or Apple directly from their official app or by typing the site name yourself.

Venmo Message

Message: Jack Sims sent $44.62 - Dinner
To accept, go to Venmo and complete the request: https://venmoapp.com/73659a0a40b8df
Long Warning Text:

- This message looks **suspicious** because it pretends to be from Venmo but sends you to **venmoapp.com** instead of Venmo's real site, **venmo.com**.
- It tries to get you to act quickly by saying money was sent and asking you to **accept** it through the link, which is a common scam tactic.
- **Do not tap the link** or enter any information; if you want to check, open the official Venmo app or type **venmo.com** yourself.

G.1.9 Conclusion Survey Message Stimuli

Formatting was shown to participants.

Kohl's Message (Legit)

Sender: 70343
Message: Don't miss this. Stack the savings with $10 + 20% off! PLUS, earn Kohl's Cash! Shop now: http://khls.us/LCI5ba | STOP to end

T-Mobile Message (Legit)

Sender: 456
Message: IMPORTANT: Your T-Mobile bill is past due. Pay now or set up a Payment Arrangement at t-mobile.com/ppy to avoid service interruption & $20 restore fee per line. CA Debt Collector License. #11141-99

Walmart Message (Legit)

Sender: 61746
Message: Walmart: Zaire started shopping and will contact you if there are changes. View order: https://w-mt.co/g/cu5e0L Reply HELP for info; STOP to opt out

Lyft Message (Legit)

Sender: *[Redacted # with PA Area Code]*
Message: We noticed a login to your Lyft account from a new device or browser on 7:33 PM EDT on August 12. If this wasn't you, let us know: https://account.lyft.com/secure?token=*[token]* &type=phone

PayPal Message (Scam)

Sender: 729725
Message: PayPal: You paid $824.66 USD to Samsung Electronics America on May 14, 2026. Log in or get the app to view recent activity: https://paypalapp.io/2C2di3

Discover Message (Scam)

Sender: 347268
Message: Discover Card Alert: A purchase occurred on your account outside of the United States or U.S. Territories. Notify Discover of your travel plans or report this transaction as suspicious by logging into https://discover.com.report-in.site/ad70d767e59d199b Reply STOP to end texts.

Target Message (Scam)

Sender: 56457
Message: Thank you for joining Target Circle! Access your 20% off coupon at retail-rewards.shop /target/circle Reply STOP to unsubscribe

FedEx Message (Scam)

Sender: 37473
Message: Free FedEx alert: Signature req'd for package delivery 11/04. For options fédex.com/ t/537013149667/en_US and select Manage Delivery. Reply STOP to stop msgs.

G.1.10 Codebooks

Effect of Warning on Actions

- **Affirmed right decision**: Participant indicates that they would have come to the same decision without the warning, but the warning made them more confident or reassured in their judgment.
  E.g., "The warning for me served as reassurance for what I already knew"
- **Highlighted particular action**: Participant indicates that the warning led them to select a particular option (e.g., because it made it easy).
  E.g., "I honestly had no idea there was a block feature, apparently google messages has one as well. The last time i had to do that I remember having to call my service to do it manually, a long time ago."
- **I don't need a warning**: Participant doesn't feel like the warnings were important for them. This is not necessarily a negative sentiment.
  E.g., "The warnings helped but I feel comfortable spotting scams either way"
- **Warning as red flag**: Participant indicates that the warning reduced trust in the message or increased suspicion
  E.g., "What it provided was a quick review of what I didnt need to confim." E.g., "As I

recall, the times that I received a warning... I appreciated the chance to pause and give a little more consideration to what was received and how I wanted to respond."

Warning Feedback

- **Accurate**: Participant perceives the warning as accurate or precise *E.g., "The app accurately flagged a text I received as a possible scam. It was a phishing attempt, and the app sowed the warning, which I already knew about, but it was helpful and accurate."*
- **Attention grabbing**: Participant describes the warning as grabbing their attention in some way
  E.g., "I believe I gave positive feedback about the look of the alert system after the first 7 days of using the system. It's extremely attention grabbing and, I believe, gives everyone a better chance of catching a scam before clicking on a link"
- **Compares warning to google messages**: Participant compares the warning to system to google messages *E.g., "I found the AI system helpful for finding fraudulent messages with a quick report and block button, but Google Messages also has that so...I don't know if its AI though"*
- **Helpful for others**: Participant indicates that the warning would be helpful for others, especially those with less knowledge.
  E.g., "Too long but helpful for people who may not be able to spot scams on their own"
- **Inaccurate**: Participant perceives the warning as inaccurate or imprecise
  E.g.,"...I will say that it did miss multiple actual scam texts I had received so I don't know that'd rely on it wholly.."
- **Likes/dislikes color**: Participant likes or dislikes the color of the warning; dislike includes recommendation to change color.
  E.g., "Red on pink might be hard for some users to view/read easily. I am not sure what a better combo would be.. maybe a light yellow background with red symbols/buttons?"
- **Likes/dislikes level of evidence**: In the positive, participant indicates that they like that the warning provides evidence or an explanation of why something appears to be spam; in the negative, this means that they feel that the warning should include more evidence as to why the message is suspicious.
  E.g., "....and I don't like how it doesn't explain how it knew it was a scam, some people like me might like to know that" (generic)
- **Likes/dislikes options**: Participant likes or dislikes the options associated with a warning/notice.
  E.g., "It was useful to me, especially the report and block button."
- **Likes/dislikes tone of warning**: Participant likes or doesn't like the tone or level of confidence of warning; dislike includes recommendations to change (e.g, should be more confident or direct)
  E.g., "I think alot of people overlook messages that might be dangerous, but the AI messages make it feel more serious, which I think is a good thing..."

- **Likes/dislikes level of detail**: Participant indicates that the warning has the appropriate or inappropriate level of detail without specifying the type of detail they would like. Dislike includes where they say there should be more or less text.
  E.g., "While I appreciate the extra information, it's a little much."
- **Likes seeing spam**: Participant mentions that they like seeing the spam messages with a warning, rather than such messages being filtered
  E.g., I liked how big and obvious the alert is when there's a scam text and that it can tell most of my regular texts or subscribed to alerts from this. I believe Google Messages just hides a lot of these, but being more aware of what's happening is a nice touch.
- **Likes/dislikes size**: In the negative: participant indicates that the warning is too large or too small; includes cases when the participant recommends that the warning be made bigger/smaller. In the positive, participant indicates that they felt the warning was the right size or that size was a positive.
  E.g., "I wouldn't make it so darn big! You could have it smaller at the top of the screen and maybe at least allow you to collapse it."
- **Request for accordion**: Participant requests a design similar to the accordion tested in the pre-studies
  E.g., "...Warning label seems way too long - maybe 1 sentence saying this looks suspicious and do not call that number, with the reason details in the Learn More section"
- **Request for in-line warning indicator**: Participant requests some sort of warning in or near the chat bubble.
  E.g., "Some people may not look at the top of the [s]creen, so i would include a warning sign next to the text as well, or some kind of flag."
- **Thinks learn more is accordion**: Participant response indicates that they believe the learn more button in the generic warning acts as an accordion
  E.g., "This is what I love to see. If it's a scam and I'm actually interested in learning more then I have the option to tap the learn more button."
- **Other/vague aesthetic criticism of warning**: Participant expresses negativity about the aesthetics of the warning in a general manner or in a specific way not covered by the other codes
  E.g., "For the design itself, I think the bolding can feel a bit arbitrary (I'm not sure the single word 'suspicious' necessarily needs to be bolded); I think the bolding would work best for instruction lines (like the 'do not reply or call the number' line) and not for scattered words, but that might just be me!"
- **Other/vague negativity/request**: Participant expresses negativity about the warnings without pointing to a specific features OR by pointing to features of the warning that are not covered by other codes
  E.g., "I would probably include a link under the "learn more" button, which could direct the user to a page explaining more tactics commonly used in scam texts. I think the information in the warning itself is great to have but more information is always good!"

- **Other/vague positivity**: Participant expresses positivity about the warnings without pointing to a specific features OR by pointing to features of the warning that are not covered by other codes
  E.g., "received a few and worked without a issue."
- **Other specific wording change**: Participant recommends a textual change to the warning that is specific to the text they viewed and not easily generalizable.
  E.g., "To shorten the text some, I would remove the second bullet point that talks about urgent warnings or possibly just shorten it."

QUIK Good Features

- **Aesthetics**: Participant discusses aesthetic features of the app like the coloring or themes as a positive
  E.g., "I like that I can change the appearance to an extra dark mode and for the accent there are really pretty and bright colors to choose from."
- **Fast**: Participant mentions that the app is fast or efficient in some way
  E.g., "The appworked and loaded fast with messages."
- **Menus**: Participant positively discusses UI features relating to the menu organization, like the details button
  E.g., "I like that it has a menu of options, such as blocked numbers, birthday reminders, and even archived messages, without having to struggle to find them."
- **QUIK Blocking**: Participant positively discusses blocking features without clearly indicating that its in the warning. QUIK includes a block list feature
  E.g., "Easy to block numbers"
- **Schedule texts**: Participant positively discusses the ability to schedule texts
  E.g., "I liked that it provided an easy and intuitive way to schedule messages"
- **Search or sorting**: Participant positively discusses some aspect of message sorting, searching, or filtering in QUIK
  E.g., "The search engine on top to filter out messages is helpful"
- **Voice to text**: Participants mentions the voice to text button, positively.
  E.g., "Again like I said last time, I really liked the text to talk icon. It made it obvious what I needed to do to use it."
- **Other/vague positivity**: Participant mentions other positive feature of their experience that was not mentioned by other participants or was too vague to categorize.
  E.g., "liked the simplicity, wasn't overly complicated. Wish some things were added to it though to make somethings easier to do but this is test so they can be added after and make the app even better."

Bugs or Complaints

- **Aesthetics (negative)**: Participant complains about an aesthetic aspect of the app or requests changes

E.g., "I get that this is an app for the study and please don't take this feedback as genuine complaining, but I HATE how boring the app looks..."

- **Crashing**: Participant mentions the app crashing or freezing, but not in a manner that totally prevents the app from working
  E.g., "the app froze many times for me and had a stutter whenever i opened msgs in the app"
- **Issues w/ sending or receiving messages**: Participant mentions not being able to send or receive some messages, especially in group messages. Includes specific types of messages (like video or GIFs). Includes having to repeatedly sync messages
  E.g., "One of the biggest issues was group chats. I had transferred over all my group chats like was suggested and I was still having issues with one in particular. I wasn't receiving messages from it but my boyfriend was on his line and it said I was a part of it."
- **Issue w/ notifications**: Participant describes issues with receiving notification; does not include message delay issues
  E.g., "Wasn't getting notifications that I had received test messages. I would have to randomly click on the app to see if I had any messages."
- **Lag**: Participant mentions issue with lag, slowdown, or other things working slowly
  E.g., "a bit laggy"
- **Link navigation warning**: Participant mentions annoyance with the warning that appears every time you try to open a link; this can be disabled in settings, but is enabled by default
  E.g., "I found it pretty annoying to have to tap a dialogue every time I tried to open a link in a text. Warn me about things from unknown senders or things that look sus but things from people I interact with frequently shouldn't have a warning like this without further reason like a suspicious url"
- **Message delay**: Participant mentions the delay in receiving messages from AI classification
  E.g., "Messages did not come thru right away"
- **Missing RCS feature(s)**: Participant mentions features that are missing from RCS, like read receipts, emoji reactions, and media quality. Also includes when participants explicitly mention RCS or "rich text"
  E.g., "I would like to be able to react to messages like I am able to do on google messages or on facebook messaging. I really like this feature."
- **Old messages synced**: Participant mentions that, upon syncing messages when starting QUIK, a number of old or spam messages appeared
  E.g., "It pulled ALL of my texts into my main screen, even the ones I had deleted long ago. So now there is just a giant morass of texts and I don't want to go through all of them to delete them. I only keep ones in my main screen that I have to respond to and now that is a mess."
- **Unable to open**: Participant mentions issue with app being unable to open or repeatedly crashing, such that it has to be reinstalled

E.g., "I had a couple of days that the app didn't load at all. I had to uninstall and reinstall, and had no problems after."

- **Voice to text (negative)**: Participant discusses annoyance with the voice to text button
  E.g., "I do wish that the speech to text button was automatically turned off."
- **Other/vague issues/requests**: Participant notes another issue/bug not described above; includes vague or unclear requests/issues
  E.g., "I just don't like the app overall..."

Other Codes

- **Potential recognition of study design**: Participant mentions an element of the study design, such as the warning presence randomization (i.e., warnings appearing or going away) or simulated scam messages. They do not have to explicitly identify it as being caused by the study
  E.g., "Everything seems to be working as intended. Though, i seem to be getting a few more spam messages than usual, may just be a coincidence."
- **Irrelevant**: Entire response is irrelevant to the study.
  E.g., "thanks for involvng me"
- **N/A**: Participant did not answer question or says something that is equivalent to not answering
  E.g., "none"

G.2 Pre-study 1 Material

G.2.1 Survey Instrument

Note: Section headings and text in italics were not shown to participants, but are used to indicate page breaks, survey flow, and text entry boxes.

General Introduction

In this study, we will ask you to use and provide feedback on a new text messaging system that is currently under development. You will first view six text messages using the new interface. We will ask you to specify what you are most likely to do in response to each message. Below is an example of the new text messaging system. You may interact with it by pressing the "three dots" in the top right corner.
The mockup text message interface was displayed here. See Appendix G.5

1. It is important to review each message before proceeding. Which option best describes the subject of the text message above?
   (a) Code for logging into Amazon
   (b) Online shopping promotion

(c) A local political campaign

(d) A package shipment

Participants who didn't select "Code for logging into Amazon" were marked as failing an attention check

Warning Introduction (Non-control conditions)

Some messages will be displayed with a red warning box, as shown below. This box is generated by the new text messaging system. You may interact with the warning by pressing *[text specific to each warning variant e.g., pressing the "Learn more" or "Report & Block" button.]*

The mockup text message interface was displayed with the appropriate warning here. See Appendix G.5.

2. It is important to review each message before proceeding. Which option best describes the subject of the text message above?
   (a) An arrest threat
   (b) An unpaid toll
   (c) A severe virus detected on the recipient's device
   (d) A package shipment

Participants who didn't select "An arrest threat" were marked as failing an attention check

Message Review

As described in the body of the paper, this section was repeated six times with the stimuli presented in Appendix G.2.2

3. Imagine you *[text unique to each message e.g., "are expecting a package via USPS"]* Please view the text message below and then select the option corresponding to what you believe you are most likely to do if you received this message on your phone.
   The mockup text message interface was displayed with the appropriate warning here. See Appendix G.5.
   (a) Reply to the message. Please specify what you would say in reply *[Text entry]*
   (b) Visit the website linked within the message
   (c) Ignore the message without doing anything
   (d) Click the "three dots" to open the menu and select "delete."
   (e) Click the "three dots" to open the menu and select "report & block."
   (f) Search online for information about this message
   (g) *(Displayed when a spam warning was present)* Click the "Not Spam" button
   (h) *(Displayed when a generic, blocking, accordion, or long warning was present)* Click the "Report & Block" button.

(i) Other. Please specify what you would do *[Text entry]*

4. Why do you think this is what you would do? *[Paragraph-entry box]*

Sentiment Likert Scale

We will now ask you several questions about the warning design you saw. For reference, an example of the warning is shown below.
The mockup text message interface was displayed with a previously unseen message and the appropriate warning here. See Appendix G.5 for visual reference.

5. How useful do you think these warnings were for you?
   (a) 1: Not at all useful
   (b) 2: Slightly useful
   (c) 3: Moderately useful
   (d) 4: Very useful
   (e) 5: Extremely useful
6. How effective were the warnings at grabbing your attention?
   (a) 1: Not at all effective
   (b) 2: Slightly effective
   (c) 3: Moderately effective
   (d) 4: Very effective
   (e) 5: Extremely effective
7. How easy or difficult was it to understand the text of the warnings?
   (a) 1: Very difficult
   (b) 2: Somewhat difficult
   (c) 3: Neither easy nor difficult
   (d) 4: Somewhat easy
   (e) 5: Very easy
8. To what extent did you find the warnings annoying?
   (a) 1: Not at all annoying
   (b) 2: Slightly annoying
   (c) 3: Moderately annoying
   (d) 4: Very annoying
   (e) 5: Extremely annoying
9. How would you rate the length of the warning text?
   (a) 1: Much too short
   (b) 2: Somewhat too short

(c) 3: Just the right length
(d) 4: Somewhat too long
(e) 5: Much too long

10. How much do you trust AI to generate accurate warnings?
    (a) 1: Do not trust at all
    (b) 2: Trust slightly
    (c) 3: Trust moderately
    (d) 4: Trust a lot
    (e) 5: Trust completely
11. How useful do you think these warnings would be for the average person?
    (a) 1: Not at all useful
    (b) 2: Slightly useful
    (c) 3: Moderately useful
    (d) 4: Very useful
    (e) 5: Extremely useful
12. If your phone identifies a text as a scam, how would you prefer to handle the message?
    (a) 1: Strongly prefer to see the message anyway
    (b) 2: Somewhat prefer to see the message anyway
    (c) 3: No preference
    (d) 4: Somewhat prefer it be moved to a spam folder without viewing it
    (e) 5: Strongly prefer it be moved to a spam folder without viewing it
13. How comfortable are you with an AI analyzing your instant messages?
    (a) 1: Very uncomfortable
    (b) 2: Somewhat uncomfortable
    (c) 3: Neither comfortable nor uncomfortable
    (d) 4: Somewhat comfortable
    (e) 5: Very comfortable

Warning Changes

The mockup text message interface was displayed with a previously unseen message and the appropriate warning here. See Appendix G.5 for visual reference.

14. Are there any changes you would make to the warning design?
    (a) Yes
    (b) No
    (c) I am not sure

15. *Displayed if yes selected*: What changes would you make to the warning design? *(text entry)*

Alternate Designs

Please consider the following warnings, which include both the warning you saw during the message review task and alternate designs. We will ask you to select the design you believe would be most useful.
Five mockup text interfaces were displayed, each corresponding to one of the warning conditions. See Appendix G.5 for visual reference. The names were chosen to reduce bias in participant responses. Detailed = Long, Expandable = Accordion, Simple = Generic, Pop-up = Blocking, Standard = Spam

16. Which warning format would be the most useful to you in assessing fraudulent instant messages?
    (a) Detailed Alert
    (b) Standard Alert
    (c) Pop-up Alert
    (d) Expandable Alert
    (e) Simple Alert
17. Why do you think this warning would be the most useful? *(text entry)*

Demographics

Thank you for completing the main portion of the survey. We will now ask you a few optional demographic questions and then conclude the survey.

18. How do you describe your gender identity?
    (a) Man
    (b) Woman
    (c) Non-binary
    (d) Prefer to self-describe
    (e) Prefer not to disclose
19. What is your age in years? *Note: For the first participants who completed the survey, one of the age buckets was missing. As a result, we used age data provided by Prolific to understand the age of participants.*
    (a) 18 to 24
    (b) 25 to 34
    (c) 35 to 44
    (d) 45 to 54
    (e) 55 to 64

(f) 65 to 74

(g) 75 or older

(h) Prefer not to disclose

20. What is your race? You may select multiple options

(a) American Indian or Alaskan Native

(b) Asian

(c) Black or African American

(d) Native Hawaiian or Other Pacific Islander

(e) White

(f) Prefer to self-describe

(g) Prefer not to disclose

21. Are you Hispanic or Latino?

(a) Yes

(b) No

(c) Prefer not to disclose

22. What is your approximate household income?

(a) Less than $25,000 per year

(b) $25,000 to $49,999 per year

(c) $50,000 to $99,999 per year

(d) $100,000 to $199,999 per year

(e) $200,000 or more per year

(f) Prefer not to disclose

23. What is the highest level of school you have completed or the highest degree you have received?

(a) Less than high school degree

(b) High school graduate (high school diploma or equivalent including GED)

(c) Some college but no degree

(d) Associate degree in college (2-year)

(e) Bachelor's degree in college (4-year)

(f) Graduate degree (e.g., MS, PhD, JD, etc.)

(g) Prefer not to disclose

24. Do you have a formal education in a computer-related field, such as computer science or IT? ("Formal education" could mean a completed degree or certificate, or classes or trainings you took towards a degree or certificate.)

(a) Yes

(b) No

(c) I'm not sure

(d) Prefer not to disclose

25. Do you have work experience in a computer-related field, such as computer science or IT?

(a) Yes

(b) No

(c) I'm not sure

(d) Prefer not to disclose

G.2.2 Message and Warning Stimuli

Formatting was shown to participants. Messages/warnings labeled as "intro" were displayed during the introduction section.

Intro Message 1

Sender: 98626
Message: Amazon: Your code is 134091. Don't share it. If you didn't request it, deny here https://amazon.com/a/c/r/SuLqMLDcVzjcmyrE0t19sZIfv

Intro Message 2

Sender: *[Redacted # with D.C. Area Code]*
Message: WARNING: (Criminal Investigation Division) I.R.S is filing lawsuit against you, for more information call on *[Redacted # with Virginia Area Code]* on urgent basis, Otherwise your arrest warrant will be forwarded to your local police department and your property and bank accounts and social benifits will be frozen by government.
Long Warning Text:

- The message uses **extreme threats** like arrest and frozen accounts to scare you into reacting quickly.
- The IRS **does not contact people by text** about lawsuits or warrants, and the wording contains obvious mistakes.
- This is a scare tactic. **Ignore and delete the message** — do not call the number.

Initial Accordion Warning Text: **Do not call the number** — real IRS notices never arrive by text.

Scam 1

Sender: *[Redacted # with Southern CA Area Code]*
Message: We have detected an issue with your recent shipment; invalid shipping information. Access http://uspsdeliveryrd.ga/ to update your shipping address.
Long Warning Text:

- The message uses **urgent language** about a shipping problem to push you into acting quickly.
- The link leads to a **fake website** that doesn't match USPS's real domain.
- The sender wants you to enter personal details; this is a common scam. **Delete the message and ignore the link.**

Initial Accordion Warning Text: **Don't click the link** — this message isn't really from USPS.

Scam 2

Sender: *[Redacted # with Western PA Area Code]*
Message: Hello, are you looking for a part-time job now? We provide sales growth services for Amazon, and you can get a stable income of $30-200 in just one hour a day. You do not need to pay any deposit or membership fee, regardless of gender, 23-68 years old, and the Commission is paid every day. If you are interested, please add my telegram. T.ME/JRE7708 or WhatsApp: *[Redacted UK Phone Number]*
Long Warning Text:

- The message promises **unrealistically high pay** for very little work, a common scam tactic.
- It asks you to reply on **Telegram or WhatsApp**, which scammers often use to stay anonymous.
- There's **no real company** named or any details about the job. **Ignore and delete the message**.

Initial Accordion Warning Text: **Don't respond** — this job offer is highly suspicious.

Scam 3

Sender: *[Redacted # with Kansas Area Code]*
Message: BofA Alert: We 've placed a hold on your debit/atm card due to some unusual activity, To recover Visit https://profib04.com and verify your identity
Long Warning Text:

- The message creates **urgency** by claiming your card is on hold, which scammers use to push quick action.
- The link goes to a **non-BofA website** that doesn't match Bank of America's real web addresses.
- The sender is asking you to **verify your identity**, a common phishing trick. **Delete the message and avoid the link**.

Initial Accordion Warning Text: **Don't open the link** — this message isn't really from Bank of America.

Legit 1

Sender: 51362

Message: MyApex: You have an upcoming appt. at Apex Health. Log in to view appt details with MyApex. https://myapex.apexhealth.com/appointments Reply XRemind to opt out

Legit 2

Sender: 46339
Message: FedEx: A package from Amazon was delivered 12/13 at 9:39A. See photo and rate delivery: fedex.com/t/*[Redacted Tracking #]*/en_US Reply HELP for help. STOP to cancel.

Legit 3

Sender: 70343
Message: Overstock Day is here with our best home deals of the year! TODAY ONLY get 20% off + 2x rewards + free shipping!* https://o.bttn.io/0nvkx2i4Cqp

Sentiment Testing Message

Sender: *[Redacted # with California Area Code]*
Message: This is a gift to express our gratitude for your exceptional payment record. T Mobile POSC9.XYZ/YgSpZZjZoh
Long Warning Text:

- The message promises a **reward you didn't ask for**, a common trick scammers use to get clicks.
- The link goes to a **fake website** that doesn't match T-Mobile's real web addresses.
- Real companies don't send gifts this way. **Delete the message and avoid the link**.

Initial Accordion Warning Text: **Don't open the link** — this "gift" message isn't really from T-Mobile.

G.2.3 Additional Results

Table G.1: Nested mixed-effects logistic regression model comparisons for and dismissal outcomes

Outcome	Model	k	AIC	BIC	$\log L$	Deviance	χ^2 (df)	p
Dismiss	Base	4	906.42	926.28	−449.21	898.42		
	+ Warning present	5	896.46	921.28	−443.23	886.46	11.96 (1)	$< .001$
	+ Warning condition	9	899.38	944.07	−440.69	881.38	5.08 (4)	.279

Table G.2: Mixed-effects logistic regression predicting whether a participant dismissed a message in pre-study 1: Dismiss $\sim$ Message_ID + Warning_Present + $(1 \mid \text{SESSION\_ID})$. The first scam message (scam1) and the no-warning condition are the reference levels. Odds ratios are e^{β}. $N = 1059$ observations from 353 participants; random intercept variance $= 1.328$ ($SD = 1.153$).

	β	OR	$S.E.$	z	p
(Intercept)	0.518	1.68	0.145	3.583	$< 0.001^{***}$
Message (ref. = scam1)					
scam2	3.250	25.79	0.502	6.476	$< 0.001^{***}$
scam3	0.096	1.10	0.325	0.297	0.767
Warning (ref. = absent)					
Present	1.212	3.36	0.354	3.425	$< 0.001^{***}$

$^{***}p < 0.001$

Table G.4: Sentiment ratings by condition with Kruskal-Wallis Test results.

.

Item	N	generic	blocking	accordion	gmessages	long	H (df)	p	ϵ^2
Useful to you	288	3.78 (1.08)	3.48 (1.25)	3.69 (1.06)	3.59 (1.08)	3.73 (1.23)	2.28 (4)	.685	0.008
Grabbed your attention	288	4.25 (0.84)	4.42 (0.64)	4.29 (0.86)	4.08 (0.88)	4.45 (0.74)	7.35 (4)	.119	0.026
Easy to understand	288	4.90 (0.36)	4.80 (0.40)	4.67 (0.82)	4.71 (0.53)	4.79 (0.45)	5.94 (4)	.204	0.021
Annoying	288	1.29 (0.74)	1.56 (0.81)	1.41 (0.80)	1.25 (0.54)	1.34 (0.70)	10.39 (4)	.034*	0.036
Length of warning text	288	2.93 (0.31)	3.26 (0.49)	3.16 (0.52)	3.12 (0.33)	3.34 (0.54)	27.38 (4)	<.001	0.095
Trust in AI to generate accurate warnings	288	3.19 (0.90)	2.94 (0.93)	3.29 (0.90)	3.02 (0.68)	3.26 (0.75)	7.39 (4)	.117	0.026
Useful to the average person	288	4.12 (0.77)	4.06 (0.68)	4.03 (0.90)	4.00 (0.87)	4.11 (0.94)	1.25 (4)	.869	0.004
Preferred handling of a detected scam	288	3.27 (1.35)	2.74 (1.38)	3.00 (1.41)	2.76 (1.50)	3.02 (1.54)	5.48 (4)	.241	0.019
Comfort with AI analyzing messages	288	3.42 (1.25)	3.18 (1.37)	3.38 (1.28)	3.39 (1.07)	3.35 (1.27)	0.87 (4)	.929	0.003

Table G.3: Cross-tabulation of *dismiss* and *interact* outcomes by warning presence and by message

		Dismiss		Interact	
		Not Dismissed	Dismissed	No Interaction	Interaction
Warning presence	No Warning	169	314	434	49
	Warning	59	517	573	3
Message	Scam 1	142	211	309	44
	Scam 2	6	347	352	1
	Scam 3	80	273	346	7

G.3 Pre-study 2 Material

G.3.1 Survey Instrument

Note: Section headings and text in italics were not shown to participants, but are used to indicate page breaks, survey flow, and text entry boxes.

General Introduction

In this study, we will ask you to review text messages and determine whether they are real or fraudulent. You will view ten text messages using a simulated interface. We will ask you to specify whether you believe each message is legitimate or fraudulent.

A legitimate message is one that is genuinely sent by the person or organization it claims to be from and has a real, non-deceptive purpose (for example, an authentic delivery update or account notification).

A fraudulent message is one that is deceptive or misleading, such as a scam, phishing attempt, or impersonation, and is intended to trick the recipient into taking an action (for example, clicking a link, sharing personal information, or sending money).

Below is an example of the text messaging interface you will see in the survey. *Participants were shown the example message presented in Appendix G.3.2.*

1. It is important to review each message before proceeding. Which option best describes the subject of the text message above?
   (a) Online shopping promotion
   (b) A local political campaign
   (c) Code for logging into Amazon
   (d) A package shipment
2. Which of the following statements best describes the definition of a fraudulent message, according to the instruction text above?
   (a) A message sent by a business that uses aggressive or annoying advertising techniques to sell a product.
   (b) A message that uses deception or impersonation to trick the recipient into performing an action, such as clicking a link or sharing sensitive data.
   (c) Any message received from an unknown sender that contains a link.
   (d) An incorrectly addressed message sent to the wrong person due to a typo

Participants who answered either attention check incorrectly were screened out of the survey. Those who answered both correctly saw the following text.

We will now begin the message review task. Click the "Next page" button below when you are ready to proceed

Message Review

As described in the body of the paper, this section was repeated ten times with the stimuli presented in Appendix G.3.2. Participants were randomly assigned to view either the legitimate or the fraudulent variant of each message, and the ten messages were presented in a random order.

3. Imagine you *[text unique to each message e.g., "have a Discover bank account"]* and receive the text message below. Select the option corresponding to your evaluation of the message.
   The mockup text message interface was displayed here. See Appendix G.5.
   (a) Definitely legitimate
   (b) Probably legitimate
   (c) I'm not sure whether it is legitimate or fraudulent.
   (d) Probably fraudulent
   (e) Definitely fraudulent
4. Briefly describe why you selected this option. *[text entry]*

Demographics

Thank you for completing the main portion of the survey. We will now ask you a few optional demographic questions and then conclude the survey.

5. Select all of the following services that you have used in the last year. *Select all that apply.*
   (a) Amazon
   (b) CVS Pharmacy
   (c) Discover Bank
   (d) DocuSign
   (e) ID.me
   (f) Indeed
   (g) OpenTable
   (h) Overstock.com
   (i) T-Mobile
   (j) Venmo
   (k) None of the above
   (l) Prefer not to disclose
6. How do you describe your gender identity?
   (a) Man
   (b) Woman
   (c) Non-binary
   (d) Prefer to self-describe *[text entry]*

(e) Prefer not to disclose

7. What is your age in years?
   (a) 18 to 24
   (b) 25 to 34
   (c) 35 to 44
   (d) 45 to 54
   (e) 55 to 64
   (f) 65 to 74
   (g) 75 or older
   (h) Prefer not to disclose
8. What is your race? You may select multiple options
   (a) American Indian or Alaskan Native
   (b) Asian
   (c) Black or African American
   (d) Native Hawaiian or Other Pacific Islander
   (e) White
   (f) Prefer to self-describe *[text entry]*
   (g) Prefer not to disclose
9. Are you Hispanic or Latino?
   (a) Yes
   (b) No
   (c) Prefer not to disclose
10. What is your approximate household income?
   (a) Less than $25,000 per year
   (b) $25,000 to $49,999 per year
   (c) $50,000 to $99,999 per year
   (d) $100,000 to $199,999 per year
   (e) $200,000 or more per year
   (f) Prefer not to disclose
11. What is the highest level of school you have completed or the highest degree you have received?
   (a) Less than high school degree
   (b) High school graduate (high school diploma or equivalent including GED)
   (c) Some college but no degree

(d) Associate degree in college (2-year)

(e) Bachelor's degree in college (4-year)

(f) Graduate degree (e.g., MS, PhD, JD, etc.)

(g) Prefer not to disclose

12. Do you have a formal education in a computer-related field, such as computer science or IT? ("Formal education" could mean a completed degree or certificate, or classes or trainings you took towards a degree or certificate.)

    (a) Yes

    (b) No

    (c) I'm not sure

    (d) Prefer not to disclose

13. Do you have work experience in a computer-related field, such as computer science or IT?

    (a) Yes

    (b) No

    (c) I'm not sure

    (d) Prefer not to disclose

G.3.2 Message Stimuli

Formatting was shown to participants. Messages labeled as "intro" were displayed during the introduction section.

Intro Message

Sender: 98626
Message: Amazon: Your code is 134091. Don't share it. If you didn't request it, deny here https://amazon.com/a/c/r/SuLqMLDcVzjcmyrE0t19sZIfv

Discover Message

Sender: 347268
Legitimate: Discover Card Alert: A purchase occurred on your account outside of the United States or U.S. Territories. Notify Discover of your travel plans by logging into app.discover.com/TR Reply STOP to end texts.
Fraudulent: Discover Card Alert: A purchase occurred on your account outside of the United States or U.S. Territories. Notify Discover of your travel plans or report this transaction as suspicious by logging into https://discover.com.report-in.site/ad70d767e59d199b. Reply STOP to end texts.

OpenTable Message

Sender: 36246
Legitimate: You booked The Commoner for 2 on 3/06/26 Fri 5:15 PM. Reply 1 to confirm, 9 to cancel. See more: https://opentable.com/l/DCYdQyC91
Fraudulent: You booked The Commoner for 2 on 3/06/26 Fri 5:15 PM. Click here to confirm or cancel: https://confirm-opentable.com/l/DCYdQyC91

Overstock Message

Sender: 70343
Legitimate: Overstock Day is here with our best home deals of the year! TODAY ONLY get 20% off + 2x rewards + free shipping!* https://o.bttn.io/0nvkx2i4Cqp
Fraudulent: Overstock Day is here with our best home deals of the year! TODAY ONLY get 20% off + 2x rewards + free shipping!* https://overstock.com@n3593.org

Amazon Message

Sender: 262966
Legitimate: Delivery attempted: Unfortunately, we couldn't complete your Amazon delivery. We'll try again. More info at https://a.co/d/3N2oipH
Fraudulent: Delivery attempted: Unfortunately, we couldn't complete your Amazon delivery. We'll try again. More info at https://www.amązon.com/track?packageId=1Z999AA10123456784

T-Mobile Message

Sender: 456
Legitimate: IMPORTANT: Your T-Mobile bill is past due. Pay now or set up a Payment Arrangement at t-mobile.com/ppy to avoid service interruption & $20 restore fee per line. CA Debt Collector License. #11141-99
Fraudulent: IMPORTANT: Your T-Mobile bill is past due. Pay now or set up a Payment Arrangement at t-mobiile.net/ppy to avoid service interruption & $20 restore fee per line. CA Debt Collector License. #11141-99

CVS Pharmacy Message

Sender: 898287
Legitimate: CVS Pharmacy: Hi <Your Name>, we know life is busy. Remember to protect yourself this season, schedule easy vaccination appointments with us: i.cvs.com/izhW4NA8P
Fraudulent: CVS Pharmacy: Hi <Your Name>, we know life is busy. Remember to protect yourself this season, schedule easy vaccination appointments with us: https://get-apt.com/cvs

DocuSign Message

Sender: 65707

Legitimate: Sarah Connolly sent you a DocuSign document "Invoice" to view and sign: https://docusign.net/1xet

Fraudulent: Sarah Connolly sent you a DocuSign document "Invoice" to view and sign: https://docusign.net.sign

ID.me Message

Sender: 95246

Legitimate: This message is from ID.me. Your identity is being used to log in to ID.me Gateway. Please click this link to upload a picture or to report unauthorized use of your identity: https://verify.idmelabs.com/en/phones/c3f9f5a72e-b3424c9e73659a0a40b8df/authorize?vfp=r_zXElAp

Fraudulent: This message is from ID.me. Your identity is being used to log in to ID.me Gateway. Please click this link to upload a picture or to report unauthorized use of your identity: https://id-me.us/en/phones/c3f9f5a72e-b3424c9e73659a0a40b8df/authorize?vfp=r_zXElAp

Indeed Message

Sender: 36666

Legitimate: Indeed: A recruiter from Apple has viewed your application for "Project Manager (Job ID: 88421)". Log in to view the update: https://indeed.com/notifications

Fraudulent: Indeed: A recruiter from Apple has viewed your profile. Log in to view the update: https://indeed.com@hiring-portal@apply-now-jobs.com/#@indeed.com

Venmo Message

Sender: 86753

Legitimate: Jack Sims requests $23.62 – Dinner. To pay, go to Venmo and complete the request. Don't have the Venmo app? Download it at venmo.com

Fraudulent: Jack Sims requests $23.62 – Dinner. To pay, go to Venmo and complete the request: https://venmoapp.com/73659a0a40b8df

G.3.3 Additional Results

Table G.5: Percentage of participants labeling each message as fraudulent, legitimate, or unsure, by message and version.

		Scam version			Legitimate version		
Message	Used By	Fraud.	Unsure	Legit.	Fraud.	Unsure	Legit.
Amazon	94.5	55.3	11.7	33.0	75.3	6.2	18.6
CVS Pharmacy	55.5	37.5	5.2	57.3	32.7	15.4	51.9
Discover Bank	19.5	47.5	18.2	34.3	40.6	18.8	40.6
DocuSign	40.0	48.0	24.5	27.5	37.8	29.6	32.7
ID.me	38,5	65.3	11.9	22.8	73.7	14.1	12.1
Indeed	44.5	52.9	15.4	31.7	14.6	13.5	71.9
OpenTable	23.0	12.0	21.0	67.0	1.0	8.0	91.0
Overstock.com	7.0	55.9	6.9	37.3	66.3	13.3	20.4
T-Mobile	31.0	74.0	9.0	17.0	54.0	16.0	30.0
Venmo	51.0	40.9	24.7	34.4	10.3	16.8	72.9

G.4 Pre-study 3 Material

G.4.1 Survey Instrument

Note: Section headings and text in italics were not shown to participants, but are used to indicate page breaks, survey flow, and text entry boxes.

General Introduction

In this study, we will ask you to use and provide feedback on a new text messaging system that is currently under development. You will first view ten text messages using the new interface. We will ask you to specify what you are most likely to do in response to each message. Below is an example of the new text messaging system. You may interact with it by pressing the "three dots" in the top right corner.
The mockup text message interface was displayed with the intro message here. See Appendix G.4.2.

1. It is important to review each message before proceeding. Which option best describes the subject of the text message above?
   (a) Online shopping promotion
   (b) A local political campaign
   (c) Code for logging into Amazon
   (d) A package shipment

Warning Introduction (Non-control conditions)

Some messages will be displayed with a red warning box, as shown below. This box is generated by the new text messaging system. You may interact with the warning by pressing the "Learn more" or "Report & Block" button.

While these warnings provide additional context, you should continue to evaluate all messages based on your own observations and intuition.

The mockup text message interface was displayed with the warning intro message and the appropriate warning here. See Appendix G.4.2 and Appendix G.5.

2. It is important to review each message before proceeding. Which option best describes the subject of the text message above?
   (a) An arrest threat
   (b) An unpaid toll
   (c) A severe virus detected on the recipient's device
   (d) A package shipment

Participants who answered either attention check incorrectly were screened out of the survey. Those who answered correctly saw the following text.

We will now begin the message review task. You will view ten messages. Click the "Next page" button below when you are ready to proceed

Message Review

As described in the body of the paper, this section was repeated ten times with the stimuli presented in Appendix G.4.2. Participants were randomly assigned to view either the legitimate or the fraudulent variant of each message, and the ten messages were presented in a random order. The scenario text and the second and third response options were adapted to the brand and to whether the message contained a link or a phone number.

3. Imagine you have a Discover bank account and receive the text message below. Select the option corresponding to what you believe you are **most likely** to do.
   The mockup text message interface was displayed with the appropriate warning here. See Appendix G.5.
   (a) Reply to the message. Please specify what you would say in reply *[text entry]*
   (b) Call the phone number listed within the message
   (c) Exit the message and visit Discover directly via their official app or website
   (d) Ignore the message without doing anything
   (e) Click the "three dots" to open the menu and select "delete."
   (f) Click the "three dots" to open the menu and select "report & block."
   (g) *(Displayed when a long, generic, or accordion warning was present)* Click the "Report & Block" button
   (h) Search online for information about this message

(i) Other. Please specify what you would do *[text entry]*

4. Why do you think this is what you would do? *[text entry]*

Sentiment Likert Scale

This subsection and the two that follow were shown only to participants in the three warning conditions.

We will now ask you several questions about the warning design you saw. For reference, an example of the warning is shown below.

The mockup text message interface was displayed with the example warning message and the appropriate warning here. See Appendix G.4.2 and Appendix G.5.

5. How useful do you think these warnings were for you?
   (a) 1: Not at all useful
   (b) 2: Slightly useful
   (c) 3: Moderately useful
   (d) 4: Very useful
   (e) 5: Extremely useful
6. How effective were the warnings at grabbing your attention?
   (a) 1: Not at all effective
   (b) 2: Slightly effective
   (c) 3: Moderately effective
   (d) 4: Very effective
   (e) 5: Extremely effective
7. How easy or difficult was it to understand the text of the warnings?
   (a) 1: Very difficult
   (b) 2: Somewhat difficult
   (c) 3: Neither easy nor difficult
   (d) 4: Somewhat easy
   (e) 5: Very easy
8. To what extent did you find the warnings annoying?
   (a) 1: Not at all annoying
   (b) 2: Slightly annoying
   (c) 3: Moderately annoying
   (d) 4: Very annoying
   (e) 5: Extremely annoying
9. How would you rate the length of the warning text?
   (a) 1: Much too short

(b) 2: Somewhat too short
(c) 3: Just the right length
(d) 4: Somewhat too long
(e) 5: Much too long

10. How much do you trust AI to generate accurate warnings?
    (a) 1: Do not trust at all
    (b) 2: Trust slightly
    (c) 3: Trust moderately
    (d) 4: Trust a lot
    (e) 5: Trust completely
11. How useful do you think these warnings would be for the average person?
    (a) 1: Not at all useful
    (b) 2: Slightly useful
    (c) 3: Moderately useful
    (d) 4: Very useful
    (e) 5: Extremely useful
12. If your phone identifies a text as a scam, how would you prefer to handle the message?
    (a) 1: Strongly prefer to see the message anyway
    (b) 2: Somewhat prefer to see the message anyway
    (c) 3: No preference
    (d) 4: Somewhat prefer it be moved to a spam folder without viewing it
    (e) 5: Strongly prefer it be moved to a spam folder without viewing it
13. How comfortable are you with an AI analyzing your instant messages?
    (a) 1: Very uncomfortable
    (b) 2: Somewhat uncomfortable
    (c) 3: Neither comfortable nor uncomfortable
    (d) 4: Somewhat comfortable
    (e) 5: Very comfortable

Warning Changes

For reference, an example of the warning is shown below.
The mockup text message interface was displayed with the example warning message and the appropriate warning here. See Appendix G.4.2 and Appendix G.5.

14. Are there any changes you would make to the warning design?
    (a) Yes

(b) No

(c) I am not sure

15. *[If answer yes to the previous question]* What changes would you make to the warning design? *[text entry]*

Alternate Designs

Please consider the following warnings, which include both the warning you saw during the message review task and alternate designs. We will ask you to select the design you believe would be most useful.
The Simple Alert, Expandable Alert, and Detailed Alert designs were displayed with the example warning message. See Appendix G.4.2 and Appendix G.5. The names were chosen to reduce bias in participant responses. Detailed = Long, Expandable = Accordion, Simple = Generic

16. Which warning format would be the most useful to you in assessing fraudulent instant messages?
    (a) Detailed Alert
    (b) Expandable Alert
    (c) Simple Alert
    (d) None of the above. I would prefer no warning
17. Why do you think this warning would be the most useful? If you would prefer no warning, please briefly explain why *[text entry]*

Demographics

Thank you for completing the main portion of the survey. We will now ask you a few optional demographic questions and then conclude the survey.

18. Select all of the following services that you have used in the last year. *Select all that apply.*
    (a) Amazon
    (b) CVS Pharmacy
    (c) Discover Bank
    (d) DocuSign
    (e) ID.me
    (f) Indeed
    (g) OpenTable
    (h) Kohl's
    (i) T-Mobile
    (j) Venmo
    (k) None of the above

(l) Prefer not to disclose

19. How do you describe your gender identity?
    (a) Man
    (b) Woman
    (c) Non-binary
    (d) Prefer to self-describe *[text entry]*
    (e) Prefer not to disclose
20. What is your age in years?
    (a) 18 to 24
    (b) 25 to 34
    (c) 35 to 44
    (d) 45 to 54
    (e) 55 to 64
    (f) 65 to 74
    (g) 75 or older
    (h) Prefer not to disclose
21. What is your race? You may select multiple options
    (a) American Indian or Alaskan Native
    (b) Asian
    (c) Black or African American
    (d) Native Hawaiian or Other Pacific Islander
    (e) White
    (f) Prefer to self-describe *[text entry]*
    (g) Prefer not to disclose
22. Are you Hispanic or Latino?
    (a) Yes
    (b) No
    (c) Prefer not to disclose
23. What is your approximate household income?
    (a) Less than $25,000 per year
    (b) $25,000 to $49,999 per year
    (c) $50,000 to $99,999 per year
    (d) $100,000 to $199,999 per year
    (e) $200,000 or more per year

(f) Prefer not to disclose

24. What is the highest level of school you have completed or the highest degree you have received?
    (a) Less than high school degree
    (b) High school graduate (high school diploma or equivalent including GED)
    (c) Some college but no degree
    (d) Associate degree in college (2-year)
    (e) Bachelor's degree in college (4-year)
    (f) Graduate degree (e.g., MS, PhD, JD, etc.)
    (g) Prefer not to disclose
25. Do you have a formal education in a computer-related field, such as computer science or IT? ("Formal education" could mean a completed degree or certificate, or classes or trainings you took towards a degree or certificate.)
    (a) Yes
    (b) No
    (c) I'm not sure
    (d) Prefer not to disclose
26. Do you have work experience in a computer-related field, such as computer science or IT?
    (a) Yes
    (b) No
    (c) I'm not sure
    (d) Prefer not to disclose

G.4.2 Message and Warning Stimuli

Formatting was shown to participants. Messages labeled as "intro" were displayed during the introduction section. Every warning used the headline "This is likely a scam." and was followed by the note "This message was generated by AI."

Intro Message

Sender: 98626
Message: Amazon: Your code is 134091. Don't share it. If you didn't request it, deny here https://amazon.com/a/c/r/SuLqMLDcVzjcmyrE0t19sZIfv

Warning Intro Message

Sender: *[Redacted # with D.C. Area Code]*

Message: WARNING: (Criminal Investigation Division) I.R.S is filing lawsuit against you, for more information call on *[Redacted # with Virginia Area Code]* on urgent basis, Otherwise your arrest warrant will be forwarded to your local police department and your property and bank accounts and social benifits will be frozen by government.
Warning: Do not call the number — real IRS notices never arrive by text.

- The message uses extreme threats like arrest and frozen accounts to scare you into reacting quickly.
- The IRS does not contact people by text about lawsuits or warrants, and the wording contains obvious mistakes.
- This is a scare tactic. Ignore and delete the message — do not call the number.

Sentiment Testing Message

Shown during the warning feedback, warning changes, and alternate design sections.
Sender: *[Redacted # with California Area Code]*
Message: This is a gift to express our gratitude for your exceptional payment record. T Mobile POSC9.XYZ/YgSpZZjZoh
Warning: Don't open the link — this "gift" message isn't really from T-Mobile.

- The message promises a reward you didn't ask for, a common trick scammers use to get clicks.
- The link goes to a fake website that doesn't match T-Mobile's real web addresses.
- Real companies don't send gifts this way. Delete the message and avoid the link.

Discover Message

Sender: 347268
Legitimate message: Discover Card Account Alert: A purchase at STARBUCKS STORE 6192 occurred outside the US. If this wasn't you, call 1-800-347-2683 immediately. Text STOP to quit.
False positive warning: Do not call the number in this text; if you're worried about the charge, contact Discover using the number on the back of your card or the official app.

- This message looks suspicious because it claims there was an overseas purchase and pushes you to call immediately, which can pressure people to act without checking first.
- Even though it mentions Discover, you should not trust contact details in the text itself unless you verify them another way.
- Do not use the number in the message. Instead, open the official Discover app or call the number printed on the back of your card to check whether the alert is real.

Fraudulent message: Discover Card Account Alert: A purchase at STARBUCKS STORE 6192 occurred outside the US. If this wasn't you, call *[Redacted # with PA Area Code]* immediately. Text STOP to quit.
Fraudulent warning: Do not call the number in this text; check your Discover account using the official app or the number on the back of your card instead.

- This message looks suspicious because it claims to be from Discover but tells you to call a regular phone number that does not clearly look like an official Discover contact.
- It also uses an urgent warning about a purchase outside the US to push you to act quickly before checking whether the message is real.
- Do not reply or call the number in the text. If you want to verify the charge, contact Discover using the number on the back of your card or the official app.

OpenTable Message

Sender: 36246
Legitimate message: You booked The Commoner for 2 on 3/06/26 Fri 5:15 PM. Reply 1 to confirm, 9 to cancel. See more: https://opentable.com/l/DCYdQyC91
False positive warning: Do not reply or tap the link unless you're sure you made this restaurant booking; instead, check your reservation directly in the OpenTable app or website.

- This message mentions a booking at The Commoner and includes a link to opentable.com, but if you didn't make this reservation, it could be a scam trying to get you to interact.
- Scammers often use unexpected booking texts to get people to tap a link or reply quickly, even when the message looks normal.
- Do not reply or open the link from the text. If you want to check, go to OpenTable directly by opening the app or typing the site name yourself.

Fraudulent message: You booked The Commoner for 2 on 3/20/26 Fri 5:15 PM. Click here to confirm or cancel: https://confirm-opentable.com/l/DCYdQyC91
Fraudulent warning: Don't tap the link—check your reservation in the official OpenTable app or by typing opentable.com yourself.

- This message is suspicious because the link does not use OpenTable's normal website name; it uses a lookalike name instead.
- It also pushes you to act on a reservation, but the link behavior doesn't clearly match a real reservation confirmation page for The Commoner.
- What to do: Do not click the link or enter any information; check the booking in the OpenTable app, go to opentable.com yourself, or call The Commoner using a trusted number.

Kohl's Message

Sender: 56457
Legitimate message: KOHL'S: Don't miss this. Stack the savings with $10 + 20% off! PLUS, earn Kohl's Cash! Shop now: http://khls.us/LCI5ba | STOP to end
False positive warning: Do not tap the link or reply—delete this message and go to Kohl's another way if you want to check any real offers.

- This message looks suspicious because it pushes you to shop now with a discount and includes a shortened link: khls.us/LCI5ba.

- Scam texts often use sales pressure and links to get you to tap before you think, even when the message appears to come from a real store like Kohl's.
- Do not tap the link or reply. Delete the message, and if you want to see Kohl's deals, open the Kohl's app or type the store's name into your browser yourself.

Fraudulent message: KOHL'S: Don't miss this. Stack the savings with $10 + 20% off! PLUS, earn Kohl's Cash! Shop now: https://kohls.com@n3593.org | STOP to end

Fraudulent warning: Do not tap the link—if you want a Kohl's deal, go to the official Kohl's app or type kohls.com yourself.

- This message is suspicious because it pretends to be from Kohl's, but the link actually leads to n3593.org, not kohls.com.
- The message uses a sales push like "Don't miss this" and big discounts to get you to act quickly, which is common in scam texts.
- Do not tap the link or enter any information; if you want to check for real offers, open the official Kohl's app or type kohls.com yourself.

Amazon Message

Sender: 262966

Legitimate message: Delivery attempted: Unfortunately, we couldn't complete your Amazon delivery. We'll try again. More info at https://a.co/d/3N2oipH

False positive warning: Do not tap the link; check your Amazon order directly in the Amazon app or website instead.

- This message claims there was a problem with an Amazon delivery and pushes you to open a link, which is a common scam tactic.
- The message creates pressure to act quickly about a missed package, and link-based delivery alerts can be used to trick people into signing in or sharing information.
- Do not use the link. Open the Amazon app or official Amazon website yourself to check your order, and delete the text if nothing looks wrong.

Fraudulent message: Delivery attempted: Unfortunately, we couldn't complete your Amazon delivery. We'll try again. More info at https://www.amązon.com/track?packageId=582-9104736-2584109

Fraudulent warning: Do not tap the link; if you want to check your delivery, open the Amazon app or type amazon.com yourself.

- This message looks suspicious because it pretends to be from Amazon but uses a lookalike website name: amązon.com instead of the normal amazon.com.
- The link leads to an Amazon-looking sign-in page, which is not normal for simply checking a delivery and may be trying to steal your login details.
- Do not tap the link or enter any information; check your order only in the Amazon app or by typing amazon.com yourself.

T-Mobile Message

Sender: 262966 (legitimate), 456 (fraudulent)
Legitimate message: IMPORTANT: Your T-Mobile bill is past due. Pay now or set up a Payment Arrangement at t-mobile.com/ppy to avoid service interruption & $20 restore fee per line. CA Debt Collector License. #11141-99
False positive warning: Do not pay from this text—open the T-Mobile app or go to T-Mobile's official website yourself to check your bill instead.

- This message looks suspicious because it pressures you to pay right away to avoid service interruption and extra fees, which is a common scam tactic.
- Even though the message mentions T-Mobile and includes a payment page, you should not trust a payment link in a text when money is involved.
- Do not tap the link or reply; instead, open the T-Mobile app or type T-Mobile's official website into your browser yourself, and contact T-Mobile using the number on your bill if you want to verify it.

Fraudulent message: IMPORTANT: Your T-Mobile bill is past due. Pay now or set up a Payment Arrangement at t-mobiile.net/ppy to avoid service interruption & $20 restore fee per line. CA Debt Collector License. #11141-99
Fraudulent warning: Do not tap the link—if you need to check your bill, open the official T-Mobile app or type t-mobile.com yourself.

- This message is suspicious because it claims to be from T-Mobile but sends you to t-mobiile.net/ppy, which is not T-Mobile's official site.
- It also tries to rush you by warning about service interruption and extra fees, which scammers often do to push quick action.
- Do not tap the link or reply. Check your account only through the official T-Mobile app or by typing t-mobile.com yourself.

CVS Pharmacy Message

Sender: 898287
Legitimate message: CVS Pharmacy: Hi <Your Name>, we know life is busy. Remember to protect yourself this season, schedule easy vaccination appointments with us: i.cvs.com/izhW4NA8P
False positive warning: Do not tap the link or reply—this message may look like a CVS reminder, but treat it as a scam and delete it.

- This message mentions CVS Pharmacy and includes a link, which can make it seem trustworthy, but scam texts often pretend to be familiar brands.
- The message tries to get you to tap a link to schedule something, which is a common way scams lure people into unsafe websites.
- Do not tap the link, reply, or share any information. If you want to check, open the official CVS app or website yourself instead of using the message.

Fraudulent message: CVS Pharmacy: Hi <Your Name>, we know life is busy. Remember to protect yourself this season, schedule easy vaccination appointments with us: https://get-apt.com/cvs

Fraudulent warning: Do not tap the link in this CVS message; if you want to book a vaccine, go directly to cvs.com instead.

- This message looks suspicious because it says it is from CVS, but the link goes to get-apt.com, not an official CVS website like cvs.com.
- That website name is a different, very new site, which is not what you would expect from a major pharmacy sending official appointment links.
- Do not tap the link or share any personal or insurance information. If you want to schedule a vaccine, open cvs.com yourself or contact CVS using an official phone number.

DocuSign Message

Sender: 65707
Legitimate message: Sarah Connolly sent you a DocuSign document "Invoice" to view and sign: https://docusign.net/1xetXRme-slkNFuo
False positive warning: Do not tap the link or sign anything from this message unless you were personally expecting an invoice from Sarah Connolly.

- This message mentions an invoice and asks you to view and sign, which scammers often use to pressure people into acting quickly.
- Even though the message looks like it is from DocuSign, you should be careful if you do not know Sarah Connolly or were not expecting any document.
- If this seems unexpected, do not open the link; instead, contact the sender or DocuSign through the app or website you normally use to check whether it is real.

Fraudulent message: Sarah Connolly sent you a DocuSign document "Invoice" to view and sign: https://docusign.net.signing-portal.secure-invoice.1xetXRme-slkNFuo-99214.etz5811-eiue348wi0li27dh8jtkku.xyz
Fraudulent warning: Do not tap the link—if this invoice is real, open DocuSign from the official app or website instead.

- This message looks suspicious because it pretends to be from DocuSign but the link goes to a long, unusual .xyz address instead of DocuSign's normal site.
- It tries to get you to open and sign an invoice, which can pressure people to act quickly, and the page appears to ask for your email sign-in details.
- Do not tap the link or reply. If you were expecting a real document, open DocuSign directly from its official app or type the company's website yourself. If you already entered your details, change your password right away.

ID.me Message

Sender: 95246
Legitimate message: This message is from ID.me. Your identity is being used to log in to ID.me Gateway. Please click this link to upload a picture or to report unauthorized use of your identity: https://verify.idmelabs.com/en/phones/c3f9f5a72e-b3424c9e73659a0a40b8df/authorize?vfp=r_zXElAp

False positive warning: Do not tap the link in this text; go to ID.me another way to check your account or report the alert safely.

- This message claims to be from ID.me, but it asks you to use a link in the text during a stressful security situation, which is a common scam tactic.
- The message creates urgency by saying your identity is being used to log in and pushes you to act right away through a link, so you should be cautious.
- If you're worried this could be real, do not use the link in the message; instead, open the ID.me app or type the company's official site yourself to check your account and report the issue.

Fraudulent message: This message is from ID.me. Your identity is being used to log in to ID.me Gateway. Please click this link to upload a picture or to report unauthorized use of your identity: https://id-me.us/en/phones/c3f9f5a72e-b3424c9e73659a0a40b8df/authorize?vfp=r_zXElAp

Fraudulent warning: Do not tap the link in this message; if you're worried about your ID.me account, open the official ID.me site or app yourself instead.

- This message looks suspicious because it says it is from ID.me, but the link goes to id-me.us instead of ID.me's official site.
- It also uses an urgent warning about someone using your identity to push you to act quickly, which is a common scam tactic.
- Do not click the link or reply; instead, go to the official ID.me website or app yourself to check your account, and if you already used the link, change your password right away.

Indeed Message

Sender: 36666

Legitimate message: Indeed: A recruiter from Apple has viewed your application for "Project Manager (Job ID: 88421)". Log in to view the update: https://indeed.com/notifications

False positive warning: Do not tap the link or reply—this message may be trying to trick you with a fake job update.

- This message looks suspicious because it creates excitement about a recruiter from Apple viewing your application to get you to act quickly.
- Even though the message mentions Indeed, scam texts often copy real company names and job-related details to seem believable.
- Do not tap the link or reply; instead, open the Indeed app or type the company's website into your browser yourself if you want to check your account safely.

Fraudulent message: Indeed: A recruiter from Apple has viewed your profile. Log in to view the update: https://indeed.com@hiring-portal@apply-now-jobs.com/#@indeed.com

Fraudulent warning: Do not tap the link—if you want to check your Indeed or Apple account, open the app or type the official site yourself.

- This message is suspicious because it claims to be about Indeed and Apple, but the link actually goes to apply-now-jobs.com, not an official Indeed or Apple site.

- The message tries to get you to log in right away, which is a common trick used to steal usernames and passwords.
- Do not tap the link or enter your login details; instead, open Indeed or Apple directly from their official app or by typing the site name yourself.

Venmo Message

Sender: 86753

Legitimate message: Jack Sims requests $23.62 - Dinner
To pay, go to Venmo and complete the request. Don't have the Venmo app? Download it at venmo.com

False positive warning: Do not pay or tap anything unless you personally know Jack Sims and were expecting this Venmo request.

- This message is suspicious because it asks for money and mentions Venmo, which scammers often use to make requests look normal.
- Be careful even though the message says venmo.com and the amount is small, because scam payment requests can still appear believable.
- Do not pay, tap the link, or download anything from this message; if you think it might be real, open the Venmo app yourself and check for the request there.

Fraudulent message: Jack Sims requests $23.62 - Dinner
To pay, go to Venmo and complete the request: https://venmoapp.com/73659a0a40b8df

Fraudulent warning: Do not tap the link or pay from this text—open the real Venmo app or go to venmo.com to check whether this request is real.

- This message looks suspicious because it asks for money using a Venmo-like website, but the link goes to venmoapp.com instead of venmo.com.
- The payment link also looks unusual and does not match a normal Venmo request, which is a common sign of a fake payment message.
- Do not tap the link or send money from this text. If you want to check the request, open the official Venmo app or type venmo.com yourself.

G.4.3 Additional Results

Table G.6: Nested mixed-effects logistic regression model comparisons for dismissal and interact outcomes for scam messages.

Outcome	Model	k	AIC	BIC	$\log L$	Deviance	χ^2 (df)	p
Dismiss	Base	11	2190.9	2251.1	−1084.5	2168.9		
	+ Warning present	12	2086.0	2151.6	−1031.0	2062.0	106.95 (1)	$< .001$
	+ Warning condition	14	2080.8	2157.3	−1026.4	2052.8	9.19 (2)	.010
Interact	Base	11	1102.2	1162.3	−540.08	1080.15		
	+ Warning present	12	1026.0	1091.6	−501.01	1002.03	78.12 (1)	$< .001$
	+ Warning condition	14	1026.2	1102.8	−499.11	998.22	3.81 (2)	.149

Table G.7: Nested mixed-effects logistic regression model comparisons for dismissal and interact outcomes for legitimate messages

Outcome	Model	k	AIC	BIC	$\log L$	Deviance	χ^2 (df)	p
Dismiss	Base	11	2046.4	2106.6	−1012.22	2024.4		
	+ Warning present	12	2001.5	2067.1	−988.74	1977.5	46.97 (1)	$< .001$
	+ Warning condition	14	2003.3	2079.8	−987.65	1975.3	2.18 (2)	.336
Interact	Base	11	1323.5	1383.6	−650.72	1301.5		
	+ Warning present	12	1287.5	1353.1	−631.76	1263.5	37.93 (1)	$< .001$
	+ Warning condition	14	1290.0	1366.5	−630.98	1262.0	1.56 (2)	.458

Table G.8: Sentiment ratings by condition with Kruskal-Wallis Test results.

Item	N	generic	accordion	long	H (df)	p	ϵ^2
Useful to you	260	3.69 (0.98)	3.61 (1.23)	3.85 (1.09)	2.34 (2)	.311	0.009
Grabbed your attention	260	4.19 (0.87)	4.25 (0.92)	4.35 (0.82)	1.80 (2)	.406	0.007
Easy to understand	260	4.76 (0.53)	4.75 (0.57)	4.79 (0.58)	0.43 (2)	.807	0.002
Annoying	260	1.54 (0.80)	1.45 (0.74)	1.50 (0.95)	1.42 (2)	.491	0.005
Length of warning text	260	3.00 (3.41)	0.06 (3.31)	3.36 (0.51)	36.95 (2)	<.001	0.143
trust in AI to generate accurate warnings?	260	3.06 (0.88)	3.17 (0.84)	3.27 (0.89)	2.52 (2)	.283	0.010
Useful to the average person	260	4.00 (0.96)	4.11 (0.96)	4.21 (0.88)	2.09 (2)	.353	0.008
Preferred handling of a detected scam	260	2.79 (1.49)	2.83 (1.48)	2.65 (1.40)	0.52 (2)	.770	0.002
Comfort with AI analyzing messages	259	3.21 (1.15)	3.33 (1.30)	3.50 (1.29)	2.97 (2)	.227	0.011

Table G.9: Mixed-effects logistic regression predicting whether a participant interacted with a simulated scam message. The Discover message is the reference level for message identity, and the control condition is the reference level for warning presence. Odds ratios are e^{β}. $N = 1750$ observations from 350 participants; random intercept variance $= 3.049$ ($SD = 1.746$).

	β	OR	$S.E.$	z	p
(Intercept)	-2.2645	0.10	0.3609	-6.274	$< 0.001^{***}$
Message identity (ref. = Discover)					
OpenTable	1.5141	4.55	0.3982	3.803	$< 0.001^{***}$
Kohl's	-0.2204	0.80	0.4450	-0.495	0.620
Amazon	0.3941	1.48	0.4171	0.945	0.345
T-Mobile	-0.3609	0.70	0.4558	-0.792	0.428
CVS	-1.1156	0.33	0.5173	-2.156	0.031^{*}
DocuSign	0.8567	2.36	0.4037	2.122	0.034^{*}
ID.me	-1.3056	0.27	0.5279	-2.473	0.013^{*}
Indeed	-0.3396	0.71	0.4465	-0.761	0.447
Venmo	0.4401	1.55	0.4048	1.087	0.277
Warning					
Warning present	-1.9716	0.14	0.2433	-8.103	$< 0.001^{***}$

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table G.10: Mixed-effects logistic regression predicting whether a participant dismissed a simulated scam message. The Discover message is the reference level for message identity, and the control condition is the reference level for warning design. Odds ratios are e^{β}. $N = 1750$ observations from 350 participants; random intercept variance $= 1.064$ ($SD = 1.031$).

	β	OR	$S.E.$	z	p
(Intercept)	-2.0064	0.13	0.2293	-8.750	$< 0.001^{***}$
Message identity (ref. = Discover)					
OpenTable	0.7769	2.17	0.2652	2.929	0.003^{**}
Kohl's	2.6572	14.26	0.2903	9.153	$< 0.001^{***}$
Amazon	0.6398	1.90	0.2683	2.384	0.017^{*}
T-Mobile	0.9580	2.61	0.2652	3.612	$< 0.001^{***}$
CVS	2.9146	18.44	0.3000	9.715	$< 0.001^{***}$
DocuSign	0.8827	2.42	0.2648	3.334	$< 0.001^{***}$
ID.me	1.5166	4.56	0.2680	5.659	$< 0.001^{***}$
Indeed	2.0313	7.62	0.2757	7.368	$< 0.001^{***}$
Venmo	0.8077	2.24	0.2653	3.045	0.002^{**}
Warning design (ref. = control)					
Generic	1.2259	3.41	0.1950	6.288	$< 0.001^{***}$
Accordion	1.1600	3.19	0.1933	6.001	$< 0.001^{***}$
Long	1.8124	6.13	0.2059	8.804	$< 0.001^{***}$

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table G.11: Mixed-effects logistic regression predicting whether a participant interacted with a legitimate message. The Discover message is the reference level for message identity, and the control condition is the reference level for warning presence. Odds ratios are e^{β}. $N = 1750$ observations from 350 participants; random intercept variance $= 1.599$ ($SD = 1.264$).

	β	OR	$S.E.$	z	p
(Intercept)	-2.4383	0.09	0.2911	-8.377	$< 0.001^{***}$
Message identity (ref. = Discover)					
OpenTable	2.6354	13.95	0.3422	7.701	$< 0.001^{***}$
Kohl's	-0.2727	0.76	0.3906	-0.698	0.485
Amazon	0.3122	1.37	0.3553	0.879	0.380
T-Mobile	-0.4287	0.65	0.4119	-1.041	0.298
CVS	-0.5309	0.59	0.4033	-1.316	0.188
DocuSign	1.1957	3.31	0.3377	3.541	$< 0.001^{***}$
ID.me	-1.2217	0.30	0.4832	-2.528	0.011^{*}
Indeed	0.9201	2.51	0.3426	2.686	0.007^{**}
Venmo	-0.5162	0.60	0.4128	-1.250	0.211
Warning					
Warning present	-1.7436	0.17	0.3342	-5.217	$< 0.001^{***}$

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table G.12: Mixed-effects logistic regression predicting whether a participant dismissed a legitimate message. The Discover message is the reference level for message identity, and the control condition is the reference level for warning presence. Odds ratios are e^{β}. $N = 1750$ observations from 350 participants; random intercept variance $= 0.505$ ($SD = 0.711$).

	β	OR	$S.E.$	z	p
(Intercept)	-2.0520	0.13	0.2358	-8.702	$< 0.001^{***}$
Message identity (ref. = Discover)					
OpenTable	0.8904	2.44	0.2904	3.066	0.002^{**}
Kohl's	3.1611	23.60	0.3043	10.388	$< 0.001^{***}$
Amazon	0.8651	2.38	0.2869	3.015	0.003^{**}
T-Mobile	0.4795	1.62	0.2985	1.606	0.108
CVS	3.1745	23.92	0.3040	10.443	$< 0.001^{***}$
DocuSign	0.9604	2.61	0.2899	3.313	$< 0.001^{***}$
ID.me	1.3730	3.95	0.2842	4.831	$< 0.001^{***}$
Indeed	1.0217	2.78	0.2891	3.535	$< 0.001^{***}$
Venmo	1.1361	3.12	0.2872	3.956	$< 0.001^{***}$
Warning					
Warning present	1.0761	2.93	0.1598	6.735	$< 0.001^{***}$

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

G.5 Pre-study Interface Designs

These interfaces were interactive:

- Pressing the "Learn More" buttons in the *generic*, *spam*, *long*, or *blocking* warnings resulted in a pop-up box appearing which stated "This would link to a page explaining the warning system in a real app."
- Pressing the "Report & Block" button in the *generic*, *long*, *accordion*, or *blocking* warnings would result in a pop-up stating "This would report and block the message in a real app."
- Pressing the "Not spam" button in the *spam* warning would cause the warning to disappear.
- Pressing the "Proceed anyway" button in the *blocking* warning would cause the warning and the grey background to disappear.
- Pressing the "three dots" in the top right corner of the mockup interface would cause a menu to appear with the following options: "Details", "Archive", "Delete", or "Report & Block." Clicking these options did not do anything.

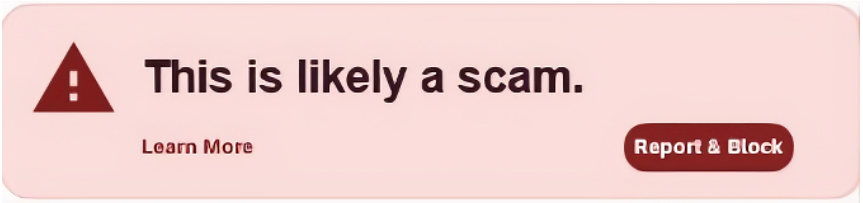


Figure G.2: Pre-study generic warning

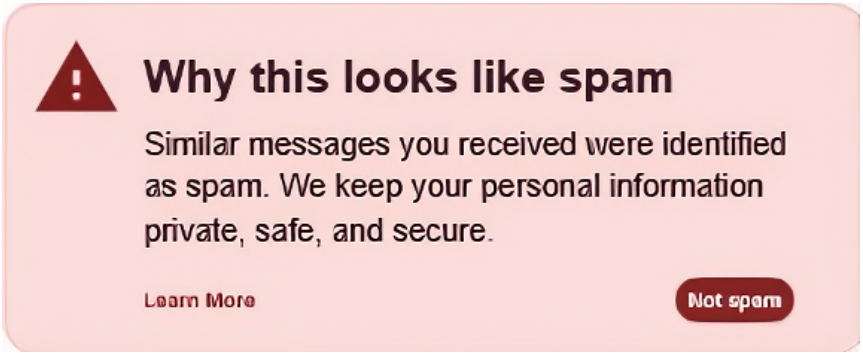


Figure G.3: Pre-study spam warning

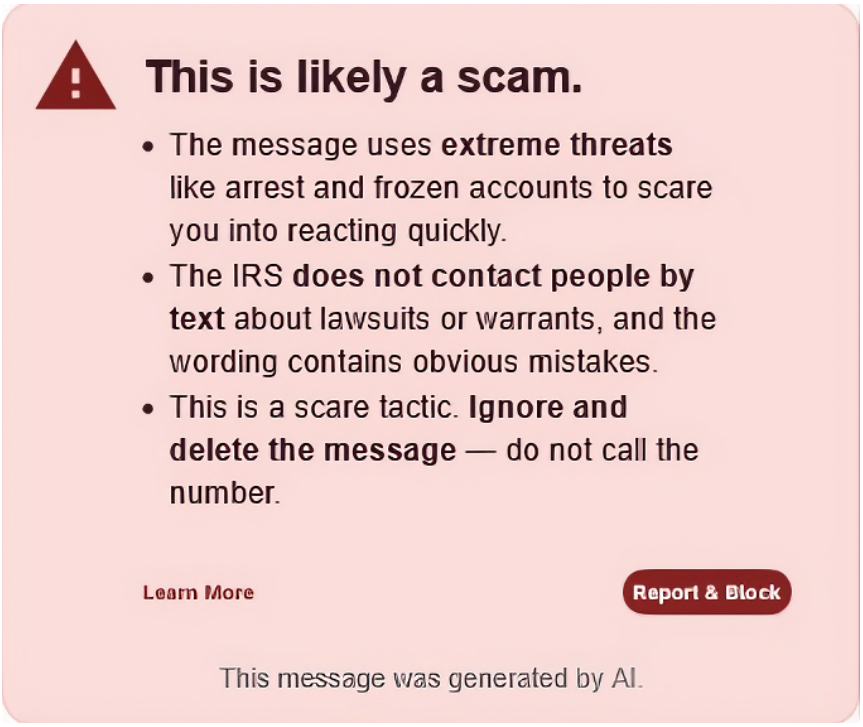


Figure G.4: Pre-study long warning

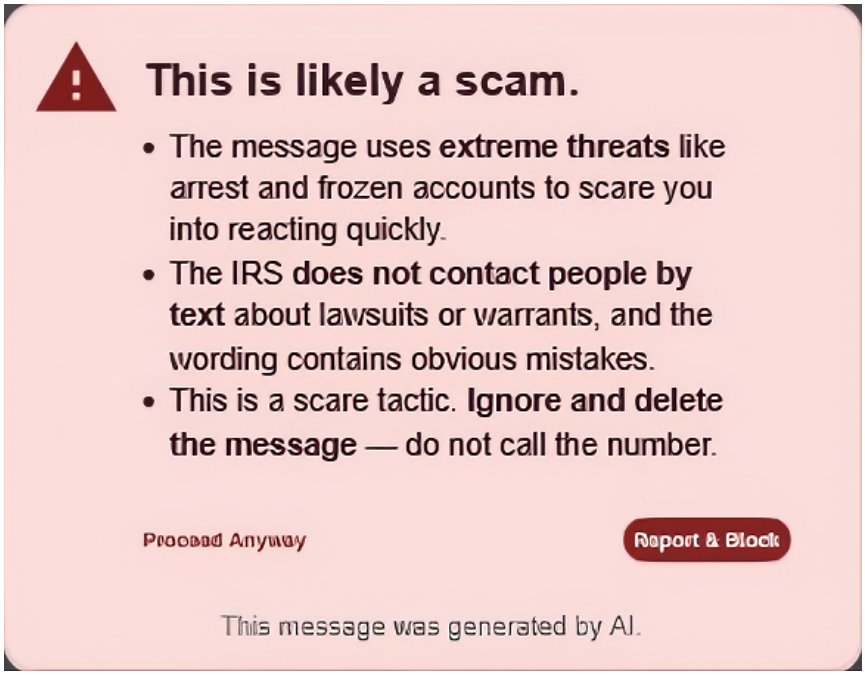


Figure G.5: Pre-study blocking warning

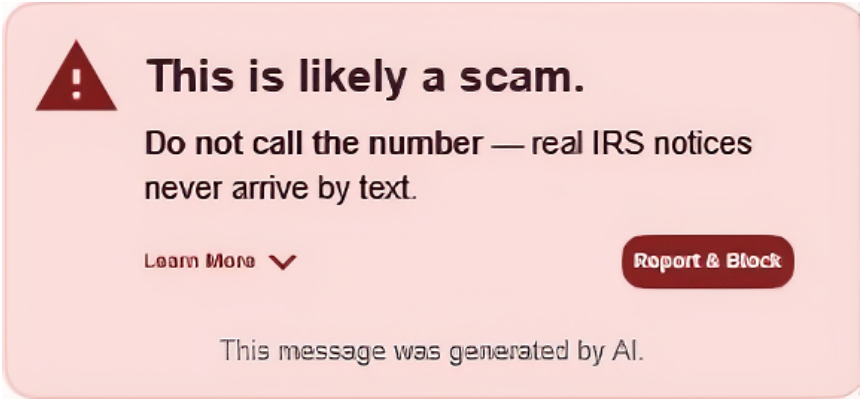


Figure G.6: Pre-study accordion warning (closed)

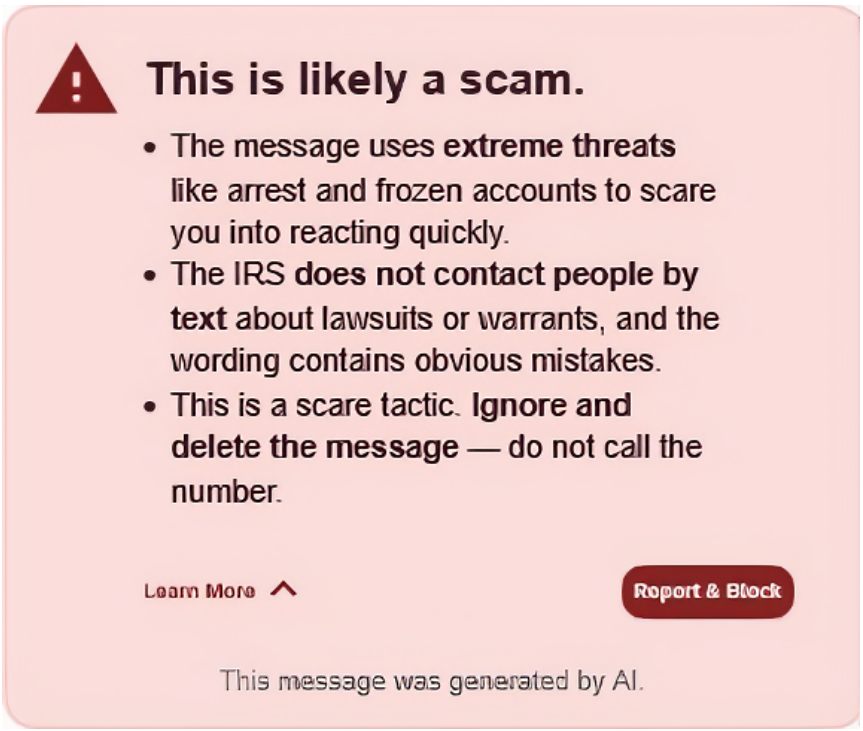


Figure G.7: Pre-study accordion warning (open)

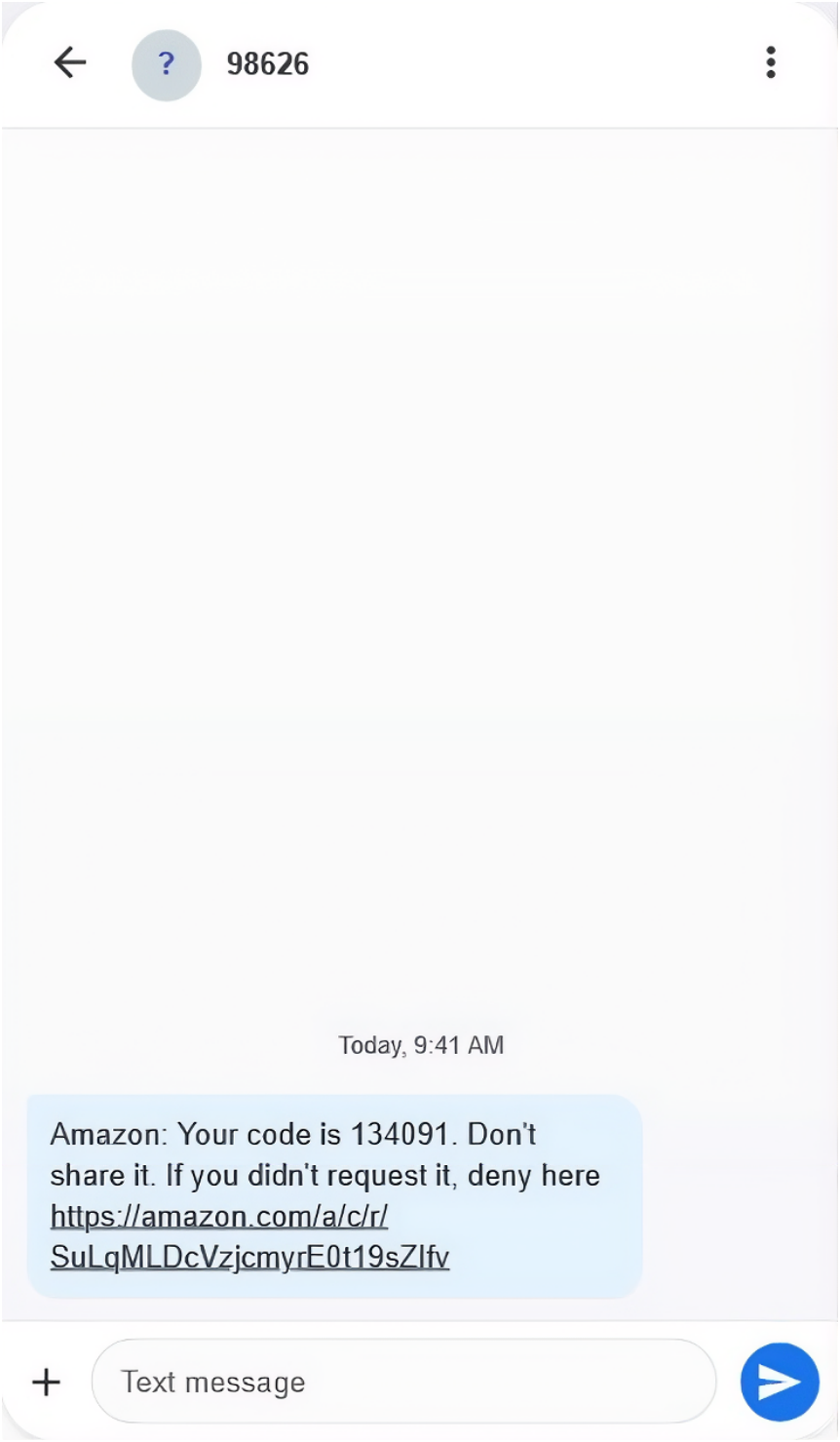


Figure G.8: Mockup texting interface shown without warning in the pre-studies.

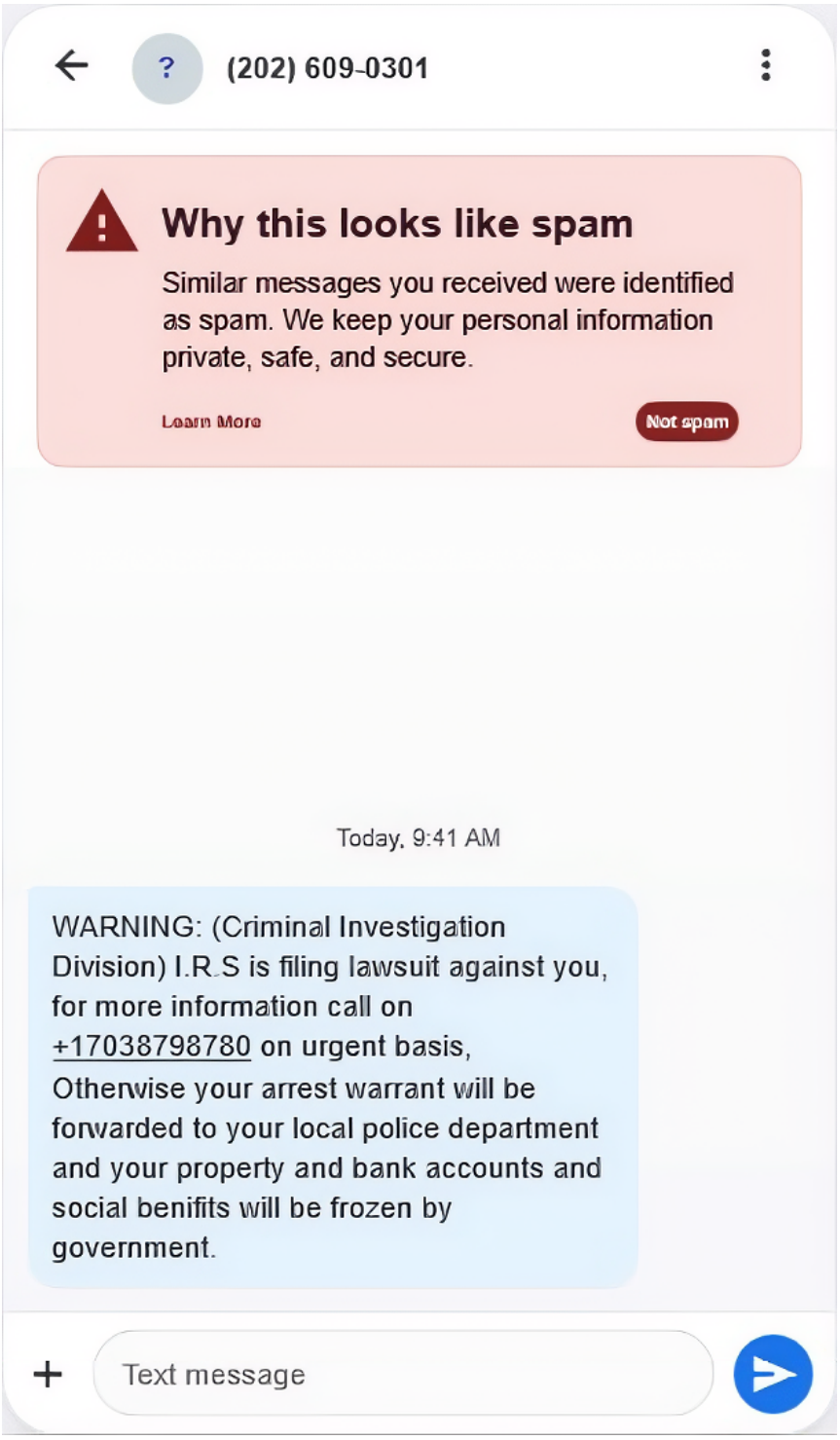


Figure G.9: Mockup texting interface shown with a spam warning in the pre-studies.

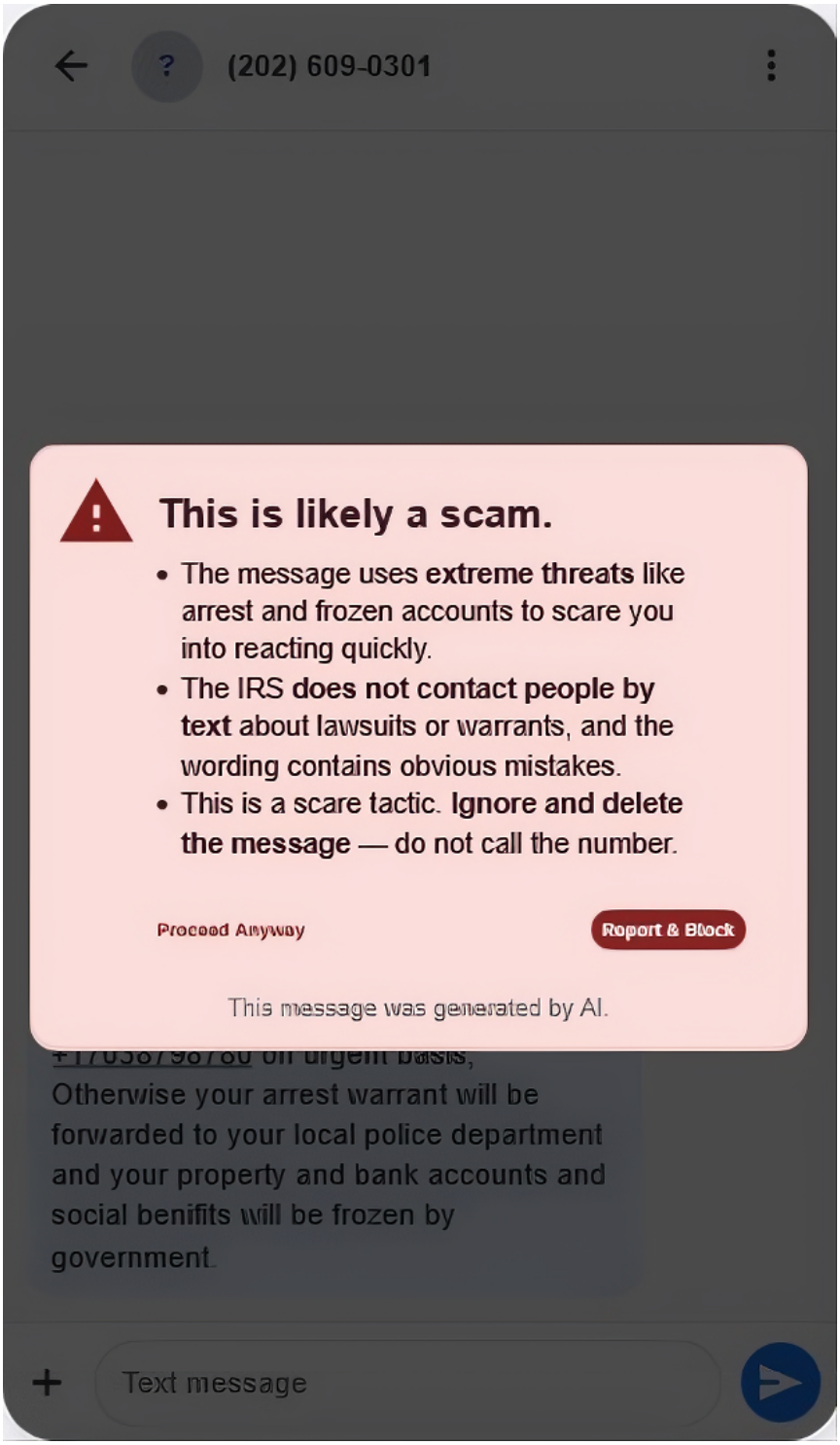


Figure G.10: Mockup texting interface shown with a blocking warning in the pre-studies.

Bibliography

[1] AARP Foundation. Off the hook: Reducing participation in telemarketing fraud, 2003. Retrieved from https://assets.aarp.org/rgcenter/consume/d17812_fraud.pdf on August 17, 2026. 2.2, 3.2.3, 10.1

[2] Norman Abrams. Uncovering the legislative histories of the early mail fraud statutes: The origin of federal auxiliary crimes jurisdiction. *Utah L. Rev.*, page 1079, 2021. 2

[3] Bhupendra Acharya, Muhammad Saad, Antonio Emanuele Cinà, Lea Schönherr, Hoang Dai Nguyen, Adam Oest, Phani Vadrevu, and Thorsten Holz. Conning the crypto conman: End-to-end analysis of cryptocurrency-based technical support scams. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 17–35, 2024. doi: 10.1109/SP54263.2024.00156. 2.1

[4] airdrop.io. airdrop.io, 2024. URL https://airdrops.io/. Retrieved from https://airdrops.io/ on August 17, 2026. 5.2.1

[5] Nurul Akbar. Analysing persuasion principles in phishing emails. Master's thesis, University of Twente, 2014. 2.1

[6] Omer Akgul, Richard Roberts, Moses Namara, Dave Levin, and Michelle L Mazurek. Investigating influencer VPN ads on YouTube. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 876–892, San Francisco, CA, May 2022. IEEE. URL https://doi.ieeecomputersociety.org/10.1109/SP46214.2022.9833633. 6.1, 6.1.2

[7] Devdatta Akhawe and Adrienne Porter Felt. Alice in warningland: a large-scale field study of browser security warning effectiveness. In *22nd USENIX security symposium (USENIX Security 13)*, pages 257–272, 2013. 2.4.1, 8.5, 9.1.2, 9.5.2

[8] Haldun Akoglu. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93, 2018. ISSN 2452-2473. doi: https://doi.org/10.1016/j.tjem.2018.08.001. URL https://www.sciencedirect.com/science/article/pii/S2452247318302164. 4.1.5

[9] Sara Albakry, Kami Vaniea, and Maria K. Wolters. What is this url's destination? empirical evaluation of users' url reading. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*, page 1–12, Honolulu, HI, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376168. 8.4.2

[10] Steve Alder. The biggest healthcare data breaches of 2024. *The HIPPA Journal*, January 2025. URL https://www.hipaajournal.com/biggest-healthcare-data-breaches-2024/. 2.3

[11] M. Alhaddad, M. Mohd, F. Qamar, and M. Imam. Study of student personality trait on spear-phishing susceptibility behavior. *International Journal of Advanced Computer Science and Applications*, 14(5):667–678, 2023. doi: 10.14569/IJACSA.2023.0140571. 3.2, 3.3.2, 3.3.5, 3.4.1

[12] Saleh Alqahtani and Priyadarsi Nanda. Effects of personal characteristics on phishing awareness, anti-phishing tool usage, and phishing avoidance behavior: A structural equation modeling approach. *2024 17th International Conference on Security of Information and Networks (SIN)*, pages 1–9, 2024. URL https://api.semanticscholar.org/CorpusID:276345239. 3.2, 3.3.5

[13] Kholoud Althobaiti, Nicole Meng, and Kami Vaniea. I don't need an expert! making url phishing features human comprehensible. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, pages 1–17. Association for Computing Machinery, 2021. 2.4.1

[14] Mohammed I Alwanain. How do children interact with phishing attacks? *International Journal of Computer Science & Network Security*, 21(3):127–133, 2021. URL http://dx.doi.org/10.22937/IJCSNS.2021.21.3.17. 2.3

[15] American Bankers Association. Banks never ask that, 2026. Retrieved from https://web.archive.org/web/20260626071907/https://www.banksneveraskthat.com/ on June 26th, 2026. 3.5

[16] Amnesty International. "I was someone else's property": Slavery, human trafficking and torture in cambodia's scamming compounds. Technical report, Amnesty International Ltd., 2025. 3.5

[17] Bonnie Brinton Anderson, Anthony Vance, C. Brock Kirwan, Jeffrey L. Jenkins, and David Eargle. From warning to wallpaper: Why the brain habituates to security warnings and what can be done about it. *Journal of Management Information Systems*, 33(3): 713–743, 2016. doi: 10.1080/07421222.2016.1243947. 8.3

[18] Keith B Anderson. To whom do victims of mass-market consumer fraud complain? *Available at SSRN 3852323*, 2021. 2.2, 10.1

[19] Monica Anderson, Michelle Faverio, and Jeffrey Gottfried. Teens, Social Media and Technology 2023. Technical report, Pew Research Center, 2023. URL https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/. https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/. 5.1.1, 5.1.4

[20] Joseph Aneke, Carmelo Ardito, and Giuseppe Desolda. Help the user recognize a phishing scam: Design of explanation messages in warning interfaces for phishing attacks. In *HCI for Cybersecurity, Privacy and Trust: Third International Conference (HCI-CPT*

'21), page 403–416, Berlin, DE, 2021. Springer-Verlag. ISBN 978-3-030-77391-5. doi: 10.1007/978-3-030-77392-2_26. 2.4.1

[21] Hibah Ansari. Lawmakers renew push to regulate kids' speech online despite speech protections. *American Civil Liberties Union*, 2025. Retrieved from https://www.aclu.org/news/privacy-technology/lawmakers-renew-push-to-regulate-kids-speech-online-despite-speech-protections on August 3rd, 2026. 10.2.1

[22] Apple. Report spam and block senders in messages on iphone, 2026. Retrieved from https://web.archive.org/web/20260302022128/https://support.apple.com/guide/iphone/report-spam-and-block-senders-iph3f94d910d/ios on March 2nd, 2026. 8.5

[23] Peter Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3): 375–386, 1955. URL https://doi.org/10.2307/3001775. 6.2.3

[24] Daniel Arp, Michael Spreitzenbarth, Malte Hubner, Hugo Gascon, Konrad Rieck, and CERT Siemens. Drebin: Effective and explainable detection of android malware in your pocket. In *Proceedings of the Network and Distributed System Security Symposium 2014 (NDSS '14)*, pages 23–26, San Diego, CA, USA, 2014. The Internet Society. 2.4.1

[25] AT&T. Fraud and scams, 2025. Retrieved from https://web.archive.org/web/20250122203502/https://about.att.com/pages/cyberaware/ae/fraud on January 22, 2025. 2.4.2

[26] Emma Avon. How to Buy Bitcoin Under 18? [Age Limit in 2023]. *CoinCodex*, July 2023. https://coincodex.com/article/30767/how-to-buy-bitcoin-under-18/, as of July 27th, 2026. 5

[27] Hapini Awang, Nur Suhaili Mansor, Mohamad Fadli Bin Zolkipli, Sarkin Tudu Shehu Malami, Khuzairi Mohd Zaini, and Ti Dun Yau. Cybersecurity awareness among special needs students: The role of parental control. *Mesopotamian Journal of CyberSecurity*, 2024. URL https://api.semanticscholar.org/CorpusID:270424883. 3.2

[28] Emad Badawi, Guy-Vincent Jourdan, Gregor Bochmann, Iosif-Viorel Onut, and Jason Flood. The "game hack" scam. In *Proceedings of the 19th International Conference on Web Engineering, (ICWE 2019)*, page 280–295, Daejeon, Republic of Korea, 2019. Springer. URL http://dx.doi.org/10.1007/978-3-030-19274-7_21. 2.3, 6, 6.1.1

[29] Jan Bailey, Louise Taylor, Paul Kingston, and Geoffrey Watts. Older adults and "scams": evidence from the mass observation archive. *Journal of Adult Protection*, 23(1):57–69, 2021. doi: 10.1108/JAP-07-2020-0030. 3.2.1, 3.2, 3.3.3

[30] Aurélien Baillon, Jeroen De Bruin, Aysil Emirmahmutoglu, Evelien Van De Veer, and Bram Van Dijk. Informing, simulating experience, or both: A field experiment on phishing risks. *PloS one*, 14(12):e0224216, 2019. URL https://doi.org/10.1371/journal.pone.0224216. 2.3

[31] Kimberly F Balsam and Jonathan J Mohr. Adaptation to sexual orientation stigma: a comparison of bisexual and lesbian/gay adults. *Journal of counseling psychology*, 54(3): 306, 2007. 4.1.2, 1, 4.2.1

[32] Luke Barr. Parents of teenager who died by suicide after sextortion scam urge 'tough' conversations with children, May 2023. URL https://abcnews.go.com/US/parents-teenager-died-by-suicide-after-sextortion-scam-urge/story?id=99047305. 4.2.2

[33] Timothy M Barry. # NotFinancialAdvice: Empowering the Federal Trade Commission to Regulate Cryptocurrency Social Media Influencers. *Ohio St. Bus. LJ*, 16:279, 2021. 5

[34] Lujo Bauer, Cristian Bravo-Lillo, Lorrie Cranor, and Eli Fragkaki. Warning design guidelines. Technical Report CMU-CyLab-13-002, Cylab, Carnegie Mellon University, Pittsburgh, PA, February 2013. URL https://kilthub.cmu.edu/articles/journal_contribution/Warning_Design_Guidelines_CMU-CyLab-13-002_/6468131?file=11896667. 8, 8.1.4, 8.5

[35] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020. 5.1.1, 7.1.1

[36] Jasika Bawa. How we're using ai to combat the latest scams, 2025. Retrieved from https://blog.google/innovation-and-ai/technology/safety-security/how-were-using-ai-to-combat-the-latest-scams/ on August 17, 2026. 3.5, 7.3, 8, 9.5.1

[37] Alycia Bayne, Elizabeth A. Mumford, Caroline Lancaster, and Jackie Sheridan-Johnson. Technology-facilitated abuse among americans age 50 and older: a latent class analysis. *Journal of Elder Abuse and Neglect*, 35(1):65–87, 2023. doi: 10.1080/08946566.2023.2197270. 3.2, 3.2.4, 3.4.1, 3.4.2

[38] Scott R. Beach, Sara J. Czaja, and Richard Schulz. Novel methods for assessment of vulnerability to financial exploitation (fe). *Journal of Elder Abuse and Neglect*, 35(4-5): 151–173, 2023. doi: 10.1080/08946566.2023.2281672. 3.2, 3.2.2, 3.3.2, 3.3.2, 3.3.2, 3.4.1

[39] Michaela Beals, Marguerite DeLiema, and Martha Deevy. Framework for a taxonomy of fraud. *Financial Fraud Research Center at the Stanford Center on Longevity and the FINRA Investor Education Foundation*, 2015. 2.1, 3.1.3, 3.2.4, 3.5, A.1.2

[40] Christopher Beckmann, Benjamin Berens, Niklas Kühl, Peter Mayer, Mattia Mossano, and Melanie Volkamer. Design and evaluation of an anti-phishing artifact based on useful transparency. In *International Workshop on Socio-Technical Aspects in Security*, pages 113–133. Springer, 2022. 3.4.2

[41] Rosanna Bellini, Emily Tseng, Noel Warford, Alaa Daffalla, Tara Matthews, Sunny Consolvo, Jill Woelfer, Patrick Kelley, Michelle Mazurek, Dana Cuomo, Nicola Dell, and Thomas Ristenpart. Sok: Safer digital-safety research involving at-risk users. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 635–654, 05 2024. doi: 10.1109/SP54263.2024.00071. 3.1.1, 3.5

[42] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: Series B (Methodological)*, 57(1):289–300, 1995. URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x. 6.2.3

[43] Manon Berney, Jan Ondrus, and Adrian Holzer. Navigating the shadows of cyber vigilantism: A preliminary analysis of social dynamics and activities of scambaiting. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703317. doi: 10.1145/3613905.3650913. URL https://doi.org/10.1145/3613905.3650913. 2.2

[44] Bet MGM Casino. The Legal Age for Casino Gambling Around the World, January 2023. Retrieved from https://casino.betmgm.com/en/blog/legal-age-casino-around-world/ on August 17, 2026. 5.2.4

[45] Jelke Bethlehem. Selection bias in web surveys. *International Statistical Review*, 78 (2):161–188, 2010. doi: https://doi.org/10.1111/j.1751-5823.2010.00112.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2010.00112.x. 4.1.8

[46] Sruti Bhagavatula, Lujo Bauer, and Apu Kapadia. "adulthood is trying each of the same six passwords that you use for everything"': The scarcity and ambiguity of security advice on social media. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), November 2022. doi: 10.1145/3555154. URL https://doi.org/10.1145/3555154. 7.2.1

[47] Arpita Bhattacharya, Calvin Liang, Emily Y. Zeng, Kanishk Shukla, Miguel E. R. Wong, Sean A. Munson, and Julie A. Kientz. Engaging Teenagers in Asynchronous Online Groups to Design for Stress Management. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, IDC '19, page 26–37, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366908. doi: 10.1145/3311927.3323140. URL https://doi.org/10.1145/3311927.3323140. 5.1.1

[48] Morvareed Bidgoli and Jens Grossklags. "hello. this is the irs calling.": A case study on scams, extortion, impersonation, and phone spoofing. In *2017 APWG Symposium on Electronic Crime Research (eCrime)*, pages 57–69, 2017. doi: 10.1109/ECRIME.2017.7945055. 2.1

[49] Behavioural Insights Team (BIT). Digital Engagement Practices in Retail Investing: Gamifcation & Other Behavioural Techniques. Technical report, Ontario Securities Commission Investor Ofce, November 2022. https://www.osc.ca/sites/default/files/2022-11/inv_research_20221117_gamification-of-retail-investing_EN.pdf as of August 17, 2026. 5.2.3

[50] Marzieh Bitaab, Haehyun Cho, Adam Oest, Zhuoer Lyu, Wei Wang, Jorij Abraham, Ruoyu Wang, Tiffany Bao, Yan Shoshitaishvili, and Adam Doupé. Beyond phish: Toward detecting fraudulent e-commerce websites at scale. In *2023 IEEE Symposium on*

Security and Privacy (SP), pages 2566–2583, San Francisco, CA, USA, 2023. doi: 10.1109/SP46215.2023.10179461. 7.2.2, 7.2.4, 7.3

[51] Mark Blythe, Helen Petrie, and John A. Clark. F for fake: four studies on how we fall for phish. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*, page 3469–3478, Vancouver, BC, Canada, 2011. Association for Computing Machinery. ISBN 9781450302289. doi: 10.1145/1978942.1979459. URL https://doi.org/10.1145/1978942.1979459. 6.4.1

[52] Mark Bohlender, Raphael Morisco, Mattia Mossano, Thorsten Schwarz, and Melanie Volkamer. Smile4vip: Intervention to support visually impaired users against phishing. In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 650–657, July 2024. doi: 10.1109/EuroSPW61312.2024.00079. 3.2, 3.4.2

[53] Mausumi Bose and Aloke Dey. *Handbook of Design and Analysis of Experiments*, chapter Chapter 4: Crossover Designs. Chapman & Hall (CRC Press), 2015. 9.3.1, 9.3.5, 9.3.8

[54] Elijah Bouma-Sims and Bradley Reaves. A first look at scams on youtube. In *the 2021 Workshop on Measurements, Attacks, and Defenses for the Web (MADWEB 2021)*, San Diego, CA, February 2021. ISBN 1-891562-67-3. URL https://dx.doi.org/10.14722/madweb.2021.23001. 2.3, 6, 6.1.1

[55] Elijah Bouma-Sims, Megan Li, Yanzi Lin, Adia Sakura-Lemessy, Alexandra Nisenoff, Ellie Young, Eleanor Birrell, Lorrie Faith Cranor, and Hana Habib. A us-uk usability evaluation of consent management platform cookie consent interface design on desktop and mobile. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, Hamburg, DE, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3580725. 8.5

[56] Elijah Bouma-Sims, Lily Klucinec, Mandy Lanyon, Lorrie Faith Cranor, and Julie Downs. Recruiting teenage participants for an online security experiment: A case study using peachjar. In *9th Workshop on Inclusive Privacy and Security (WIPS 2024)*, 2024. URL https://arxiv.org/abs/2408.00864. 6.2.2

[57] Elijah Bouma-Sims, Sanjnah Ananda Kumar, and Lorrie Faith Cranor. Exploring the privacy experiences of closeted users of online dating services in the us. *Proceedings on Privacy Enhancing Technologies*, 2024, 2024. 1, (document)

[58] Elijah Bouma-Sims, Lily Klucinec, Mandy Lanyon, Lorrie Faith Cranor, and Julie Downs. The kids are all right: Investigating the susceptibility of teens and adults to youtube giveaway scams. In *Proceedings of the Network and Distributed System Security (NDSS) Symposium 2025*, February 2025. 1, (document)

[59] Elijah Bouma-Sims, Mandy Lanyon, and Lorrie Faith Cranor. 'Is this a scam?': The nature and quality of reddit discussion about scams. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pages 2444–2458, 2025. 1, (document)

[60] Elijah Bouma-Sims, Enze Liu, Alexandra Xinran Li, and Lorrie Faith Cranor. From "be careful" to "here's why": Investigating user reasoning with context-specific sms scam

warnings. In *2026 IEEE Symposium on Security and Privacy (SP)*, pages 2589–2608, 2026. doi: 10.1109/SP63933.2026.00200. 1, (document)

[61] Elijah Robert Bouma-Sims, Hiba Hassan, Alexandra Nisenoff, Lorrie Faith Cranor, and Nicolas Christin. "It was honestly just gambling": Investigating the experiences of teenage cryptocurrency users on reddit. In *Proceedings of the Twentieth Symposium on Usable Privacy and Security*, 2024. URL https://www.usenix.org/system/files/soups2024-bouma-sims.pdf. 1, 3.2, 3.2.2, (document)

[62] Cristian Bravo-Lillo, Lorrie Faith Cranor, Julie Downs, and Saranga Komanduri. Bridging the gap in computer security warnings: A mental model approach. *IEEE Security & Privacy*, 9(2):18–26, 2011. doi: 10.1109/MSP.2010.198. URL https://doi.org/10.1109/MSP.2010.198. 6.2.1

[63] Cristian Bravo-Lillo, Saranga Komanduri, Lorrie Faith Cranor, Robert W. Reeder, Manya Sleeper, Julie Downs, and Stuart Schechter. Your attention please: designing security-decision uis to make genuine risks harder to ignore. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450323192. doi: 10.1145/2501604.2501610. URL https://doi.org/10.1145/2501604.2501610. 2.4.1

[64] Casey Breen, Cormac Herley, and Elissa M. Redmiles. A large-scale measurement of cybercrime against individuals. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517613. URL https://doi.org/10.1145/3491102.3517613. 3.5

[65] Susan W Brenner. Cybercrime jurisdiction. *Crime, law and social change*, 46(4):189–206, 2006. 10.2.1

[66] Curtis Briddick, Pamela Briggs, and James Nicholson. Using breach and attack demonstrations to explain spear phishing attacks to young adults. In *Information Security Education - Challenges in the Digital Age*, 2024. URL https://api.semanticscholar.org/CorpusID:270739115. 3.2, 3.4.1

[67] Antonio Briola, David Vidal-Tomás, Yuanrong Wang, and Tomaso Aste. Anatomy of a Stablecoin's failure: The Terra-Luna case. *Finance Research Letters*, 51:103358, 2023. ISSN 1544-6123. doi: https://doi.org/10.1016/j.frl.2022.103358. URL https://www.sciencedirect.com/science/article/pii/S1544612322005359. C.2

[68] R. Broadhurst, K. Skinner, N. Sifniotis, B. Matamoros-Macias, and Y. Ipsen. Phishing risks in a university student community. *Trends and Issues in Crime and Criminal Justice*, (587), 2020. 3.2, 3.2.2, 3.3.1, 3.3.2

[69] John Brooke. SUS: a 'quick and dirty' usability scale. *Usability evaluation in industry*, 189(3):189–194, 1996. 9.3.2

[70] Justin Brookman, Mat t Schwartz, and Grace Gedye. Consumer reports calls on the ftc and state attorneys general to take action against meta for its failure to mitigate harmful scam advertisements. *Consumer Reports Advocacy*, 2025. Retrieved from https://web.archive.org/web/20260803042744/https://advocacy.consumerr

eports.org/press_release/consumer-reports-calls-on-the-ftc-and-state-attorneys-general-to-take-action-against-meta-for-its-failure-to-mitigate-harmful-scam-advertisements/ on August 3rd, 2026. 10.2.1

[71] Samuel Hoy VII Brown. Gambling on the Blockchain: How the Unlawful Internet Gambling Enforcement Act Has Opened the Door for Offshore Crypto Casinos. *Vand. J. Ent. & Tech. L.*, 24:535, 2021. 5.2.4

[72] Stacey M. Brumbaugh-Johnson and Kathleen E. Hull. Coming out as transgender: Navigating the social implications of a transgender identity. *Journal of Homosexuality*, 66 (8):1148–1177, 2019. doi: 10.1080/00918369.2018.1493253. URL https://doi.org/10.1080/00918369.2018.1493253. PMID: 30052497. 2.3

[73] J.-W. Bullee, L. Montoya, M. Junger, and P. Hartel. Spear phishing in organisations explained. *Information and Computer Security*, 25(5):593–613, 2017. doi: 10.1108/ICS-03-2017-0009. 3.2, 3.2.2, 3.2.3, 3.3.1, 3.3.2, 3.3.3

[74] Bumble. Sec form 10-q, May 2023. URL https://ir.bumble.com/financials/sec-filings/default.aspx. 4.3

[75] Paolo Buono, Giuseppe Desolda, Francesco Greco, and Antonio Piccinno. Let warnings interrupt the interaction and explain: designing and evaluating phishing email warnings. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*, pages 1–6, Hamburg, DE, 2023. Association for Computing Machinery. 2.4.1

[76] Pavlo Burda, Luca Allodi, and Nicola Zannone. Cognition in social engineering empirical research: A systematic literature review. *ACM Trans. Comput.-Hum. Interact.*, 31(2), January 2024. ISSN 1073-0516. doi: 10.1145/3635149. URL https://doi.org/10.1145/3635149. 6

[77] United States Consumer Financial Protection Bureau. Cfpb sues jpmorgan chase, bank of america, and wells fargo for allowing fraud to fester on zelle, 2024. Retrieved from https://wayback.archive-it.org/23481/20250102080540/https://www.consumerfinance.gov/about-us/newsroom/cfpb-sues-jpmorgan-chase-bank-of-america-and-wells-fargo-for-allowing-fraud-to-fester-on-zelle/ on January 20th, 2025. 10.2.2

[78] Matt Burgess. A huge scam targeting kids with roblox and fortnite 'offers' has been hiding in plain sight. *WIRED*, 2023. Retrieved from https://web.archive.org/web/20241126073331/https://www.wired.com/story/poison-pdf-scam-fortnite-roblox/ on November 26th, 2024. 6, 6.4.2

[79] Thomas Burke and Ballard CFS Group. Ny attorney general sues early warning services for zelle network's lax security measures. *Consumer Finance Monitor*, 2025. Retrieved from https://web.archive.org/web/20260513132917/https://www.consumerfinancemonitor.com/2025/09/29/ny-attorney-general-sues-early-warning-services-for-zelle-networks-lax-security-measures/ on May 13th, 2026. 10.2.2

[80] Thomas Burke, Alan Kaplinsky, and Adam Maarec. New york court refuses to dismiss attorney general's sweeping fraud case against zelle operator early warning services. *Consumer Finance Monitor*, 2026. 10.2.2

[81] Mark Button and Cassandra Cross. *Cyber frauds, scams and their victims*. Routledge, 2017. 2.1

[82] Mark Button, David Shepherd, Chloe Hawkins, and Jacki Tapley. Disseminating fraud awareness and prevention advice to older adults: perspectives on the most effective means of delivery. *Crime Prevention and Community Safety*, 26:385 – 400, 2024. URL https://api.semanticscholar.org/CorpusID:273556717. 3.2, 3.4.1, 3.5

[83] Julia Camargo, Denise Cogo, and Amanda Alencar. Venezuelan refugees in brazil: Communication rights and digital inequalities during the covid-19 pandemic. *Media and Communication*, 10(2):230–240, 2022. doi: 10.17645/mac.v10i2.5051. 3.2, 3.2.2, 3.3.2

[84] Rebecca J. Campbell and Peter A. Lichtenberg. A short form of the financial exploitation vulnerability scale. *Clinical Gerontologist*, 44(5):594–603, 2021. doi: 10.1080/07317115.2020.1836108. 3.2, 3.2.4, 3.3.2, 3.3.2

[85] Canadian Centre for Child Protection. An analysis of financial sextortion victim posts published on r/sextortion, 2022. Retrieved from https://protectchildren.ca/en/resources-research/an-analysis-of-financial-sextortion-victim-posts-published-on-sextortion/ on August 17, 2026. 2.1, 2.2, 7, 7.1.3, 7.2.2, 7.2.3, 7.2.4, 7.3

[86] Deanna D. Caputo, Shari Lawrence Pfleeger, Jesse D. Freeman, and M. Eric Johnson. Going spear phishing: Exploring embedded training and awareness. *IEEE Security & Privacy*, 12(1):28–38, 2014. doi: 10.1109/MSP.2013.106. 2.4.2

[87] Nicole Carle and Jacques Ophoff. Using visual cues to enhance phishing education for children. In *Information Security Education - Challenges in the Digital Age*, 2024. URL https://api.semanticscholar.org/CorpusID:270738846. 3.2, 3.4.1

[88] Logan S. Casey, Sari L. Reisner, Mary G. Findling, Robert J. Blendon, John M. Benson, Justin M. Sayde, and Carolyn Miller. Discrimination in the united states: Experiences of lesbian, gay, bisexual, transgender, and queer americans. *Health Services Research*, 54 (S2):1454–1466, 2019. doi: https://doi.org/10.1111/1475-6773.13229. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-6773.13229. 2.3

[89] Cagatay Catal, Görkem Giray, Bedir Tekinerdogan, Sandeep Kumar, and Suyash Shukla. Applications of deep learning for phishing detection: a systematic literature review. *Knowledge and Information Systems*, 64(6):1457–1500, 2022. URL https://doi.org/10.1007/s10115-022-01672-x. 6.4.2

[90] Federico Maria Cau, Giuseppe Desolda, Francesco Greco, Lucio Davide Spano, and Luca Viganò. Can large language models improve phishing defense? a large-scale controlled experiment on warning dialogue explanations. *arXiv preprint arXiv:2507.07916*, 2025. 2.4.1, 8, 9

[91] E. A. Cech and T. J. Waidzunas. Systemic inequalities for lgbtq professionals in stem. *Science Advances*, 7(3):eabe0933, 2021. doi: 10.1126/sciadv.abe0933. URL https://www.science.org/doi/abs/10.1126/sciadv.abe0933. 2.3

[92] FINRA Investor Education Foundation CFA Institute. Gen Z and Investing: Social Media, Crypto, FOMO, and Family, May 2023. URL https://rpc.cfainstitute.org/en/research/reports/2023/gen-z-investing. Retrieved from https://rpc.cfainstitute.org/en/research/reports/2023/gen-z-investing on July 27th, 2026. 5

[93] Shelly Chaiken. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39(5): 752–766, 1980. 2.1

[94] Jake Chanenson, Tara Matthews, Sunny Consolvo, Patrick Gage Kelley, Jessica McClearn, Sarah Meiklejohn, Abhishek Roy, Renee Shelby, Kurt Thomas, and Amelia Hassoun. "it didn't feel right but i needed a job so desperately": Understanding people's emotions and help needs during scams. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3790556. URL https://doi.org/10.1145/3772318.3790556. 2.2, 3.5, 7, 10.3.3

[95] Cheng-Min Chao, Kai-Yun Kao, and Tai-Kuei Yu. Reactions to problematic internet use among adolescents: Inappropriate physical and mental health perspectives. *Frontiers in Psychology*, 11, 2020. ISSN 16641078. doi: 10.3389/fpsyg.2020.01782. 3.2

[96] Michelle Chapman. Cfpb drops lawsuit against bank of america, jpmorgan chase and wells fargo over zelle fraud. *The Associated Press*, 2025. 10.2.2

[97] Sayan Chaudhry and Chinmay Kulkarni. Design Patterns of Investing Apps and Their Effects on Investing Behaviors. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, DIS '21, page 777–788, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384766. doi: 10.1145/3461778.3462008. URL https://doi.org/10.1145/3461778.3462008. 5.3

[98] Hongliang Chen, Yunsha Pu, and David Atkin. Migration stress, risky internet uses, and scam victimization: An empirical study among chinese migrant workers. *Telemat. Inf.*, 83(C), September 2023. ISSN 0736-5853. doi: 10.1016/j.tele.2023.102022. 3.2, 3.3.1, 3.3.2, 3.3.2, 3.3.2, 3.3.5, 3.4.2

[99] Y. Chen, I. YeckehZaare, and A.F. Zhang. Real or bogus: Predicting susceptibility to phishing with economic experiments. *PLoS ONE*, 13(6), 2018. doi: 10.1371/journal.pone.0198213. 3.2, 3.3.2, 3.3.5, 3.4.1

[100] Zhihu Chen, Jing Wen, Chun Gou, Shuo Wang, and Jingjin Shao. Death anxiety and older adults' vulnerability to fraud: the mediating role of materialism. *Journal of Elder Abuse & Neglect*, 36:291 – 309, 2024. URL https://api.semanticscholar.org/CorpusID:269602204. 3.2

[101] Kimberly Ann Chew, Jennie Louise Ponsford, and Kate Rachel Gould. "this is a lifesaver": co-designing a novel cyberscam psychosocial recovery intervention framework for people with acquired brain injury. *Disability and Rehabilitation*, 47(19):5054–5062, 2025. 3.5

[102] Kimberly Ann Chew, Jennie Louise Ponsford, and Kate Rachel Gould. "a massive burden off my shoulders"': participant experiences of a novel cyberscam psychosocial recovery intervention for people with acquired brain injury. *Disability and Rehabilitation*, pages 1–16, 2025. 3.5

[103] Kimberly Ann Chew, Alyxandra Cazanis, Sue Hymus, Samantha Grant, Jennie Louise Ponsford, and Kate Rachel Gould. Evaluating the feasibility of a co-designed psychosocial recovery intervention for cyberscam survivors with acquired brain injury (abi): a single-case experimental design. *Disability and Rehabilitation*, 0(0):1–23, 2026. doi: 10.1080/09638288.2025.2605172. URL https://doi.org/10.1080/09638288.2025.2605172. PMID: 41477699. 3.5

[104] Robert Chew, Caroline Kery, Laura Baum, Thomas Bukowski, Annice Kim, Mario Navarro, et al. Predicting Age Groups of Reddit Users Based on Posting Behavior and Metadata: Classification Model Development and Validation. *JMIR Public Health and Surveillance*, 7(3):e25807, 2021. 5.1.1

[105] Andrew Childs. 'i guess that's the price of decentralisation...': Understanding scam victimisation experiences in an online cryptocurrency community. *International Review of Victimology*, page 02697580231215840, 2024. 2.2, 7

[106] Yan-Ming Chiou, Chien-Chung Shen, Chrystalla Mouza, and Teomara Rutherford. Augmented reality-based cybersecurity education on phishing. *Proceedings of the 4th IEEE International Conference on Artificial Intelligence and Virtual Reality, AIVR 2021*, pages 228–231, 2021. doi: 10.1109/AIVR52153.2021.00052. 3.2, 3.2.2, 3.4.1

[107] Clare Y. Cho. Kids online safety act. *Congressional Research Service*, (IF12730), 2026. 10.2.1

[108] Esther Yin-Nei Cho. A qualitative investigation of financial decision-making and enabling factors among ethnic minority young adults in hong kong. *Humanities and Social Sciences Communications*, 11, 2024. URL https://api.semanticscholar.org/CorpusID:272272444. 3.2

[109] Usman W Chohan. A history of Dogecoin. *Discussion Series: Notes on the 21st Century*, 2021. URL https://ssrn.com/abstract=3091219. 5.2.2

[110] Brad Sears Christy Mallory. Evidence of housing discrimination based on sexual orientation and gender identity: An analysis of complaints filed with state enforcement agencies, 2008-2014. Technical report, UCLA School of Law Williams Institute, feb 2016. URL https://escholarship.org/uc/item/50c6h3qf. 2.3

[111] Andrew Chu, Arjun Arunasalam, Muslum Ozgur Ozmen, and Z. Berkay Celik. Behind the tube: Exploitative monetization of content on YouTube. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2171–2188, Boston, MA, August 2022. USENIX Association. ISBN 978-1-939133-31-1. URL https://www.usenix.org/conference/usenixsecurity22/presentation/chu. 2.3, 6

[112] Edwin Ka Hung Chung and Dannii Yuen-lan Yeung. Reducing older people's risk of fraud victimization through an anti-scam board game. *Journal of Elder Abuse and Neglect*, 35 (2-3):121–138, 2023. doi: 10.1080/08946566.2023.2240005. 3.2, 3.4.1

[113] Robert B Cialdini. *Influence: The psychology of persuasion*, volume 55. Collins New York, 2007. 3.2.4

[114] Jason W. Clark and Damon McCoy. There are no free iPads: An analysis of survey scams as a business. In *6th USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 13)*, Washington, D.C., August 2013. USENIX Association. URL https://www.usenix.org/conference/leet13/workshop-program/presentation/clark. 2.3, 6, 6.1.1

[115] Camille Cobb and Tadayoshi Kohno. How public is my private life? privacy in online dating. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1231–1240, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052592. URL https://doi.org/10.1145/3038912.3052592. 4, 4.1.2, 4.2.2, 4.2.3

[116] William G Cochran. The comparison of percentages in matched samples. *Biometrika*, 37 (3/4):256–266, 1950. URL https://doi.org/10.2307/2332378. 6.2.3

[117] Lawrence E. Cohen and Marcus Felson. Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 44(4):588–608, 1979. ISSN 00031224. URL http://www.jstor.org/stable/2094589. 2.1

[118] Coinbase. User agreement, November 2023. URL https://www.coinbase.com/legal/user_agreement/united_states. Retrieved from https://www.coinbase.com/legal/user_agreement/united_states on January 16th, 2024. 2

[119] Treasury Committee. Regulating crypto. Technical report, UK Parliament, May 2022. URL https://publications.parliament.uk/pa/cm5803/cmselect/cmtreasy/615/report.html. 5.3

[120] Heith Copes, Kent R Kerley, Karen A Mason, and Judy Van Wyk. Reporting behavior of fraud victims and black's theory of law: An empirical assessment. *Justice Quarterly*, 18 (2):343–363, 2001. URL http://dx.doi.org/10.1080/07418820100094931. 2.3

[121] Juliet Corbin and Anselm Strauss. *Basics of Qualitative Research Techniques and Procedures for Developing Grounded Theory*. SAGE, Thousand Oaks, CA, USA, 4 edition, 2014. 4.1.1

[122] Giulio Cornelli, Sebastian Doerr, Jon Frost, and Leonardo Gambacorta. Crypto shocks and retail losses. In *BIS Bulletin*. Bank for International Settlement, February 2023. 5, 5.2.3

[123] Craig Newmark Philanthropies. Take9, 2026. Retrieved from https://web.archive.org/web/20260609031051/https://pausetake9.org/ on June 9th, 2026. 3.5

[124] Harald Cramér. *Mathematical methods of statistics*. Princeton Landmarks in Mathematics. Princeton University Press, 1999. ISBN 9780691005478. 6.2.3

[125] Criminal Law Magazine and Reporter. Bi-monthly digest of recent criminal cases, american and english. *Criminal Law Magazine and Reporter*, 10, July 1888. 2

[126] Robert Crosnoe and Clea McNeely. Peer relations, adolescent behavior, and public health research and practice. *Family & Community Health*, 31:S71–S80, 2008. 5

[127] Cassandra Cross. No laughing matter: Blaming the victim of online fraud. *International Review of Victimology*, 21(2):187–204, 2015. doi: 10.1177/0269758015571471. URL https://doi.org/10.1177/0269758015571471. 2.2, 7.1.3

[128] Cassandra Cross. Victims' motivations for reporting to the 'fraud justice network'. *Police Practice and Research*, 19(6):550–564, 2018. doi: 10.1080/15614263.2018.1507891. URL https://doi.org/10.1080/15614263.2018.1507891. 7.2.1, 7.3

[129] Cassandra Cross. 'oh we can't actually do anything about that': The problematic nature of jurisdiction for online fraud victims. *Criminology & Criminal Justice*, 20(3):358–375, 2020. 7.3

[130] Cassandra Cross and Thomas J Holt. More than money: examining the potential exposure of romance fraud victims to identity crime. *Global Crime*, 24(2):107–121, 2023. 2.1

[131] Cassandra Cross, Kelly Richards, and Russell G Smith. The reporting experiences and support needs of victims of online fraud. *Trends and issues in crime and criminal justice*, (518):1–14, 2016. 2.1, 2.2, 7.1.3, 7.3

[132] Anthony Cuthbertson. Student accidentally becomes a millionaire after turning selfies into NFT as a joke. *The Independent*, 2022. Retrieved from https://www.independent.co.uk/life-style/gadgets-and-tech/nft-cryptocurrency-selfie-crypto-b1996276.html on August 17, 2026. 5.2.3

[133] T. Daengsi, P. Pornpongtechavanich, P. Sirawongphatsara, T. Thimthong, and N. Sukniyom. A study of online scams associated with age, gender and loss of value in thailand. In *Proceedings of the 2022 International Conference on Data Analytics for Business and Industry*, pages 219–222, 2022. doi: 10.1109/ICDABI56818.2022.10041456. 3.2, 3.3.2

[134] Shanee Dawkins, Shanee Dawkins, and Jody Jacobs. NIST phish scale user guide. Technical report, US Department of Commerce, National Institute of Standards and Technology, 2023. 3.5

[135] DeleteMe. Stop spear phishing at the source, 2026. Retrieved from https://web.archive.org/web/20260423172157/https://joindeleteme.com/business/spear-phishing-prevention/ on April 23, 2026. 3.5

[136] M. Deliema, M. Deevy, A. Lusardi, and O.S. Mitchell. Financial fraud among older americans: Evidence and implications. *Journals of Gerontology - Series B Psychological*

Sciences and Social Sciences, 75(4):861–868, 2020. doi: 10.1093/geronb/gby151. 3.2.1, 3.2, 3.3.2, 3.3.2, 3.3.2

[137] M. Deliema, D. Shadel, and K. Pak. Profiling victims of investment fraud: Mindsets and risky behaviors. *Journal of Consumer Research*, 46(5):904–914, 2020. ISSN 00935301 (ISSN). doi: 10.1093/jcr/ucz020. 3.2, 3.2.3, 3.3.2, 3.3.2, 3.3.5

[138] M. DeLiema, Y. Li, and G. Mottola. Correlates of responding to and becoming victimized by fraud: Examining risk factors by scam type. *International Journal of Consumer Studies*, 47(3):1042–1059, 2023. ISSN 14706423 (ISSN). doi: 10.1111/ijcs.12886. 3.2, 3.2.4, 3.3.1, 3.3.2, 3.3.2, 3.3.2, 3.3.5, 7.1.2

[139] Marguerite DeLiema, Clifford A Robb, and Stephen Wendel. What does trust have to do with it? training consumers to detect digital imposter scams. *Journal of Financial Crime*, 32(1):77–97, 2025. 2.4.2

[140] Doygun Demirol, Gurkan Tuna, and Resul Das. A simple logging system for safe internet use. In *Proceedings of the 2017 International Artificial Intelligence and Data Processing Symposium*. IEEE, 2017. ISBN 978-1-5386-1880-6. doi: 10.1109/IDAP.2017.8090252. 3.2, 3.2.2, 3.4.2

[141] Natalia Denisenko, Valerio La Gatta, Marco Postiglione, Lirika Sola, Youdinghuan Chen, and V.S. Subrahmanian. Ai-generated phishing: Combining human behavior with post content to assess susceptibility. *ACM Trans. Internet Technol.*, 26(2), March 2026. ISSN 1533-5399. doi: 10.1145/3799891. URL https://doi.org/10.1145/3799891. 3.5

[142] Giuseppe Desolda, Lauren S. Ferro, Andrea Marrella, Tiziana Catarci, and Maria Francesca Costabile. Human factors in phishing attacks: A systematic literature review. *ACM Comput. Surv.*, 54(8), October 2021. ISSN 0360-0300. doi: 10.1145/3469886. 2.1, 3.3.2, 9.3.5

[143] Giuseppe Desolda, Francesco Greco, and Luca Viganò. APOLLO: A gpt-based tool to detect phishing emails and generate explanations that warn users. *Proceedings of the ACM on Human-Computer Interaction*, 9(4):1–33, 2025. 2.4.1, 8, 9

[144] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (MindTrek '11)*, page 9–15, Tampere, Finland, 2011. Association for Computing Machinery. ISBN 9781450308168. doi: 10.1145/2181037.2181040. URL https://doi.org/10.1145/2181037.2181040. 6.4.2

[145] Michael A. DeVito, Ashley Marie Walker, and Jeremy Birnholtz. 'too gay for facebook': Presenting lgbtq+ identity throughout the personal social media ecosystem. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), nov 2018. doi: 10.1145/3274313. URL https://doi.org/10.1145/3274313. 2.3

[146] Rachna Dhamija, J. D. Tygar, and Marti Hearst. Why phishing works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 581–590, New York, NY, USA, 2006. Association for Computing Machinery. ISBN

1595933727. doi: 10.1145/1124772.1124861. URL https://doi.org/10.1145/1124772.1124861. 2.4.1

[147] Carmen L. DiMario, Rio C. Bacha, and Brian K. Butka. Combatting senior scams using a large language model-created rubric: This paper explains a novel approach to reinforcement learning designed for the detection and prevention of scams aimed at senior citizens. In *Proceeding of the 2024 5th Asia Service Sciences and Software Engineering Conference*, ASSE '24, pages 130–136, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400717543. doi: 10.1145/3702138.3702140. URL https://doi.org/10.1145/3702138.3702140. 3.2, 3.4.2

[148] Verena Distler, Matthias Fassl, Hana Habib, Katharina Krombholz, Gabriele Lenzini, Carine Lallemand, Lorrie Faith Cranor, and Vincent Koenig. A systematic literature review of empirical methods and risk representation in usable privacy and security research. *ACM Trans. Comput.-Hum. Interact.*, 28(6), December 2021. ISSN 1073-0516. doi: 10.1145/3469845. 3.1.3, 3.2.2, 10.1, A.1.4

[149] W. Diteeyont and H.-Y. Ku. Internet literacy among the elderly in thailand. *Educational Media International*, 58(3):248–260, 2021. ISSN 09523987 (ISSN). doi: 10.1080/09523987.2021.1976829. 3.2, 3.2.2

[150] Dinil Mon Divakaran and Adam Oest. Phishing detection leveraging machine learning and deep learning: A review. *IEEE Security & Privacy*, 20(5):86–95, 2022. doi: 10.1109/MSEC.2022.3175225. 2.4

[151] M. Dixon, J. Nicholson, D. Branley-Bell, P. Briggs, and L. Coventry. Holding your hand on the danger button: Observing user phish detection strategies across mobile and desktop. *Proceedings of the ACM on Human-Computer Interaction*, 6(MHCI), 2022. ISSN 25730142 (ISSN). doi: 10.1145/3546730. 3.2, 3.2.3

[152] DocsBot. Gpt-5.4 vs gpt-4o, 2026. Retrieved from https://docsbot.ai/models/compare/gpt-5-4/gpt-4o on August 1st, 2026. 9.2.2

[153] M. Dolev-Cohen, I. Nezer, and A.A. Zumt. A qualitative examination of school counselors' experiences of sextortion cases of female students in israel. *Sexual Abuse*, 35(8):903–926, 2023. doi: 10.1177/10790632221145925. 3.2.1, 3.2, 3.2.4, 3.3.3, 3.4.2

[154] Lea Duesterwald, Ian Yang, and Norman Sadeh. Can a cybersecurity question answering assistant help change user behavior? an in situ study. In *Proceedings of the Symposium on Usable Security and Privacy (USEC) 2025*, San Diego, CA, USA, February 2025. The Internet Society. URL https://usableprivacy.org/static/files/duesterwald_usec_2025.pdf. 7.3

[155] S. Duke Hana, P.A. Boyle, B.D. James, L. Yu, and D.A. Bennett. Mild cognitive impairment and susceptibility to scams in old age. *Journal of Alzheimer's Disease*, 49(3):845–851, 2015. ISSN 13872877 (ISSN). doi: 10.3233/JAD-150442. 1, 2.3, 3.2, 3.2.2, 3.3.2, 3.3.2

[156] Marta Dynel and Andrew S. Ross. You don't fool me: On scams, scambaiting, deception, and epistemological ambiguity at r/scambait on reddit. *Social Media + Society*, 7(3):

20563051211035698, 2021. doi: 10.1177/20563051211035698. URL https://doi.org/10.1177/20563051211035698. 2.2, 7.1.1

[157] Valerie A. Earnshaw, Sari L. Reisner, Jaana Juvonen, Mark L. Hatzenbuehler, Jeff Perrotti, and Mark A. Schuster. LGBTQ Bullying: Translating Research to Action in Pediatrics. *Pediatrics*, 140(4):20 – 29, 10 2017. ISSN 0031-4005. doi: 10.1542/peds.2017-0432. URL https://doi.org/10.1542/peds.2017-0432. e20170432. 2.3

[158] eBay. Stay safe from scammers, 2025. Retrieved from https://web.archive.org/web/20250313012354/https://pages.ebay.com/securitycenter/stay_safe.html on March 13th, 2025. 2.4.2

[159] Natalie C Ebner, Donovan M Ellis, Tian Lin, Harold A Rocha, Huizi Yang, Sandeep Dommaraju, Adam Soliman, Damon L Woodard, Gary R Turner, R Nathan Spreng, et al. Uncovering susceptibility risk to online deception in aging. *The Journals of Gerontology: Series B*, 75(3):522–533, 2020. 3.2.1, 3.2, 3.2.2, 3.2.3, 3.3.2, 3.3.5

[160] Matthew Edwards, Claudia Peersman, and Awais Rashid. Scamming the scammers: Towards automatic detection of persuasion in advance fee frauds. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1291–1299, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349147. doi: 10.1145/3041021.3053889. URL https://doi.org/10.1145/3041021.3053889. 2.2

[161] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1065–1074, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580111. doi: 10.1145/1357054.1357219. URL https://doi.org/10.1145/1357054.1357219. 2.4.1

[162] I. Elueze and A. Quan-Haase. Privacy attitudes and concerns in the digital lives of older adults: Westin's privacy attitude typology revisited. *American Behavioral Scientist*, 62 (10):1372–1391, 2018. ISSN 00027642 (ISSN). doi: 10.1177/0002764218787026. 3.2

[163] emmeans package. Comparisons and contrasts in emmeans, 2026. Retrieved from https://web.archive.org/web/20260728145813/https://cran.r-project.org/web/packages/emmeans/vignettes/comparisons.html on July 28th, 2026. 9.3.5

[164] Audrey Engvalson. 18 dating app convos that are so bad they're funny, Jun 2019. URL https://www.buzzfeed.com/audreyworboys/tinder-convos-that-are-so-awkward-theyre-funny. 4.2.2

[165] Julie Ryan Evans. Tales From the Crypto Crypt: 38% of Investors Have Lost More Money Than Made It. Technical report, Lending Tree, January 2023. URL https://www.lendingtree.com/debt-consolidation/crypto-investors-survey/. 5, 5.2.3

[166] Facebook. Avoiding scams on facebook, 2025. Retrieved from https://web.archive.org/web/20250214180009/https://www.facebook.com/help/1674717642789671 on February 14th, 2025. 2.4.2

[167] Sascha Fahl, Marian Harbach, Yasemin Acar, and Matthew Smith. On the ecological validity of a password study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS '13)*, Newcastle, UK, 2013. Association for Computing Machinery. ISBN 9781450323192. doi: 10.1145/2501604.2501617. 8.4.2

[168] Muskan Fatima, Naheem Noah, and Sanchari Das. Exploring older adults' perceptions and experiences with online dating. *2024 Conference on Building a Secure & Empowered Cyberspace (BuildSEC)*, pages 57–65, 2024. URL https://api.semanticscholar.org/CorpusID:273655187. 3.2

[169] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4):1149–1160, 2009. doi: 10.3758/BRM.41.4.1149. 6.2.2, 9.3.3

[170] FBI El Paso. Fbi tech tuesday: Building a digital defense against "oops, wrong number!" texts, September 2023. Retrieved from https://web.archive.org/web/20241129093636/https://www.fbi.gov/contact-us/field-offices/elpaso/news/fbi-tech-tuesday-building-a-digital-defense-against-oops-wrong-number-texts on November 29th, 2024. 7.2.2

[171] Alessandro Fedele, Mirco Tonin, and Matteo Valerio. Phishing attacks: An analysis of the victims' characteristics based on administrative data. *Economics Letters*, 237:111663, 03 2024. doi: 10.1016/j.econlet.2024.111663. 3.2, 3.3.2

[172] Federal Deposit Insurance Company. Q: Will i be liable for unauthorized transactions made on business credit/debit cards?, 2026. Retrieved from https://ask.fdic.gov/fdicinformationandsupportcenter/s/article/Q-Will-I-be-liable-for-unauthorized-transactions-made-on-business-credit-debit-cards?language=en_US, on August 3rd, 2026. 10.2.2

[173] Brian A Feinstein and Roberto Rentería. Where is the line between being in versus out of the closet? *Archives of Sexual Behavior*, 52:1–5, 2023. 4.1.1, 4.1.2

[174] D. Feng, H. Rafih, and C. Munteanu. Understanding older adults' safety perceptions and risk mitigation strategies when accessing online services. In *Lect. Notes Comput. Sci.*, volume 14045 LNCS, pages 467–491. Springer, 2023. ISBN 03029743 (ISSN); 978-303135821-0 (ISBN). doi: 10.1007/978-3-031-35822-7_31. 3.2, 3.4.1

[175] L.-A. Fenge and S. Lee. Understanding the risks of financial scams as part of elder abuse prevention. *British Journal of Social Work*, 48(4):906–923, 2018. ISSN 00453102 (ISSN). doi: 10.1093/bjsw/bcy037. 3.2, 3.2.4, 3.3.2, 3.3.3, 3.4.1, 3.4.2

[176] Ana Ferreira, Lynne Coventry, and Gabriele Lenzini. Principles of persuasion in social engineering and their use in phishing. In *Human Aspects of Information Security, Privacy, and Trust: Third International Conference, HAS 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015. Proceedings 3*, pages 36–47. Springer, 2015. 2.1

[177] Casey Fiesler and Nicholas Proferes. "participant" perceptions of twitter research ethics. *Social Media+ Society*, 4(1):2056305118763366, 2018. 3.5, 5.1.3, 7.1.3, 10.3.3

[178] Casey Fiesler, Michael Zimmer, Nicholas Proferes, Sarah Gilbert, and Naiyan Jones. Remember the human: A systematic review of ethical considerations in reddit research. *Proc. ACM Hum.-Comput. Interact.*, 8(GROUP), February 2024. doi: 10.1145/3633070. URL https://doi.org/10.1145/3633070. 3.1.3, 3.5, 7.1.3, 7.3, 10.1

[179] Financial Conduct Authority. Fg21/1 guidance for firms on the fair treatment of vulnerable customers, 2021. Retrieved from http://www.fca.org.uk/publication/finalised-guidance/fg21-1.pdf on July 3rd, 2026. 10.2.2

[180] FINRA. Crypto assets, 2026. Retrieved from https://www.finra.org/rules-guidance/key-topics/crypto-assets on July 27th, 2026. 5

[181] Ronald A Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922. URL https://doi.org/10.1111/j.2397-2335.1922.tb00768.x. 6.2.3, 9.3.5

[182] Brock Floyd, Jared Jackson, Emma Probst, Hong Liu, Nischal Mishra, and Chen Zhong. Understanding Learners' Interests in Cybersecurity Competitions on Reddit. In *Proceedings of the 13th International Conference on Education Technology and Computers*, ICETC '21, page 444–449, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450385114. doi: 10.1145/3498765.3498835. URL https://doi.org/10.1145/3498765.3498835. 5.1.1

[183] Society for Worldwide Interbank Financial Telecommunication. What is KYC?, 2024. Retrieved from https://www.swift.com/your-needs/financial-crime-cyber-security/know-your-customer-kyc/meaning-kyc on August 17, 2026. 3

[184] Alain Forget, Saranga Komanduri, Alessandro Acquisti, Nicolas Christin, Lorrie Cranor, and Rahul Telang. Security behavior observatory: Infrastructure for long-term monitoring of client machines (cmu-cylab-14-009). Technical report, Carnegie Mellon University, 2014. 3.5

[185] Jesse Fox and Rachel Ralston. Queer identity online: Informal learning and teaching experiences of lgbtq individuals on social media. *Computers in Human Behavior*, 65: 635–642, 2016. 2.3

[186] Anjuli Franz, Verena Zimmermann, Gregor Albrecht, Katrin Hartwig, Christian Reuter, Alexander Benlian, and Joachim Vogt. SoK: Still plenty of phish in the sea — a taxonomy of User-Oriented phishing interventions and avenues for future research. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 339–358. USENIX Association, August 2021. ISBN 978-1-939133-25-0. 3.5

[187] Karen I. Fredriksen-Goldsen, Hyun-Jun Kim, Chengshi Shiu, Jayn Goldsen, and Charles A. Emlet. Successful Aging Among LGBT Older Adults: Physical and Mental Health-Related Quality of Life by Age Group. *The Gerontologist*, 55(1):154–168, 09 2014. ISSN 0016-9013. doi: 10.1093/geront/gnu081. URL https://doi.org/10.1093/geront/gnu081. 4.1.2

[188] Linton C Freeman. A further note on freeman's measure of association. *Psychometrika*, 41:273–275, 1976. URL https://doi.org/10.1007/BF02291845. 6.2.3

[189] Frontier Economics. Payment scam policies reduce fraud and improve consumer protection, 2026. Retrieved from https://www.frontier-economics.com/uk/en/news-and-insights/news/news-article-i22383-app-payment-scam-policies-reducing-fraud/ on August 3rd, 2026. 10.2.2

[190] B.E. Gavett, R. Zhao, S.E. John, C.A. Bussell, J.R. Roberts, and C. Yue. Phishing suspiciousness in older and younger adults: The role of executive functioning. *PLoS ONE*, 12 (2), 2017. 3.2, 3.3.2, 3.3.5

[191] Christine Geeng, Mike Harris, Elissa Redmiles, and Franziska Roesner. "like lesbians walking the perimeter": Experiences of U.S. LGBTQ+ folks with online security, safety, and privacy advice. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 305–322, Boston, MA, USA, August 2022. USENIX Association. ISBN 978-1-939133-31-1. URL https://www.usenix.org/conference/usenixsecurity22/presentation/geeng. 4

[192] Roland Gemayel and Alex Preda. Performance and learning in an ambiguous environment: A study of cryptocurrency traders. *International Review of Financial Analysis*, 77: 101847, 2021. ISSN 1057-5219. doi: https://doi.org/10.1016/j.irfa.2021.101847. URL https://www.sciencedirect.com/science/article/pii/S1057521921001794. 5, 5.2.3

[193] Diptiben Ghelani. What is Non-fungible token (NFT)? A short discussion about NFT Terms used in NFT. *Authorea Preprints*, 2022. 9

[194] Amy Gonzales, Dana Mastro, Laurent Wang, Jeannine Bell, Philip Anderson, and Jelani Ince. Dis"like": How race and education may influence the perceived costs of internet use in the u.s. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021. doi: 10.1145/3449131. 3.2, 3.3.2

[195] Google. Safe browsing, 2025. Retrieved from https://web.archive.org/web/20250428093410/https://safebrowsing.google.com/ on April 28th, 2025. 2.4.1

[196] Google. Accessibilityservice, 2026. Retrieved from https://web.archive.org/web/20260621051227/https://developer.android.com/reference/android/accessibilityservice/AccessibilityService on June 21st, 2026. 9.5.2

[197] Google. Learn about restricted settings, 2026. Retrieved from https://web.archive.org/web/20260720145710/https://support.google.com/android/answer/12623953?hl=en on July 20th, 2026. 9.3.2

[198] Jeffrey Gottfried, Monica Anderson, and Haley Nolan. Americans' social media use. Technical report, Pew Research Center, January 2024. URL https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/. 7.2.2

[199] K.R. Gould, J.-Y.J. Carminati, and J.L. Ponsford. "they just say how stupid i was for being conned". cyberscams and acquired brain injury: A qualitative exploration of the lived experience of survivors and close others. *Neuropsychological Rehabilitation*, 33(2): 325–345, 2023. doi: 10.1080/09602011.2021.2016447. 1, 2.3, 3.2.1, 3.2, 3.3.2, 3.3.2, 3.3.3, 3.3.3, 3.4.1

[200] Andy Greenberg, Lily Hay Newman, and Matt Burgess. Revealed: Leaked chats expose the daily life of a scam compound's enslaved workforce. *WIRED*, 2026. 3.5

[201] Frank L. Greitzer, Wanru Li, Kathryn B. Laskey, James Lee, and Justin Purl. Experimental investigation of technical and human factors related to phishing susceptibility. *Trans. Soc. Comput.*, 4(2), June 2021. doi: 10.1145/3461672. 2.3, 3.3.2

[202] M.D. Grilli, K.S. McVeigh, Z.M. Hakim, A.A. Wank, S.J. Getz, B.E. Levin, N.C. Ebner, and R.C. Wilson. Is this phishing? older age is associated with greater difficulty discriminating between safe and malicious emails. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, 76(9):1711–1715, 2021. ISSN 10795014 (ISSN). doi: 10.1093/geronb/gbaa228. 2.3, 3.2.1, 3.2, 3.2.4

[203] T.L.H. Grillo and A.F. Ward. How childhood adversity shapes susceptibility to covid-19 scams. *Journal of the Association for Consumer Research*, 7(1):54–62, 2022. ISSN 23781815 (ISSN). doi: 10.1086/711833. 3.2, 3.3.2

[204] Pamela Grimm. *Wiley international encyclopedia of marketing*, chapter Social desirability bias, page 1. John Wiley & Sons, Ltd, Chichester, UK, 2010. 4.1.8, 6.2.6

[205] Grindr. Albums, 2023. URL https://help.grindr.com/hc/en-us/articles/4414580688787-Albums. 2

[206] Grindr. Sec form 10-q, May 2023. URL https://investors.grindr.com/financials/sec-filings/default.aspx. 4.3

[207] The Match Group. Privacy, 2023. URL https://match.com/privacy. 4.3

[208] GSMA. RCS, 2026. Retrieved from https://web.archive.org/web/20260730201659/https://www.gsma.com/solutions-and-impact/technologies/networks/rcs/ on July 30th, 2026. 9.2.1

[209] I. Guedes, M. Martins, and C.S. Cardoso. Exploring the determinants of victimization and fear of online identity theft: an empirical study. *Security Journal*, 36(3):472–497, 2023. ISSN 09551662 (ISSN). doi: 10.1057/s41284-022-00350-5. 3.2, 3.2.2, 3.3.2, 3.3.5

[210] Andreas Hackethal, Tobin Hanspal, Dominique M Lammer, and Kevin Rink. The Characteristics and Portfolio Behavior of Bitcoin Investors: Evidence from Indirect Cryptocurrency Investments*. *Review of Finance*, 26(4):855–898, December 2021. ISSN 1572-3097. doi: 10.1093/rof/rfab034. URL https://doi.org/10.1093/rof/rfab034. 5.2.3

[211] Oliver L. Haimson and Tiffany C. Veinot. Coming out to doctors, coming out to "everyone": Understanding the average sequence of transgender identity disclosures using social media data. *Transgender Health*, 5(3):158–165, 2020. doi: 10.1089/trgh.2019.0045. URL https://doi.org/10.1089/trgh.2019.0045. 2.3

[212] S.D. Han, L.L. Barnes, S. Leurgans, L. Yu, C.C. Stewart, M. Lamar, C.M. Glover, D.A. Bennett, and P.A. Boyle. Susceptibility to scams in older black and white adults. *Frontiers in Psychology*, 12, 2021. ISSN 16641078 (ISSN). doi: 10.3389/fpsyg.2021.685258. 3.2, 3.2.2, 3.3.2, 3.3.2, 3.3.2

[213] MD Romael Haque and Sabirat Rubya. An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. *JMIR mHealth and uHealth*, 11(1):e44838, 2023. 7.3

[214] M. Harfath, R. Amrith, N. Dulanaka, P. Perera, L. Rupersinga, and C. Liyanapathirana. Intelligent cyber safe framework for children. In *IEEE Annu. Ubiquitous Comput., Electron. Mob. Commun. Conf., UEMCON*, pages 23–29. IEEE, 2021. ISBN 978-166540690-1. 3.2, 3.2.2, 3.4.2

[215] Ayako A. Hasegawa, Naomi Yamashita, Tatsuya Mori, Daisuke Inoue, and Mitsuaki Akiyama. Understanding Non-Experts' security- and Privacy-Related questions on a Q&A site. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 39–56, Boston, MA, August 2022. USENIX Association. ISBN 978-1-939133-30-4. URL https://www.usenix.org/conference/soups2022/presentation/hasegawa. 7.2.1, 7.2.3

[216] Benjamin Havers, Kartikeya Tripathi, Alexandra Burton, Wendy M Martin, and Claudia Cooper. Exploring the factors preventing older adults from reporting cybercrime and seeking help: A qualitative, semistructured interview study. *Health & Social Care in the Community*, 2024. URL https://api.semanticscholar.org/CorpusID:273585603. 3.2, 3.3.2, 3.3.3, 3.4.1, 3.4.2

[217] Benjamin Havers, Kartikeya Tripathi, Alexandra Burton, Sally McManus, and Claudia Cooper. Cybercrime victimisation among older adults: A probability sample survey in England and Wales. *PLOS ONE*, 19, 2024. URL https://api.semanticscholar.org/CorpusID:274854947. 3.2, 3.2.2, 3.3.2, 3.3.2, 3.3.5

[218] Jiahui He, Pete Snyder, Hamed Haddadi, Fabián E Bustamante, and Gareth Tyson. Measuring the accuracy and effectiveness of pii removal services. *arXiv preprint arXiv:2505.06989*, 2025. 10.3.2

[219] Amber Heemskerk, Tian Lin, Didem Pehlivanoglu, Ziad Hakim, Pedro A. Valdes-Hernandez, Leanne ten Brinke, Matthew D. Grilli, Robert C. Wilson, Gary R. Turner, R. Nathan Spreng, and Natalie C. Ebner. Interoceptive accuracy enhances deception detection in older adults. *The journals of gerontology. Series B, Psychological sciences and social sciences*, 2024. URL https://api.semanticscholar.org/CorpusID:272743060. 3.2, 3.2.4

[220] Michael V Heinz, Daniel M Mackin, Brianna M Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A Banta, Abi D Jewett, Abigail J Salzhauer, Tess Z Griffin, and Nicholas C Jacobson. Randomized trial of a generative ai chatbot for mental health treatment. *NEJM AI*, 2(4):AIoa2400802, 2025. 7.3

[221] M Todd Henderson and Max Raskin. A regulatory classification of digital assets: toward an operational howey test for cryptocurrencies, icos, and other digital assets. *Colum. Bus. L. Rev.*, page 443, 2019. 5.3

[222] Audrey Heng, Clare Heal, Jennifer Banks, and Robyn Preston. Transgender peoples' experiences and perspectives about general healthcare: A systematic review. *International Journal of Transgenderism*, 19(4):359–378, 2018. doi: 10.1080/15532739.2018.1502711. URL https://doi.org/10.1080/15532739.2018.1502711. 2.3

[223] Joe Hernandez and Jonathan Franklin. The SEC charges Lindsay Lohan, Jake Paul and others with illegally promoting crypto. *NPR*, March 2023. URL https://www.npr.org/2023/03/22/1165477713/lindsay-lohan-jake-paul-sec-crypto. 5

[224] Rebecca Hersher. Meet Mafiaboy, The 'Bratty Kid' Who Took Down The Internet. *NPR*, 2015. 5.2.4

[225] Sharon Heung, Claire Florence Weizenegger, Mo Houtti, Sunny Consolvo, Patrick Gage Kelley, Tara Matthews, Renee Shelby, Kurt Thomas, and Ashley Marie Walker. Who is at risk? examining the prevalence of digital-safety attacks and contextual risk factors in the united states. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3791463. URL https://doi.org/10.1145/3772318.3791463. 1, 2.3, 3.5

[226] Stan Higgins. Bitcoin Tipping Service ChangeTip to Shut Down. *CoinDesk*, September 2016. Retrieved from https://www.coindesk.com/markets/2016/11/18/bitcoin-tipping-service-changetip-to-shut-down/ on August 17, 2026. 8

[227] Grant Ho, Ariana Mirian, Elisa Luo, Khang Tong, Euyhyun Lee, Lin Liu, Christopher A Longhurst, Christian Dameff, Stefan Savage, and Geoffrey M Voelker. Understanding the efficacy of phishing training in practice. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 76–76. IEEE Computer Society, 2025. 2.4.2, 9.3.5

[228] Nguyen Phong Hoang, Yasuhito Asano, and Masatoshi Yoshikawa. Your neighbors are my spies: Location and other privacy concerns in glbt-focused location-based dating applications. In *2017 19th International Conference on Advanced Communication Technology (ICACT)*, pages 851–860, New York, NY, USA, 2017. IEEE. doi: 10.23919/ICACT.2017.7890236. 4

[229] K.W. Hong, C.M. Kelley, R. Tembe, E. Murphy-Hill, and C.B. Mayhorn. Keeping up with the joneses: Assessing phishing susceptibility in an email task. In *Proc Hum Factors Ergon Soc*, pages 1012–1016, 2013. ISBN 10711813 (ISSN); 978-094528943-2 (ISBN). doi: 10.1177/1541931213571226. 3.2, 3.3.5

[230] Jeff Horwitz. Meta is earning a fortune on a deluge of fraudulent ads, documents show. *Reuters*, 2025. Retrieved from https://www.reuters.com/investigations/meta-is-earning-fortune-deluge-fraudulent-ads-documents-show-2025-11-06/ on August 2nd, 2026. 10.2.1

[231] Mohamed Houtti, Abhishek Roy, Venkata Narsi Reddy Gangula, and Ashley Walker. A survey of scam exposure, victimization, types, vectors, and reporting in 12 countries. *Journal of Online Trust and Safety*, 2(4), Sep. 2024. doi: 10.54501/jots.v2i4.204. URL https://www.tsjournal.org/index.php/jots/article/view/204. 2.2, 10.1

[232] HSBC UK. What is an authorised push payment (app) scam?, 2026. Retrieved from https://web.archive.org/web/20260710051850/https://www.hsbc.co.uk/help/security-centre/what-is-an-authorised-push-payment-scam/, on July 10th, 2026. 10.2.2

[233] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. Measuring misinformation in video search platforms: An audit study on youtube. *Proc. ACM Hum.-Comput. Interact.*, 4 (CSCW1), may 2020. doi: 10.1145/3392854. URL https://doi.org/10.1145/3392854. 6.4.2

[234] Maciej Hyzy, Raymond Bond, Maurice Mulvenna, Lu Bai, Alan Dix, Simon Leigh, and Sophie Hunt. System usability scale benchmarking for digital health apps: meta-analysis. *JMIR mHealth and uHealth*, 10(8):e37290, 2022. 9.4.4

[235] International Organization for Standardization. Information security, cybersecurity and privacy protection – Information security controls. Standard, International Organization for Standardization, Geneva, CH, 2022. 2.4.2

[236] Iulia Ion, Rob Reeder, and Sunny Consolvo. "...No one can hack my Mind": Comparing expert and Non-Expert security practices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 327–346, Ottawa, July 2015. USENIX Association. ISBN 978-1-931971-249. URL https://www.usenix.org/conference/soups2015/proceedings/presentation/ion. 7.2.4

[237] C. Iuga, J.R.C. Nurse, and A. Erola. Baiting the hook: factors impacting susceptibility to phishing attacks. *Human-centric Computing and Information Sciences*, 6(1), 2016. ISSN 21921962 (ISSN). doi: 10.1186/s13673-016-0065-2. 3.2, 3.3.2, 3.3.2

[238] Mohieddin Jafari and Naser Ansari-Pour. Why, when and how to adjust your p values? *Cell Journal (Yakhteh)*, 20(4):604, 2019. URL https://dx.doi.org/10.22074/cellj.2019.5992. 6.2.3

[239] Pranjal Jain, Rama Adithya Varanasi, and Nicola Dell. "who is protecting us? no one!" vulnerabilities experienced by low-income indian merchants using digital payments. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '21, pages 261–274, New York, NY, 2021. Association for Computing Machinery. ISBN 978-1-4503-8453-7. doi: 10.1145/3460112.3471961. event-place: Virtual Event, Australia. 3.2, 3.2.2, 3.3.2, 3.3.4

[240] N. Jakobsson and M. Gerell. Evaluating the impact of an informational postcard campaign on telephone scams targeting the elderly. *Nordic Journal of Criminology*, 25(1), 2023. ISSN 2578983X (ISSN). doi: 10.18261/njc.25.1.2. 2.4.2, 3.2, 3.2.2, 3.4.1

[241] B.D. James, P.A. Boyle, and D.A. Bennett. Correlates of susceptibility to scams in older adults without dementia. *Journal of Elder Abuse and Neglect*, 26(2):107–122, 2014. ISSN

15404129 (ISSN). doi: 10.1080/08946566.2013.821809. 3.2, 3.2.2, 3.2.3, 3.3.2, 3.3.2, 3.3.2, 3.3.5, 3.4.1

[242] Daniel Jampen, Gürkan Gür, Thomas Sutter, and Bernhard Tellenbach. Don't click: towards an effective anti-phishing training. a comparative literature review. *Human-centric Computing and Information Sciences*, 10(1):1–41, 2020. 2.4.2

[243] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–35, 2019. 7.2.4, 7.3

[244] Benjamin Johnson, Steven Co, Tianze Sun, Carmen CW Lim, Daniel Stjepanović, Janni Leung, John B Saunders, and Gary CK Chan. Cryptocurrency trading and its associations with gambling and mental health: A scoping review. *Addictive Behaviors*, 136:107504, 2023. 5

[245] Benjamin Johnson, Daniel Stjepanović, Tianze Sun Janni Leung, and Gary C. K. Chan. Cryptocurrency trading, mental health and addiction: a qualitative analysis of reddit discussions. *Addiction Research & Theory*, 31, 2023. doi: 10.1080/16066359.2023.2174259. 5.1.1, 5.2.3

[246] Diana Novak Jones. Meta says us states are seeking $1.4 trillion in penalties in august youth safety trial. *Reuters*, 2026. Retrieved from `https://www.reuters.com/business/meta-says-us-states-are-seeking-14-trillion-penalties-august-youth-safety-trial-2026-07-07/` on August 3rd, 2026. 10.2.1

[247] R.A. Judges, S.N. Gallant, L. Yang, and K. Lee. The role of cognition, personality, and trust in fraud victimization in older adults. *Frontiers in Psychology*, 8(APR), 2017. ISSN 16641078 (ISSN). doi: 10.3389/fpsyg.2017.00588. 1, 2.3, 3.2, 3.3.2, 3.3.5

[248] Michael J. Kalsher, Michael S. Wogalter, and Bernadette M. Racicot. Pharmaceutical container labels: enhancing preference perceptions with alternative designs and pictorials. *International Journal of Industrial Ergonomics*, 18(1):83–90, 1996. ISSN 0169-8141. doi: 10.1016/0169-8141(95)00033-X. 8.5

[249] Alifiya Kapasi, Lei Yu, Christopher Stewart, Julie A. Schneider, David A. Bennett, and Patricia A. Boyle. Association of amyloid-β pathology with decision making and scam susceptibility. *Journal of Alzheimer's Disease*, 83(2):879–887, 2021. doi: 10.3233/JAD-210356. 3.2, 3.2.2, 3.3.2, 3.3.2

[250] Eleni Alexandra Katsarakes, Morgan Edwards, and Jeremiah D Still. Where do users look when deciding if a text message is safe or malicious? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 68, pages 221–225, Phoenix, AZ, USA, 2024. SAGE Publications. 2.1

[251] Claudia Kawai, Yang Zhang, Gáspár Lukács, Wenyi Chu, Chaoyi Zheng, Cijun Gao, Davood Gozli, Yonghui Wang, and Ulrich Ansorge. The good, the bad, and the red: implicit color-valence associations across cultures. *Psychological Research*, 87(3):704–724, 2023. 8.5

[252] Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. Privacy as part of the app decision-making process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, page 3393–3402, Paris, France, 2013. Association for Computing Machinery. ISBN 9781450318990. doi: 10.1145/2470654.2466466. URL https://doi.org/10.1145/2470654.2466466. 6.2.1

[253] Steven Kemp. Fraud reporting in catalonia in the internet era: Determinants and motives. *European Journal of Criminology*, 19(5):994–1015, 2022. URL https://doi.org/10.1177/1477370820941405. 2.2, 2.3

[254] Steven Kemp, Esther Sitges Maciá, and Nieves Erades-Pérez. Worry about online fraud and older adults: Avoidant and protective responses. *Journal of Criminology*, 58:446 – 465, 2024. URL https://api.semanticscholar.org/CorpusID:274039828. 3.2, 3.2.2, 3.2.2

[255] Kent R Kerley and Heith Copes. Personal fraud victims and their official responses to victimization. *Journal of Police and Criminal Psychology*, 17(1):19–35, 2002. URL https://doi.org/10.1007/BF02802859. 2.3

[256] Emaan Bilal Khan, Emaan Atique, and Mobin Javed. Investigating phishing threats in the email browsing experience of visually impaired individuals. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024. URL https://api.semanticscholar.org/CorpusID:269748068. 3.2, 3.3.2

[257] Amin Kharraz, William Robertson, and Engin Kirda. Surveylance: Automatically detecting online survey scams. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 70–86, San Francisco, CA, May 2018. URL https://dx.doi.org/10.1109/SP.2018.00044. 2.3, 6, 6.1.1

[258] Victoria Killion. The take it down act: A federal law prohibiting the nonconsensual publication of intimate images. *Congressional Research Service*, (LSB11314), 2025. 10.2.1

[259] Sara Kingsley, Michael Six Silberman, Clara Wang, Robert Lambeth, Jiayin Zhi, Motahhare Eslami, Beibei Li, and Jeffrey Bigham. 'your duties are to sweep a floor remotely': Low information quality in job advertisements is a barrier to low-income job-seekers' successful use of digital platforms. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, CHIWORK '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400710179. doi: 10.1145/3663384.3663403. URL https://doi.org/10.1145/3663384.3663403. 1, 3.2, 3.3.1, 3.3.2, 3.3.5

[260] G.H. Kirwan, C. Fullwood, and B. Rooney. Risk factors for social networking site scam victimization among malaysian students. *Cyberpsychology, Behavior, and Social Networking*, 21(2):123–128, 2018. ISSN 21522715 (ISSN). doi: 10.1089/cyber.2016.0714. 3.2, 3.3.5

[261] Lily Klucinec, Ellie Young, Elijah Bouma-Sims, and Lorrie Faith Cranor. "i wanted to buy robux but got scammed for 10 dollars in bitcoin"': Emerging adults' experiences with crypto assets as teens. In *Proceedings of the Symposium on Usable Security and Privacy (USEC) 2026*, San Diego, CA, USA, 2026. 5.3, 10.3.3

[262] Jennifer Klütsch, Jasmin Schwab, Christian Böffel, Verena Zimmermann, and Sabine J. Schlittmeier. Friend or phisher: how known senders and fear of missing out affect young adults' phishing susceptibility on social media. *Humanities and Social Sciences Communications*, 11, 2024. URL https://api.semanticscholar.org/CorpusID:272790456. 3.2, 3.3.5

[263] Takashi Koide, Naoki Fukushi, Hiroki Nakano, and Daiki Chiba. Chatspamdetector: Leveraging large language models for effective phishing email detection. *arXiv preprint arXiv:2402.18093*, 2024. 7.3

[264] Luka Koning, Marianne Junger, and Bernard Veldkamp. Risk factors for fraud victimization: The role of socio-demographics, personality, mental, general, and cognitive health, activities, and fraud knowledge. *International Review of Victimology*, 30(3):443–479, 2024. doi: 10.1177/02697580231215839. URL https://doi.org/10.1177/02697580231215839. 3.1.1, 3.2, 3.2.3, 3.3.1, 3.3.2, 3.3.2, 3.3.2, 3.3.3, 3.3.5, 3.5

[265] K. Kopecky. Online blackmail of czech children focused on so-called "sextortion" (analysis of culprit and victim behaviors). *Telematics and Informatics*, 34(1):11–19, 2017. ISSN 07365853 (ISSN). doi: 10.1016/j.tele.2016.04.004. 3.2, 3.2.2, 3.2.4, 3.3.3

[266] Klaus Krippendorff. Computing krippendorff's alpha-reliability. Technical report, Annenberg School for Communication, 2011. 7.1.2

[267] Soren Kristiansen and Aksel Vassard Jensen. Victimization in online gaming-related trade scams: A study among young danes. *Nordic Journal of Criminology*, 24(2):1–17, 2023. ISSN 2578983X. doi: 10.18261/njc.24.2.6. 3.2, 3.3.5

[268] Katharina Krombholz, Aljosha Judmayer, Matthias Gusenbauer, and Edgar Weippl. The other side of the coin: User experiences with bitcoin security and privacy. In *Financial Cryptography and Data Security: 20th International Conference, FC 2016, Christ Church, Barbados, February 22–26, 2016, Revised Selected Papers 20*, pages 555–580. Springer, 2017. 5, 5.2.4, 5.3

[269] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion analysis of variance. *Journal of the American Statistical Association*, 47(260):583–621, 1952. 9.1.2

[270] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Protecting people from phishing: the design and evaluation of an embedded training email system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*, pages 905–914, San Jose, CA, 2007. Association for Computing Machinery. ISBN 9781595935939. doi: 10.1145/1240624.1240760. URL https://doi.org/10.1145/1240624.1240760. 2.4.2

[271] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. School of phish: A real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security (SOUPS 2009)*, Mountain View, California, USA, 2009. Association for Computing Machinery. ISBN 9781605587363. doi: 10.1145/1572532.1572536. URL https://doi.org/10.1145/1572532.1572536. 2.4.2

[272] Daniele Lain, Kari Kostiainen, and Srdjan Čapkun. Phishing in organizations: Findings from a large-scale and long-term study. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 842–859, 2022. doi: 10.1109/SP46214.2022.9833766. 2.4.2, 9.3.5

[273] Daniele Lain, Tarek Jost, Sinisa Matetic, Kari Kostiainen, and Srdjan Capkun. Content, nudges and incentives: A study on the effectiveness and perception of embedded phishing training. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pages 4182–4196, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706363. doi: 10.1145/3658644.3690348. URL https://doi.org/10.1145/3658644.3690348. 2.4.2, 9.3.5

[274] Daniele Lain, Yoshimichi Nakatsuka, Kari Kostiainen, Gene Tsudik, and Srdjan Capkun. URL inspection tasks: Helping users detect phishing links in emails. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 1435–1454, Seattle, WA, August 2025. USENIX Association. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/lain. 2.4.1, 9.1.3, 9.3.1, 10.3.1

[275] E. Lastdrager, I.C. Gallardo, P. Hartel, and M. Junger. How effective is anti-phishing training for children? In *Proceedings of the Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 229–239. USENIX Association, 2017. ISBN 978-193197139-3. 2.3, 3.2, 3.4.1, 6.3.2, 9.4.3

[276] Elaine Lau and Zachary Peterson. A research framework and initial study of browser security for the visually impaired. In *Proceedings of the 32nd USENIX Security Symposium*, pages 4679–4696, 2023. 1, 3.2, 3.3.2, 3.4.2

[277] Carolyn Lauckner, Natalia Truszczynski, Danielle Lambert, Varsha Kottamasu, Saher Meherally, Anne Marie Schipani-McLaughlin, Erica Taylor, and Nathan Hansen. "catfishing," cyberbullying, and coercion: An exploration of the risks associated with dating app use among rural sexual minority males. *Journal of Gay & Lesbian Mental Health*, 23 (3):289–306, 2019. 4.2.2

[278] Adena B Meyers Laura E Berk. *Infants, Children, and Adolescents*. Pearson, 8 edition, 2016. ISBN 9780133936735; 0133936732; 0134035674; 9780134035673; 0134035658; 9780134035659. 2.3, 3.3.5, 5

[279] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research Methods in Human-Computer Interaction*, chapter Chapter 11: Analyzing qualitative data, pages 340–372. Elsevier Science & Technology, 2017. 3.1.3, 4.1.4, 9.3.6

[280] Ari Lazarus. How to spot extortion scams on lgbtq+ dating apps, September 2021. Retrieved from https://web.archive.org/web/20241212210754/https://consumer.ftc.gov/consumer-alerts/2021/09/how-spot-extortion-scams-lgbtq-dating-apps on December 12th, 2024. 2.3, 4, 4.2.2

[281] Suleman Lazarus, Mina Chiang, and Mark Button. Assessing human trafficking and cybercrime intersections through survivor narratives. *Deviant Behavior*, 47(7):1252–1269, 2026. doi: 10.1080/01639625.2025.2470402. URL https://doi.org/10.1080/01639625.2025.2470402. 3.5

[282] Van Thanh Le. How to Buy Crypto Under 18: Your Ultimate Guide to Financial Freedom. *Coin360*, October 2023. https://coin360.com/news/how-to-buy-crypto-under-18, as of July 27th, 2026. 5

[283] Esperance Ngo Mbedeg Le-Nye, Charles Yaacoub, and Jalal Joseph Possik. Evaluating phishing awareness strategies: A comparative study of education-based approaches and game-based learning. In *EUSPN/ICTH*, 2024. URL https://doi.org/10.1016/j.procs.2024.11.166. 3.2, 3.3.2

[284] K. Lee, W. Tang, S. Jones, L. Xu, and Z. Cong. The money smart for older adults program: A qualitative study of the participants' financial well-being. *Journal of Gerontological Social Work*, 64(2):120–134, 2021. ISSN 01634372 (ISSN). doi: 10.1080/01634372.2020.1814477. 3.2, 3.3.2

[285] Y.Y. Lee, C.L. Gan, and T.W. Liew. Phishing victimization among malaysian young adults: cyber routine activities theory and attitude in information sharing online. *Journal of Adult Protection*, 24(3-4):179–194, 2022. ISSN 14668203 (ISSN). doi: 10.1108/JAP-06-2022-0011. 3.4.1

[286] Y.Y. Lee, C.L. Gan, and T.W. Liew. Susceptibility to instant messaging phishing attacks: does systematic information processing differ between genders? *Crime Prevention and Community Safety*, 25(2):179–203, 2023. ISSN 14603780 (ISSN). doi: 10.1057/s41300-023-00176-2. 3.2, 3.4.1

[287] Nicole Legate, Richard M. Ryan, and Netta Weinstein. Is coming out always a "good thing"'? exploring the relations of autonomy support, outness, and wellness for lesbian, gay, and bisexual individuals. *Social Psychological and Personality Science*, 3(2):145–152, 2012. doi: 10.1177/1948550611411929. URL https://doi.org/10.1177/1948550611411929. 4, 4.1.2, 1

[288] Ada Lerner, Helen Yuxun He, Anna Kawakami, Silvia Catherine Zeamer, and Roberto Hoyle. Privacy and activism in the transgender community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376339. URL https://doi.org/10.1145/3313831.3376339. 4.2.3

[289] Alexandra Xinran Li, Elijah Bouma-Sims, Lily Klucinec, Ray Liu, Ayesha Binte Mostofa, Arjun Arunasalam, Lorrie Faith Cranor, Pubali Datta, Lucy Simko, and Karen Sowon. Quantifying risk perception and scam response among international and domestic u.s. university students. In *Proceedings of the Twenty-Second Symposium on Usable Privacy and Security*, 2026. 3.5

[290] Dennis H Li, Shruta Rawat, Jayson Rhoton, Pallav Patankar, Maria L Ekstrand, BR Simon Rosser, and J Michael Wilkerson. Harassment and violence among men who have sex with men (msm) and hijras after reinstatement of india's "sodomy law"'. *Sexuality research and social policy*, 14(3):324–330, 2017. 2.3

[291] Hanlin Li, Brent Hecht, and Stevie Chancellor. All that's happening behind the scenes: Putting the spotlight on volunteer moderator labor in reddit. *Proceedings of the Inter-*

national AAAI Conference on Web and Social Media, 16(1):584–595, May 2022. doi: 10.1609/icwsm.v16i1.19317. URL https://ojs.aaai.org/index.php/ICWSM/article/view/19317. 7.3

[292] W. Li, J. Lee, J. Purl, F.L. Greitzer, B. Yousefi, and K.B. Laskey. Experimental investigation of demographic factors related to phishing susceptibility. In *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, volume 2020-January, pages 2240–2249. IEEE Computer Society, 2020. ISBN 15301605 (ISSN); 978-099813313-3 (ISBN). 3.2, 3.4.1

[293] P.A. Lichtenberg, L. Stickney, and D. Paulson. Is psychological vulnerability related to the experience of fraud in older adults? *Clinical Gerontologist*, 36(2):132–146, 2013. ISSN 15452301 (ISSN). doi: 10.1080/07317115.2012.749323. 3.2.1, 3.2, 3.3.2, 3.3.2, 3.3.2, 3.4.1

[294] Peter A Lichtenberg, Maggie Tocco, and LaToya Hall. Financial decision-making deficits in scam cases: how frequent are they? *Journal of Elder Abuse & Neglect*, 36:84 – 89, 2024. URL https://api.semanticscholar.org/CorpusID:267409634. 3.2, 3.2.2, 3.2.2

[295] Bryan Lim, Roman Huerta, Alejandro Sotelo, Anthonie Quintela, and Priyanka Kumar. Explicate: Enhancing phishing detection through explainable ai and llm-powered interpretability. *arXiv preprint arXiv:2503.20796*, 2025. 2.4.1, 8, 9

[296] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. Does domain highlighting help people identify phishing sites? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2075–2084, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302289. doi: 10.1145/1978942.1979244. URL https://doi.org/10.1145/1978942.1979244. 2.4.1

[297] T. Lin, D.E. Capecci, D.M. Ellis, H.A. Rocha, S. Dommaraju, D.S. Oliveira, and N.C. Ebner. Susceptibility to spear-phishing emails: Effects of internet user demographics and email content. *ACM Transactions on Computer-Human Interaction*, 26(5), 2019. doi: 10.1145/3336141. 2.3, 3.2, 3.2.3, 3.2.4, 3.3.2

[298] Enze Liu, Lu Sun, Alex Bellon, Grant Ho, Geoffrey M. Voelker, Stefan Savage, and Imani N. S. Munyaka. Understanding the viability of gmail's origin indicator for identifying the sender. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*, pages 77–95, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-36-6. URL https://www.usenix.org/conference/soups2023/presentation/liu. 2.4.1

[299] Z. Liu, L. Zhou, and D. Zhang. Effects of demographic factors on phishing victimization in the workplace. In *Proc. Pac. Asia Conf. Inf. Syst.: Inf. Syst. (IS) Future, PACIS*. Association for Information Systems, 2020. 3.2, 3.3.2, 3.4.1

[300] Kenji Logie and Sumita Das. Lessons learned from dread darknet communities: How and why are fraudsters targeting the elderly to be victims or accomplices? *Criminology & Public Policy*, 24(2):237–271, 2025. doi: https://doi.org/10.1111/1745-9133.12684. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1745-9133.12684. 3.1.1, 3.2.1, 3.2, 3.2.2, 3.3.1, 3.3.4, 3.5, 10.3.2

[301] Kenji Logie, Katheryne Pugliese, and Alexis Acevedo. An examination of harm reduction strategies in oxycodone and adderall buyer feedback on alphabay. *Criminology & Public Policy*, 22(4):695–733, 2023. doi: https://doi.org/10.1111/1745-9133.12652. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1745-9133.12652. 3.1.1, 3.2, 3.2.2, 3.3.1

[302] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. "could you define that in bot terms"? requesting, creating and using bots on reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3488–3500, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3025830. URL https://doi.org/10.1145/3025453.3025830. 7.2.4

[303] Elizabeth Lopatto. Binance really loved telling people to use vpns, allegedly. *The Verge*, March 2023. URL https://www.theverge.com/2023/3/27/23659109/binance-cftc-vpn-signal-notes. 5.2.2

[304] Allison Lu, Bernardo BP Medeiros, Kevin RB Butler, and Patrick Traynor. Read this paper to get $50 million:* an analysis of mobile messaging scams using reddit data. *arXiv preprint arXiv:2605.16656*, 2026. 9.5.1

[305] Christoph Lutz and Giulia Ranzini. Where dating meets data: Investigating social and institutional privacy concerns on tinder. *Social Media + Society*, 3(1):2056305117697735, 2017. doi: 10.1177/2056305117697735. URL https://doi.org/10.1177/2056305117697735. 4.1.2, 4.3, 10.1

[306] Shaurya Malwa. Reddit Starts Airdrop of Polygon-Based 'Collectible Avatars'. *CoinDesk*, August 2022. URL https://www.coindesk.com/business/2022/08/25/reddit-starts-airdrop-of-polygon-based-collectible-avatars/. Retrieved from https://www.coindesk.com/business/2022/08/25/reddit-starts-airdrop-of-polygon-based-collectible-avatars/ on August 17, 2026. 5.2.1, 5.2.3

[307] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947. 9.3.5

[308] Jimmie Manning. Communicating sexual identities: A typology of coming out. *Sexuality & culture*, 19(1):122–138, 2015. 2.3

[309] Hannah Marder. 31 dating profiles so horribly cringe-worthy, i can't believe someone actually posted them, feb 2023. URL https://www.buzzfeed.com/hannahmarder/terrible-dating-profiles. 4.2.2

[310] Jessica Marshall. The challenges posed by scammers to online support groups: The 'deserving' and the 'undeserving' victims of scams. *New perspectives on cybercrime*, pages 213–240, 2017. 2.2, 7, 7.2.4

[311] Nina Marshall, Daniel Sturman, and Jaime C. Auton. Exploring the evidence for email phishing training: A scoping review. *Computers & Security*, 139:103695, 2024. ISSN

0167-4048. doi: https://doi.org/10.1016/j.cose.2023.103695. URL https://www.sciencedirect.com/science/article/pii/S0167404823006053. 2.4.2

[312] A.P. Martin, H. Keskes, W. Heni, P. Burton, A. Deliege, C. Rabhi, and R. Ayari. From entertainment to extortion: Tunisian children's online experiences. *Child Abuse and Neglect*, 145, 2023. doi: 10.1016/j.chiabu.2023.106405. 3.2, 3.3.3, 3.4.2

[313] G. Mba, G. Stringhini, J. Onaolapo, and L. Cavallaro. Flipping 419 cybercrime scams: Targeting the weak and the vulnerable. In *Int. World Wide Web Conf. , WWW Companion*, pages 1301–1310. International World Wide Web Conferences Steering Committee, 2017. ISBN 978-145034914-7. doi: 10.1145/3041021.3053892. 3.2.1, 3.2, 3.3.1, 3.3.2

[314] Colleen McClain and Risa Gelles-Watnick. From looking for love to swiping the field: Online dating in the u.s. Technical report, Pew Research Center, Washington, D.C., USA, Feb 2023. URL https://www.pewresearch.org/internet/2023/02/02/from-looking-for-love-to-swiping-the-field-online-dating-in-the-u-s/. 4.2.1

[315] Gordon McDonald. Multi-label agreement, August 2023. https://gdmcdonald.github.io/multi-label-inter-rater-agreement/Multi-Label_Agreement.html. 7.1.2

[316] Nora McDonald and Andrea Forte. The politics of privacy theories: Moving from norms to vulnerabilities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–14, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376167. URL https://doi.org/10.1145/3313831.3376167. 3.1.1

[317] Nora Mcdonald and Helena M. Mentis. Don't fall for this': Communications about cybersafety from the aarp. *Proceedings of the ACM on Human-Computer Interaction*, 7 (CSCW2), 2023. doi: 10.1145/3610039. 3.2, 3.4.1

[318] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019. doi: 10.1145/3359174. 8.1.5

[319] Peter C Meidlinger and Debra A Hope. Differentiating disclosure and concealment in measurement of outness for sexual minorities: The nebraska outness scale. *Psychology of Sexual Orientation and Gender Diversity*, 1(4):489, 2014. 4.1.2, 4.2.1

[320] Sarah Meiklejohn, Marjori Pomarole, Grant Jordan, Kirill Levchenko, Damon McCoy, Geoffrey M Voelker, and Stefan Savage. A fistful of Bitcoins: characterizing payments among men with no names. In *Proceedings of the ACM/USENIX Internet measurement conference*, pages 127–140, Barcelona, Spain, October 2013. 5

[321] Lucas Ramon Mendos and Dhia Rezki Rohaizad. Laws on us: A global overview of legal progress and backtracking on sexual orientation, gender identity, gender expression, and sex characteristics. Technical report, ILGA World, 2024. 2.3

[322] Jonathan Meng and Feng Fu. Understanding gambling behaviour and risk attitudes using cryptocurrency-based casino blockchain data. *Royal Society open science*, 7(10):201446, 2020. 5.2.4

[323] H.M. Mentis, G. Madjaroff, and A.K. Massey. Upside and downside risk in online security for older adults with mild cognitive impairment. In *Conf Hum Fact Comput Syst Proc*. Association for Computing Machinery, 2019. ISBN 978-145035970-2. doi: 10.1145/3290605.3300573. 3.2.1, 3.2, 3.3.2, 3.4.2

[324] Najmeh Miramirkhani, Oleksii Starov, and Nick Nikiforakis. Dial one for scam: Analyzing and detecting technical support scams. In *22nd Annual Network and Distributed System Security Symposium (NDSS*, 2016. 2.1

[325] Mehrnoosh Mirtaheri, Sami Abu-El-Haija, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. Identifying and analyzing cryptocurrency manipulations in social media. *IEEE Transactions on Computational Social Systems*, 8(3):607–617, 2021. 5

[326] David Modic and Ross Anderson. It's all over but the crying: The emotional and financial impact of internet fraud. *IEEE Security & Privacy*, 13(5):99–103, 2015. doi: 10.1109/MSP.2015.107. 2.1

[327] Jonathan Mohr and Ruth Fassinger. Measuring dimensions of lesbian and gay male experience. *Measurement and Evaluation in Counseling and Development*, 33(2):66–90, 2000. doi: 10.1080/07481756.2000.12068999. URL https://doi.org/10.1080/07481756.2000.12068999. 4.1.2, 1

[328] P.L. Morgan, E.J. Williams, N.A. Zook, and G. Christopher. Exploring older adult susceptibility to fraudulent computer pop-up interruptions. In *Adv. Intell. Sys. Comput.*, volume 782, pages 56–68. Springer Verlag, 2019. ISBN 21945357 (ISSN); 978-331994781-5 (ISBN). doi: 10.1007/978-3-319-94782-2_6. 3.2, 3.3.2, 3.4.1

[329] Mattia Mossano, Oksana Kulyk, Benjamin Maximillian Berens, Elena Marie Häußler, and Melanie Volkamer. Influence of url formatting on users' phishing url detection. In *Proceedings of the 2023 European Symposium on Usable Security*, EuroUSEC '23, page 318–333, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400708145. doi: 10.1145/3617072.3617111. URL https://doi.org/10.1145/3617072.3617111. 2.4.1

[330] E.A. Mueller, S.A. Wood, Y. Hanoch, Y. Huang, and C.L. Reed. Older and wiser: age differences in susceptibility to investment fraud: the protective role of emotional intelligence. *Journal of Elder Abuse and Neglect*, 32(2):152–172, 2020. ISSN 08946566 (ISSN). doi: 10.1080/08946566.2020.1736704. 3.2, 3.3.2, 3.3.5

[331] Kate Muir and Adam Joinson. An exploratory study into the negotiation of cyber-security within the family home, 2020. ISSN 16641078. 3.2, 3.2.2

[332] Joe Mullin. The kids online safety act will make the internet worse for everyone. *Electronic Frontier Foundation*, 2025. Retrieved from https://www.eff.org/deeplinks/2025/05/kids-online-safety-act-will-make-internet-worse-everyone on August 2nd, 2026. 10.2.1

[333] Joe Mullin. The take it down act: A flawed attempt to protect victims that will lead to censorship. *Electronic Frontier Foundation*, 2025. Retrieved from www.congress.gov/crs-product/LSB11314 on August 2nd 2026. 10.2.1

[334] Collins Munyendo, Yasemin Acar, and Adam Aviv. "in eighty percent of the cases, i select the password for them": Security and privacy challenges, advice, and opportunities at cybercafes in kenya. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 570–587, 05 2023. doi: 10.1109/SP46215.2023.10179410. 3.1.1, 3.2, 3.3.2, 3.3.4

[335] Mahjabin Nahar, Haeseung Seo, Eun-Ju Lee, Aiping Xiong, and Dongwon Lee. Fakes of varying shades: How warning affects human perception and engagement regarding llm hallucinations. *arXiv preprint arXiv:2404.03745*, 2024. 8.5

[336] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system, 2008. URL https://bitcoin.org/en/bitcoin-paper. 5.2.3, 5.2.4

[337] Hiroki Nakano, Daiki Chiba, Takashi Koide, Naoki Fukushi, Takeshi Yagi, Takeo Hariu, Katsunari Yoshioka, and Tsutomu Matsumoto. Understanding characteristics of phishing reports from experts and non-experts on twitter. *IEICE Transactions on Information and Systems*, E107.D(7):807 – 824, 2024. doi: 10.1587/transinf.2023EDP7221. 7.3

[338] D. Napoli, K. Baig, S. Maqsood, and S. Chiasson. "i'm literally just hoping this will work:" obstacles blocking the online security and privacy of users with visual disabilities. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security*, pages 263–280, 2021. 1, 2.3, 3.2, 3.3.2, 3.4.1

[339] National Cybersecurity Alliance. Oh behave! cybersecurity attitudes and behaviors report: 2021-2025. Technical report, National Cybersecurity Alliance, 2026. Retrieved from https://www.staysafeonline.org/articles/oh-behave-cybersecurity-attitudes-and-behaviors-report-2021%E2%80%932025 on August 17, 2026. 3.5

[340] National Cybersecurity Alliance. Then & now, 2026. Retrieved from https://web.archive.org/web/20260416150916/https://www.staysafeonline.org/thenandnow on April 16th, 2026. 3.5

[341] John C Navarro and George E Higgins. Familial identity theft. *American Journal of Criminal Justice*, 42(1):218–230, 2017. 5.3

[342] Philip WS Newall and Leonardo Weiss-Cohen. The gamblification of investing: How a new generation of investors is being born to lose. *International Journal of Environmental Research and Public Health*, 19(9):5391, 2022. 5

[343] P. Ni and W. Yu. A victim-based framework for telecom fraud analysis: A bayesian network model. *Computational Intelligence and Neuroscience*, 2022, 2022. doi: 10.1155/2022/7937355. 3.2

[344] J. Nicholson, Y. Javed, M. Dixon, L. Coventry, O.D. Ajayi, and P. Anderson. Investigating teenagers' ability to detect phishing messages. In *Proc. - IEEE Euro. Symp. Secur. Priv. Workshops, Euro S PW*, pages 140–149. IEEE, 2020. ISBN 978-172818597-2. doi: 10.1109/EuroSPW51379.2020.00027. 2.3, 3.2, 3.4.1

[345] Nina Grgurić Čop and Barbara Culiberg. Business Is Business: The Difference in Perception of Influencer's Morality Between Generation Y and Z. In Francisco J. Martínez-López and Steven D'Alessandro, editors, *Advances in Digital Marketing and eCommerce*, pages 56–61, Cham, 2020. Springer International Publishing. ISBN 978-3-030-47595-6. 5

[346] J.-R. Nino, G. Enström, and A.R. Davidson. Factors in fraudulent emails that deceive elderly people. In *Lect. Notes Comput. Sci.*, volume 10297 LNCS, pages 360–368. Springer Verlag, 2017. ISBN 03029743 (ISSN); 978-331958529-1 (ISBN). doi: 10.1007/978-3-319-58530-7_28. 3.2

[347] Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, Maurizio Tesconi, and Emilio Ferrara. Charting the Landscape of Online Cryptocurrency Manipulation. *IEEE Access*, 8:113230–113245, 2020. doi: 10.1109/ACCESS.2020.3003370. 5.2.4

[348] J. Nolte, Y. Hanoch, S. Wood, and D. Hengerer. Susceptibility to covid-19 scams: The roles of age, individual difference measures, and scam-related perceptions. *Frontiers in Psychology*, 12, 2021. ISSN 16641078 (ISSN). doi: 10.3389/fpsyg.2021.789883. 3.2, 3.3.2, 3.3.5

[349] Gareth Norris, Alexandra Brookes, and David Dowell. The psychology of internet fraud victimisation: A systematic review. *Journal of Police and Criminal Psychology*, 34:231–245, 2019. 2.1, 9.3.5

[350] Katie Notopoulos. Reddit's traffic is way, way up. like, banoodles up. *Business Insider*, June 2024. URL https://www.businessinsider.com/reddit-traffic-surges-from-google-search-2024-6. 7.2.4

[351] Rajvardhan Oak and Zubair Shafiq. Victims, vigilantes, and advice givers: An analysis of Scam-Related discourse on reddit. In *Proceedings of the Twenty-First Symposium on Usable Privacy and Security (SOUPS 2025)*, pages 57–71, 2025. 2.2, 3.5, 7, 10.3.3

[352] A.M. O'Connor, R.A. Judges, K. Lee, and A.D. Evans. Can adults discriminate between fraudulent and legitimate e-mails? examining the role of age and prior fraud experience. *Journal of Elder Abuse and Neglect*, 33(3):181–205, 2021. ISSN 08946566 (ISSN). doi: 10.1080/08946566.2021.1934767. 3.2

[353] Consumer Federation of America. Consumer federation of america sues meta for failing to protect users from scam advertisements, 2026. Retrieved from https://web.archive.org/web/20260803042339/https://consumerfed.org/news/press-releases/consumer-federation-of-america-sues-meta-for-failing-to-protect-users-from-scam-advertisements/ on August 3rd, 2026. 10.2.1

[354] OGAds. The takedown problem, 2023. Retrieved from https://web.archive.org/web/20231001161836/https://ogads.com/blog/youtube-takedowns on October 1st, 2023. 6.4.2

[355] Sebeom Oh, Samuel Rosen, and Anthony Lee Zhang. Digital Veblen Goods. *SSRN*, December 2023. https://ssrn.com/abstract=4042901. 5.2.3

[356] Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam Soliman, Tian Lin, and Natalie Ebner. Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing, 2017. 3.2, 3.2.3, 3.2.4, 3.3.2

[357] D.S. Oliveira, T. Lin, H. Rocha, D. Ellis, S. Dommaraju, H. Yang, D. Weir, S. Marin, and N.C. Ebner. Empirical analysis of weapons of influence, life domains, and demographic-targeting in modern spam: an age-comparative perspective. *Crime Science*, 8(1), 2019. ISSN 21937680 (ISSN). doi: 10.1186/s40163-019-0098-8. 3.2.1, 3.2, 3.2.2, 3.3.1, 3.3.2

[358] Roberta Liggett O'Malley and Karen M. Holt. Cyber sextortion: An exploratory analysis of different perpetrators engaging in a similar crime. *Journal of Interpersonal Violence*, 37(1-2):258–283, 2022. doi: 10.1177/0886260520909186. URL https://doi.org/10.1177/0886260520909186. PMID: 32146856. 4.2.2

[359] Roberta Liggett O'Malley and Katelyn Smith. Suicidal ideation among male victim-survivors of financial sextortion. *Victims & Offenders*, 20:1285 – 1306, 2024. URL https://api.semanticscholar.org/CorpusID:271396987. 3.2, 3.3.3

[360] Tully O'Neill. 'today i speak': Exploring how victim-survivors use reddit. *International Journal for Crime, Justice and Social Democracy*, 7(1):44–59, 2018. URL https://search.informit.org/doi/10.3316/informit.160821193555976. 7.3

[361] John E Pachankis and Skyler D Jackson. A developmental model of the sexual minority closet: Structural sensitization, psychological adaptations, and post-closet growth. *Archives of Sexual Behavior*, 52(5):1869–1895, 2023. 2.3, 4

[362] John E Pachankis, Conor P Mahon, Skyler D Jackson, Benjamin K Fetzner, and Richard Bränström. Sexual orientation concealment and mental health: A conceptual and meta-analytic review. *Psychological Bulletin*, 146(10):831, 2020. 4

[363] André Palheiros Da Silva, Winnie Mbaka, Johann Mayer, Jan-Willem Bullee, and Katja Tuma. Does trainer gender make a difference when delivering phishing training? a new experimental design to capture bias. In *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering*, EASE '24, pages 130–139, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400717017. doi: 10.1145/3661167.3661232. URL https://doi.org/10.1145/3661167.3661232. 3.2, 3.2.2, 3.4.1

[364] Rosemary Cosmas Tlatlaa Panga, Janeth Marwa, and Jema David Ndibwile. A game or notes? the use of a customized mobile game to improve teenagers' phishing knowledge, case of tanzania. *Journal of Cybersecurity and Privacy*, 2(3):466–489, 2022. doi: 10.3390/jcp2030024. 3.2, 3.4.1

[365] Kostantinos Papadamou, Jay Patel, Jeremy Blackburn, Philipp Jovanovic, and Emiliano De Cristofaro. From HODL to MOON: Understanding Community Evolution, Emotional Dynamics, and Price Interplay in the Cryptocurrency Ecosystem. *arXiv*, 2023. https://arxiv.org/abs/2312.08394. 5.1.1

[366] Masarah Paquet-Clouston, Matteo Romiti, Bernhard Haslhofer, and Thomas Charvat. Spams meet cryptocurrencies: Sextortion in the bitcoin ecosystem. In *Proceedings of the 1st ACM Conference on Advances in Financial Technologies*, AFT '19, pages 76–88, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367325. doi: 10.1145/3318041.3355466. URL https://doi.org/10.1145/3318041.3355466. 2.1, 2.2

[367] Charlie Parker, Sam Scott, and Alistair Geddes. Snowball sampling. In *SAGE research methods foundations*. Sage, Thousand Oaks, CA, USA, 2019. 4.1.1

[368] Kathryn Parsons, Agata McCormac, Malcolm Pattinson, Marcus Butavicius, and Cate Jerram. Phishing for the truth: A scenario-based experiment of users' behavioural response to emails. In *Proceedings of the 28th IFIP TC 11 International Information Security and Privacy Conference (SEC 2013)*, pages 366–378, Auckland, New Zealand, July 2013. Springer. URL https://doi.org/10.1007/978-3-642-39218-4_27. 2.3, 6.4.1

[369] Kathryn Parsons, Marcus Butavicius, Malcolm Pattinson, Dragana Calic, Agata Mccormac, and Cate Jerram. Do users focus on the correct cues to differentiate between phishing and genuine emails? In *Proceedings of the Australasian Conference on Information Systems (ACIS 2015)*, Adelaide, SA, Australia, 2016. URL https://aisel.aisnet.org/acis2015/6. 6.4.1

[370] Katalin Parti. What is a capable guardian to older fraud victims? comparison of younger and older victims' characteristics of online fraud utilizing routine activity theory. *Frontiers in Psychology*, 14, 2023. doi: 10.3389/fpsyg.2023.1118741. 3.2, 3.2.3, 3.3.2, 3.3.2, 3.3.5

[371] Katalin Parti and Faika Tahir. "if we don't listen to them, we make them lose more than money:" exploring reasons for underreporting and the needs of older scam victims. *Social Sciences*, 12(5):264, 2023. 7.2.3

[372] Payment Systems Regulatory. App scams, 2026. Retreived from web.archive.org/web/20260411105332/https://www.psr.org.uk/our-work/app-scams/ on April 11, 2026. 10.2.2

[373] B.K. Payne. Criminals work from home during pandemics too: a public health approach to respond to fraud and crimes against those 50 and above. *American Journal of Criminal Justice*, 45(4):563–577, 2020. ISSN 10662316 (ISSN). doi: 10.1007/s12103-020-09532-6. 3.2, 3.2.2

[374] PayPal. Learn about scams, 2025. Retrieved from https://web.archive.org/web/20250313010125/https://www.paypal.com/us/security/learn-about-scams March 13th, 2025. 2.4.2

[375] Juan Manuel Pedroza. Making noncitizens' rights real: Evidence from immigration scam complaints. *Law & Policy*, 44(1):44–69, 2022. doi: https://doi.org/10.1111/lapo.12180. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/lapo.12180. 3.1.1, 3.2, 3.2.2, 3.3.1, 3.3.3, 3.4.2, 8.4.1

[376] Didem Pehlivanoglu, Alayna Shoenfelt, Ziad M. Hakim, Amber Heemskerk, Jialong Zhen, Mario Mosqueda, Robert C. Wilson, Matt J Huentelman, Matthew D. Grilli, Gary R

Turner, R. Nathan Spreng, Natalie C. Ebner, and Paulo Boggio. Phishing vulnerability compounded by older age, apolipoprotein e e4 genotype, and lower cognition. *PNAS Nexus*, 3, 2024. URL https://api.semanticscholar.org/CorpusID:271663073. 1, 2.3, 3.2, 3.2.3, 3.2.4, 3.3.2

[377] Peng Peng, Chao Xu, Luke Quinn, Hang Hu, Bimal Viswanath, and Gang Wang. What happens after you leak your password: Understanding credential sharing on phishing sites. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, Asia CCS '19, page 181–192, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367523. doi: 10.1145/3321705.3329818. URL https://doi.org/10.1145/3321705.3329818. 2.1, 10.3.2

[378] Fahad Pervaiz, Rai Shah Nawaz, Muhammad Umer Ramzan, Maryem Zafar Usmani, Shrirang Mare, Kurtis Heimerl, Faisal Kamiran, Richard Anderson, and Lubna Razaq. An assessment of sms fraud in pakistan. In *Proceedings of the 2019 Conference on Computing and Sustainable Societies*, pages 195–205. Association for Computing Machinery, Inc, 2019. ISBN 978-1-4503-6714-1. doi: 10.1145/3314344.3332500. 3.2, 3.2.2, 3.3.1, 3.3.2

[379] Justin Petelka, Yixin Zou, and Florian Schaub. Put your warning where your link is: Improving and evaluating email phishing warnings. In *Proceedings of the 2019 CHI conference on human factors in computing systems (CHI '19)*, pages 1–15, Glasgow, UK, 2019. Association for Computing Machinery. 2.4.1, 9.4.4

[380] Richard E. Petty and John T. Cacioppo. The elaboration likelihood model of persuasion. volume 19 of *Advances in Experimental Social Psychology*, pages 123–205. Academic Press, 1986. doi: https://doi.org/10.1016/S0065-2601(08)60214-2. URL https://www.sciencedirect.com/science/article/pii/S0065260108602142. 2.1

[381] L.M. Podila, J.P. Bandreddi, J.I. Campos, Q. Niyaz, X. Yang, A. Trekles, C. Czerniak, and A.Y. Javaid. Practice-oriented smartphone security exercises for developing cybersecurity mindset in high school students. In *Proc. IEEE Int. Conf. Teach., Assess., Learn. Eng., TALE*, pages 303–310. IEEE, 2020. ISBN 978-172816942-2. doi: 10.1109/TALE48869.2020.9368440. 3.2, 3.4.1

[382] Sathvik Prasad, Elijah Bouma-Sims, Athishay Kiran Mylappan, and Bradley Reaves. Who's calling? characterizing robocalls through audio and metadata analysis. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 397–414. USENIX Association, August 2020. ISBN 978-1-939133-17-5. URL https://www.usenix.org/conference/usenixsecurity20/presentation/prasad. 7.2.2

[383] Suzanne Prior and Jacques Ophoff. Lessons learnt from using educational phishing materials with uk primary school children. In *WISE*, 2024. URL https://api.semanticscholar.org/CorpusID:270739041. 3.2, 3.4.1

[384] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society*, 7(2):20563051211019004, 2021. doi: 10.1177/2056

3051211019004. URL https://doi.org/10.1177/20563051211019004. 5.1.1

[385] Prolific. What are authenticity checks?, May 2026. Retrieved from https://researcher-help.prolific.com/en/articles/445144-what-are-authenticity-checks on July 28th, 2026. 9.1.1

[386] PwC UK. Psr sets the final rules on app scam reimbursements, 2024. Retrieved from https://web.archive.org/web/20250914062139/https://www.pwc.co.uk/industries/financial-services/understanding-regulatory-developments/psr-sets-the-final-rules-on-app-scam-reimbursements.html on September 14th, 2025. 10.2.2

[387] Md Lutfor Rahman, Daniel Timko, Hamid Wali, and Ajaya Neupane. Users really do respond to smishing. In *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy*, CODASPY '23, page 49–60, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700675. doi: 10.1145/3577923.3583640. URL https://doi.org/10.1145/3577923.3583640. 9.2.1, 9.4.2, 9.5.1, 9.5.2

[388] Ivana Rakovic and Yavuz Inal. Dark finance: exploring deceptive design in investment apps. In *IFIP Conference on Human-Computer Interaction*, pages 339–348. Springer, 2023. 5.2.3

[389] Fariza Rashid, Nishavi Ranaweera, Ben Doyle, and Suranga Seneviratne. LLMs are one-shot url classifiers and explainers. *Computer Networks*, 258:1–15, 2025. 2.4.1, 9

[390] D. Raval. Who is victimized by fraud? evidence from consumer protection cases. *Journal of Consumer Policy*, 44(1):43–72, 2021. doi: 10.1007/s10603-020-09466-w. 2.3, 3.2, 3.2.2, 3.3.2

[391] Devesh Raval. Which communities complain to policymakers? evidence from consumer sentinel. *Economic Inquiry*, 58(4):1628–1642, 2020. 2.2, 10.1

[392] Kate Raynes-Goldie. Aliases, creeping, and wall cleaning: Understanding privacy in the age of facebook. *First Monday*, 2010. 10.1

[393] Joseph Reagle. Disguising reddit sources and the efficacy of ethical research. *Ethics and Information Technology*, 24(3):41, 2022. 7.1.3, 10.1

[394] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How I learned to be secure: a census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*, pages 666–677, Vienna, Austria, 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978307. URL https://doi.org/10.1145/2976749.2978307. 1, 2.4.2

[395] Elissa M. Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L. Mazurek. A comprehensive quality evaluation of security and privacy advice on the web. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 89–108. USENIX Association, August 2020. ISBN 978-1-

939133-17-5. URL https://www.usenix.org/conference/usenixsecurity20/presentation/redmiles. 7.2.4, 9.5.3

[396] Rob Reeder, Ellen Cram Kowalczyk, and Adam Shostack. Poster: Helping engineers design neat security warnings. In *Proceedings of the Symposium On Usable Privacy and Security (SOUPS '11)*, Pittsburgh, PA, 2011. Association for Computing Machinery. 8

[397] Robert W. Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. An experience sampling study of user reactions to browser warnings in the field. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174086. URL https://doi.org/10.1145/3173574.3174086. 2.4.1, 9.1.2

[398] Francisco Regalado, Carlos Santos, and Ana Isabel Veloso. Game-based solutions for impacting digital competences in aged populations: Insights from an empathetic workshop. In *Interacción*, 2024. URL https://api.semanticscholar.org/CorpusID:270392521. 3.2

[399] Oliver D Reithmaier, Thorsten Thiel, Anne Vonderheide, and Markus Durmuth. Vision: Retiring scenarios—enabling ecologically valid measurement in phishing detection research with phishymailbox. In *Proceedings of the Symposium on Usable Security and Privacy (USEC) 2025*, pages 1–7, 2025. 10.3.1

[400] Marc Eric S Reyes, Nickaella B Bautista, Gemaima Reign A Betos, Kirby Ivan S Martin, Sophia Therese N Sapio, Ma Criselda T Pacquing, and John Manuel R Kliatchko. In/out of the closet: Perceived social support and outness among lgb youth. *Sexuality & Culture*, 27(1):290–309, 2023. 4.1.2

[401] Joshua Reynolds, Deepak Kumar, Zane Ma, Rohan Subramanian, Meishan Wu, Martin Shelton, Joshua Mason, Emily Stark, and Michael Bailey. Measuring identity confusion with uniform resource locators. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376298. URL https://doi.org/10.1145/3313831.3376298. 9.1.3, 9.3.1

[402] B.W. Reyns. Online routines and identity theft victimization: Further expanding routine activity theory beyond direct-contact offenses. *Journal of Research in Crime and Delinquency*, 50(2):216–238, 2013. ISSN 1552731X (ISSN). doi: 10.1177/0022427811425539. 3.2, 3.3.2, 3.3.2, 3.3.5

[403] Liliana Ribeiro, Inês Guedes, and Carla Cardoso. Which factors predict susceptibility to phishing? an empirical study. *Computers & Security*, 136, 11 2023. doi: 10.1016/j.cose.2023.103558. 3.2, 3.3.2, 3.3.2, 3.3.5

[404] Victoria Rideout, Alanna Peebles, Supreet Mann, and Michael B. Robb. The common sense census: Media use by tweens and teens. Technical report, Common Sense Media, 2021. URL https://www.commonsensemedia.org/research/the-common-sense-census-media-use-by-tweens-and-teens-2021. 6.1.1, 6.2.1

[405] Ellen DB Riggle, Sharon S Rostosky, Whitney W Black, and Danielle E Rosenkrantz. Outness, concealment, and authenticity: Associations with lgb individuals' psychological distress and well-being. *Psychology of Sexual Orientation and Gender Diversity*, 4(1):54, 2017. 4.1.2

[406] CA Robb and S Wendel. Who can you trust? assessing vulnerability to digital imposter scams. *Journal of Consumer Policy*, 46(1):27–51, 2023. URL https://doi.org/10.1007/s10603-022-09531-6. 2.4.2

[407] Jemima Robinson and Matthew Edwards. Fraudsters target the elderly: Behavioural evidence from randomised controlled scam-baiting experiments. *Security Journal*, pages 1–24, 2024. URL https://api.semanticscholar.org/CorpusID:267019986. 3.2.1, 3.2, 3.3.1

[408] Neal J Roese and Kathleen D Vohs. Hindsight bias. *Perspectives on psychological science*, 7(5):411–426, 2012. 9.3.8

[409] Jeremy Rose. *Communication in Practice*, chapter 6, pages 89–112. University of Minnesota Libraries Publishing, Twin Cities, MN, USA, 2025. 2.1

[410] Andrew S Ross and Lorenzo Logi. 'hello, this is martha': Interaction dynamics of live scambaiting on twitch. *Convergence*, 27(6):1789–1810, 2021. doi: 10.1177/13548565211015453. URL https://doi.org/10.1177/13548565211015453. 2.2

[411] Emma Roth. Online age verification is coming, and privacy is on the chopping block, May 2023. URL https://www.theverge.com/23721306/online-age-verification-privacy-laws-child-safety. 4.3

[412] Sayak Saha Roy, Cesar Torres, and Shirin Nilizadeh. 'Explain, Don't Just Warn!'–a real-time framework for generating phishing warnings with contextual cues. *arXiv preprint arXiv:2505.06836*, 2025. 2.4.1, 9

[413] Sayak Saha Roy, Elham Pourabbas Vafa, Kobra Khanmohamaddi, and Shirin Nilizadeh. DarkGram: A Large-Scale analysis of cybercriminal activity channels on telegram. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 4839–4858, Seattle, WA, August 2025. USENIX Association. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/roy. 10.3.2

[414] Andrew T. Rozema and James C. Davis. Anti-phishing training (still) does not work: A reproduction of phishing training inefficacy grounded in the nist phish scale. In *Proceedings of the ACM Web Conference 2026*, WWW '26, page 3147–3158, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400723070. doi: 10.1145/3774904.3792467. URL https://doi.org/10.1145/3774904.3792467. 3.5, 9.3.5

[415] Evan Ryan. Save the kids: The need for regulation of cryptocurrency to protect adolescents from fraud. *Family Court Review*, 2023. 5, 5.3

[416] Muhammad Salman, Muhammad Ikram, Nardine Basta, and Mohamed Ali Kaafar. Spallm-guard: Pairing sms spam detection using open-source and commercial llms. *arXiv preprint arXiv:2501.04985*, 2025. 7.3

[417] Premankit Sannd and David M. Cook. Older adults and the authenticity of emails: Grammar, syntax, and compositional indicators of social engineering in ransomware and phishing attacks. In *Proceedings of the 14th International Conference on Information Processing: Internet of Things, ICInPro 2018 - Proceedings*, 2018. doi: 10.1109/ICINPRO43533.2018.9096878. 3.2

[418] Shruti Sannon and Andrea Forte. Privacy research with marginalized groups: What we know, what's needed, and what's next. *Proceedings of the ACM on Human-Computer Interaction*, 6:1–33, 11 2022. doi: 10.1145/3555556. 3.1.1

[419] Dawn M. Sarno, Joanna E. Lewis, Corey J. Bohil, and Mark B. Neider. Which phish is on the hook? phishing vulnerability for older versus younger adults. *Human Factors*, 62(5): 704–717, 2020. doi: 10.1177/0018720819855570. 2.3, 3.2, 3.3.2, 3.3.2, 9.4.3

[420] D.M. Sarno, J.E. Lewis, C.J. Bohil, M.K. Shoss, and M.B. Neider. Who are phishers luring?: A demographic analysis of those susceptible to fake emails. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 2017-October, pages 1735–1739, 2017. doi: 10.1177/1541931213601915. 2.3, 3.2, 3.2.3, 3.3.2

[421] Elissa L Sarno, Jonathan J Mohr, Skyler D Jackson, and Ruth E Fassinger. When identities collide: Conflicts in allegiances among lgb people of color. *Cultural Diversity and Ethnic Minority Psychology*, 21(4):550, 2015. 4.1.2

[422] Heather Sarsons and Guo Xu. Confidence men? gender and confidence: Evidence among top economists. *Harvard University, Department of Economics, Littauer Center*, pages 1–26, 2015. URL `https://scholar.harvard.edu/files/sarsons/files/confidence_final.pdf`. 6.3.4

[423] Stuart E. Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The emperor's new security indicators. In *2007 IEEE Symposium on Security and Privacy (SP '07)*, pages 51–65, 2007. doi: 10.1109/SP.2007.35. 2.4.1

[424] Charlotte Schluger, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. Proactive moderation of online discussions: Existing practices and the potential for algorithmic support. *Proceedings of the ACM on Human-Computer Interaction*, 6 (CSCW2):1–27, 2022. 7.2.4

[425] Bruce Schneier. The third wave of network attacks. *ZDNet News*, 2000. Retrieved from `https://web.archive.org/web/20260803195447/https://www.zdnet.com/article/the-third-wave-of-network-attacks/` on August 3rd, 2026. 2.1

[426] S. Schoebel, R. Roepke, and U. Schroeder. Phishing academy: Evaluation of a digital educational game on urls and phishing. In *Lect. Notes Comput. Sci.*, volume 13134 LNCS, pages 44–53. Springer, 2021. ISBN 03029743 (ISSN); 978-303092181-1 (ISBN). doi: 10.1007/978-3-030-92182-8_5. 3.2, 3.4.1

[427] Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. An analysis of domestic abuse discourse on Reddit. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2577–2583, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1309. URL https://aclanthology.org/D15-1309/. 7.3

[428] Margrit Schreier. *The SAGE handbook of qualitative data collection*, chapter Sampling and generalization. SAGE Publications Ltd, 2018. 8.1.1

[429] Eve Kosofsky Sedgwick. *Epistemology of the Closet*. Univ of California Press, Berkeley, CA, USA, 2008. 4

[430] Ronald C Serlin, John Carr, and Leonard A Marascuilo. A measure of association for selected nonparametric procedures. *Psychological Bulletin*, 92(3):786, 1982. URL https://doi.org/10.1037/0033-2909.92.3.786. 6.2.3

[431] Yuxi Shang, Zhongxian Wu, Xiaoyu Du, Yanbin Jiang, Beibei Ma, and Meihong Chi. The psychology of the internet fraud victimization of older adults: A systematic review. *Frontiers in Psychology*, 13, 2022. URL https://doi.org/10.3389/fpsyg.2022.912242. 2.3

[432] Filipo Sharevski and Aziz Zeidieh. Assessing suspicious emails with banner warnings among blind and Low-Vision users in realistic settings. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2083–2100, Philadelphia, PA, August 2024. USENIX Association. ISBN 978-1-939133-44-1. URL https://www.usenix.org/conference/usenixsecurity24/presentation/sharevski. 3.1.1, 3.2, 3.3.2

[433] Filipo Sharevski, Aziz Zeidieh, Jennifer Vander Loop, and Peter Jachim. Blind and low-vision individuals' detection of audio deepfakes. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pages 4867–4881, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706363. doi: 10.1145/3658644.3690305. URL https://doi.org/10.1145/3658644.3690305. 3.2, 3.3.2

[434] Filipo Sharevski, Jennifer Vander Loop, Bill Evans, and Alexander Ponticello. (blind) users really do heed aural telephone scam warnings. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 74–92, 2025. doi: 10.1109/SP61157.2025.11097278. 3.5

[435] Ashley Sheil, Jacob Camilleri, Moya Cronin, Melanie Gruben, Michelle O'Keeffe, Hazel Murray, and Sanchari Das. Enhancing personalised cybersecurity guidance for older adults in ireland. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW Companion '24, pages 478–484, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711145. doi: 10.1145/3678884.3681894. URL https://doi.org/10.1145/3678884.3681894. 3.2, 3.4.1

[436] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now"': Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Com-*

munications Security, CCS '24, page 1671–1685, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706363. doi: 10.1145/3658644.3670388. URL https://doi.org/10.1145/3658644.3670388. 7.1.1

[437] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: The design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd Symposium on Usable Privacy and Security (SOUPS 2007)*, page 88–99, Pittsburgh, PA, 2007. Association for Computing Machinery. ISBN 9781595938015. doi: 10.1145/1280680.1280692. URL https://doi.org/10.1145/1280680.1280692. 6.4.2

[438] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*, pages 373–382, Atlanta, GA, 2010. Association for Computing Machinery. ISBN 9781605589299. doi: 10.1145/1753326.1753383. URL https://doi.org/10.1145/1753326.1753383. 2.3

[439] Imani Sherman, Jasmine Bowers, Keith McNamara, Juan Gilbert, Jaime Ruiz, and Patrick Traynor. Are you going to answer that? measuring user responses to anti-robocall application indicators. In *Proceedings of the 2020 Network and Distributed Systems Security Symposium (NDSS '20)*, San Diego, CA, USA, February 2020. The Internet Society. 8.5

[440] Hossein Siadati, Toan Nguyen, Payas Gupta, Markus Jakobsson, and Nasir Memon. Mind your smses: Mitigating social engineering in second factor authentication. *Computers & Security*, 65:14–28, 2017. 2.1

[441] Hossein Siadati, Sean Palka, Avi Siegel, and Damon McCoy. Measuring the effectiveness of embedded phishing exercises. In *10th USENIX Workshop on Cyber Security Experimentation and Test (CSET 17)*, Vancouver, BC, August 2017. USENIX Association. URL https://www.usenix.org/conference/cset17/workshop-program/presentation/siadatii. 2.4.2

[442] L. Simko, A. Lerner, S. Ibtasam, F. Roesner, and T. Kohno. Computer security and privacy for refugees in the united states. In *Proceedings of the 39th IEEE Symposium on Security and Privacy*, volume 2018-May, pages 409–423, 2018. doi: 10.1109/SP.2018.00023. 3.2.1, 3.2, 3.3.2, 3.3.3, 3.3.4

[443] Menna Singergy, John Ehab, Mohamed Rasslan, Caroline Sabty, and Alia El Bolock. Identity shield: Cultivating privacy awareness through ar for young adults. In *International Conference on Methodologies and Intelligent Systems for Technology Enhanced Learning*, 2024. URL https://api.semanticscholar.org/CorpusID:275516493. 3.2, 3.4.1

[444] J. Slupska, S. Cho, M. Begonia, R. Abu-Salma, N. Prakash, and M. Balakrishnan. "they look at vulnerability and use that to abuse you": Participatory threat modelling with migrant domestic workers. In *Proceedings of the 31st USENIX Security Symposium*, pages 323–340, 2022. 3.2.1, 3.2, 3.2.2, 3.3.3, 3.3.4, 3.4.2

[445] Adebayo Benedict Soares and Suleman Lazarus. Examining fifty cases of convicted online romance fraud offenders. *Criminal Justice Studies*, 37:328 – 351, 2024. 3.2.1, 3.2, 3.2.2, 3.3.1, 3.5

[446] Adebayo Benedict Soares, Suleman Lazarus, and Mark Button. Love, lies, and larceny: One hundred convicted case files of cybercriminals with eighty involving online romance fraud. *Deviant Behavior*, pages 1–24, 2025. 3.5

[447] G. Sonowal, K.S. Kuppusamy, and A. Kumar. Usability evaluation of active anti-phishing browser extensions for persons with visual impairments. In *Proceedings of the 4th International Conference on Advanced Computing and Communication Systems*, 2017. doi: 10.1109/ICACCS.2017.8014654. 3.2, 3.3.2

[448] Kyle Soska, Jin-Dong Dong, Alex Khodaverdian, Ariel Zetlin-Jones, Bryan Routledge, and Nicolas Christin. Towards understanding cryptocurrency derivatives: A case study of BitMEX. In *Proceedings of the 30th Web Conference (WWW'21)*, Ljubljana, Slovenia (online), April 2021. 5, 5.2.3

[449] Andreas Sotirakopoulos, Kirstie Hawkey, and Konstantin Beznosov. On the challenges in usable security lab studies: lessons learned from replicating a study on SSL warnings. In *Proceedings of the Seventh Symposium on Usable Privacy and Security (SOUPS '11)*, Pittsburgh, PA, USA, 2011. Association for Computing Machinery. ISBN 9781450309110. doi: 10.1145/2078827.2078831. 8.4.2

[450] Hannah Sparks and Samantha Ibrahim. Who is 'west elm caleb' and why do people on tiktok care about him?, jan 2022. URL https://nypost.com/2022/01/21/who-is-west-elm-caleb-and-why-do-people-care-about-him/. 4.2.2

[451] Eric Spero, Isa Seow, Lucas Betts, Eddie Fuatimau, Robert Biddle, Danielle Lottridge, and Giovanni Russello. Language as lure: A naturalistic study on pasifika phishing susceptibility. In *Twenty-First Symposium on Usable Privacy and Security (SOUPS 2025)*, pages 19–35, Seattle, WA, August 2025. USENIX Association. ISBN 978-1-939133-51-9. URL https://www.usenix.org/conference/soups2025/presentation/spero. 3.5

[452] Katta Spiel, Oliver L. Haimson, and Danielle Lottridge. How to do better with gender on surveys: A guide for hci researchers. *Interactions*, 26(4):62–65, jun 2019. ISSN 1072-5520. doi: 10.1145/3338283. URL https://doi.org/10.1145/3338283. 4.1.2

[453] Clay Spinuzzi. The methodology of participatory design. *Technical communication*, 52 (2):163–174, 2005. 3.5

[454] Bharat Srinivasan, Athanasios Kountouras, Najmeh Miramirkhani, Monjur Alam, Nick Nikiforakis, Manos Antonakakis, and Mustaque Ahamad. Exposing search and advertisement abuse tactics and infrastructure of technical support scammers. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 319–328, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356398. doi: 10.1145/3178876.3186098. URL https://doi.org/10.1145/3178876.3186098. 2.1

[455] Harold Stanislaw and Natasha Todorov. Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1):137–149, 1999. 3.5, 9.3.5

[456] Jemimah Steinfeld. Forced out of the closet: As people live out more of their lives online right now, our report highlights how lgbtq dating apps can put people's lives at risk. *Index on Censorship*, 49(2):101–104, 2020. doi: 10.1177/0306422020935360. URL https://doi.org/10.1177/0306422020935360. 4

[457] A. Steinmaurer, A. Bajramovic, D. Pollhammer, and C. Gutl. Learning security awareness in email communication using a platform for digital skill teaching. In *Proceedings of the 2022 IEEE International Conference on Teaching, Assessment and Learning for Engineering*, pages 38–45. IEEE, 2022. ISBN 978-166549117-4. doi: 10.1109/TALE54877.2022.00015. 3.2, 3.4.1

[458] Fred Steinmetz, Marc von Meduna, Lennart Ante, and Ingo Fiedler. Ownership, uses and perceptions of cryptocurrency: Results from a population survey. *Technological Forecasting and Social Change*, 173:121073, 2021. ISSN 0040-1625. doi: https://doi.org/10.1016/j.techfore.2021.121073. URL https://www.sciencedirect.com/science/article/pii/S0040162521005059. 5.3

[459] Hannah R Stevens, Irena Acic, and Sofia Rhea. Natural language processing insight into LGBTQ+ youth mental health during the COVID-19 pandemic: Longitudinal content analysis of anxiety-provoking topics and trends in emotion in LGBTeens microcommunity subreddit. *JMIR public health and surveillance*, 7(8):e29029, 2021. 5.1.1

[460] Elizabeth Stobert, David Barrera, Valérie Homier, and Daniel Kollek. Understanding cybersecurity practices in emergency departments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–8, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376881. URL https://doi.org/10.1145/3313831.3376881. 2.3

[461] Chris Stokel-Walker. How a Squid Game Crypto Scam Got Away With Millions. *WIRED*, 2021. Retrieved from https://www.wired.com/story/squid-game-coin-crypto-scam/ on August 17, 2026. C.2

[462] S.S. Sudha, J.P. Bandreddi, L.M. Podila, R. Govindula, A. Richardson, Q. Niyaz, X. Yang, and A.Y. Javaid. Impact of smartphone-based interactive learning modules on cybersecurity learning at the high-school level. In *IEEE Global Eng. Edu. Conf., EDUCON*, volume 2023-May. IEEE Computer Society, 2023. ISBN 21659559 (ISSN); 979-835039943-1 (ISBN). doi: 10.1109/EDUCON54358.2023.10125124. 3.2, 3.4.1

[463] F. Sudzina and A. Pavlicek. Propensity to click on suspicious links: Impact of gender, of age, and of personality traits. In *Proceedings of the Bled eConference*, pages 593–602. Association for Information Systems Electronic Library, AISeL, 2017. ISBN 978-961286043-1. doi: 10.18690/978-961-286-043-1.41. 3.2, 3.3.2, 3.3.5

[464] F. Sudzina, R. Novak, and A. Pavlicek. Impact of personality traits and gender on experiencing virtual offenses. In *IDIMT: Strateg. Model. Manag., Econ. Soc. - Interdiscip. Inf.*

Manag. Talks, pages 243–249. Trauner Verlag Universitat, 2018. ISBN 978-399062339-8. 3.2, 3.3.5

[465] J.C.-Y. Sun, S.-J. Yu, S.S.J. Lin, and S.-S. Tseng. The mediating effect of anti-phishing self-efficacy between college students' internet self-efficacy and anti-phishing behavior and gender difference. *Computers in Human Behavior*, 59:249–257, 2016. ISSN 07475632 (ISSN). doi: 10.1016/j.chb.2016.02.004. 3.2, 3.3.5, 3.4.1

[466] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. Crying wolf: An empirical study of ssl warning effectiveness. In *USENIX security symposium*, pages 399–416. Montreal, Canada, 2009. 2.4.1

[467] Thomas Sutter, Ahmet Selman Bozkir, Benjamin Gehring, and Peter Berlich. Avoiding the hook: Influential factors of phishing awareness training on click-rates and a data-driven approach to predict email difficulty perception. *IEEE Access*, 10:100540–100565, 2022. doi: 10.1109/ACCESS.2022.3207272. 2.4.2

[468] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34, 2000. 4.2.3

[469] Madiha Tabassum, Alana Mackey, and Ada Lerner. 'custodian of online communities': How moderator mutual support in communities help fight hate and harassment online. In *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*, pages 297–314, Philadelphia, PA, August 2024. USENIX Association. ISBN 978-1-939133-42-7. URL https://www.usenix.org/conference/soups2024/presentation/tabassum-madiha. 7.3

[470] Sarah Tabassum, Cori Faklaris, and Heather Richter Lipford. What drives SMiShing susceptibility? a US. interview study of how and why mobile phone users judge text messages to be real or fake. In *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*, pages 393–411, Philadelphia, PA, 2024. URL https://www.usenix.org/conference/soups2024/presentation/tabassum-sarah. 2.1, 8.1.1, 8.1.3, 8.4.2, 8.4.2, 8.4.3, 8.5, 10.3.1

[471] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How well do my results generalize now? the external validity of online privacy and security surveys. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 367–385, Boston, MA, August 2022. USENIX Association. ISBN 978-1-939133-30-4. URL https://www.usenix.org/conference/soups2022/presentation/tang. 6.2.6

[472] Xinru Tang, Gabriel Lima, Li Jiang, Lucy Simko, and Yixin Zou. Beyond "vulnerable populations": A unified understanding of vulnerability from a socio-ecological perspective. *Proc. ACM Hum.-Comput. Interact.*, 9(2), May 2025. doi: 10.1145/3710935. 2.3

[473] Xuyan Tang, Joyce S.T. Lam, and Bernie Garrett. Validation of the brief version of the susceptibility to persuasion-ii scale and evaluation of measurement invariance across age and gender in the canadian general population. *Heliyon*, 10, 2024. URL https://api.semanticscholar.org/CorpusID:268396861. 3.2, 3.2.2, 3.3.2, 3.3.5

[474] H. Tao. Loving strangers, avoiding risks: Online dating practices and scams among chinese lesbian (lala) women. *Media, Culture and Society*, 44(6):1199–1214, 2022. doi: 10.1177/01634437221088952. 3.1.1, 3.2, 3.2.2, 3.3.3

[475] Sam Taube. What Is a Custodial Account? UGMAs, UTMAs and More. *NerdWallet*, March 2024. Retrieved from https://www.nerdwallet.com/article/investing/what-is-a-custodial-account on August 17, 2026. 5.3

[476] Samuel Hardman Taylor, Jevan Alexander Hutson, and Tyler Richard Alicea. Social consequences of grindr use: Extending the internet-enhanced self-disclosure hypothesis. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 6645–6657, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3025775. URL https://doi.org/10.1145/3025453.3025775. 2.3

[477] F. Tazi, S. Shrestha, D. Norton, K. Walsh, and S. Das. Parents, educators, & caregivers cybersecurity & privacy concerns for remote learning during covid-19. In *Proceedings of the 1st International Conference of the ACM Greek SIGCHI Chapter*. Association for Computing Machinery, 2021. ISBN 978-145038578-7. doi: 10.1145/3489410.3489426. 3.2.1, 3.2

[478] Kurt Thomas, Frank Li, Ali Zand, Jacob Barrett, Juri Ranieri, Luca Invernizzi, Yarik Markov, Oxana Comanescu, Vijay Eranti, Angelika Moscicki, Daniel Margolis, Vern Paxson, and Elie Bursztein. Data breaches, phishing, or malware? understanding the risks of stolen credentials. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 1421–1434, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349468. doi: 10.1145/3133956.3134067. URL https://doi.org/10.1145/3133956.3134067. 10.3.2

[479] Sharen Thumboo and Sudeshna Mukherjee. Digital romance fraud targeting unmarried women. *Discover Global Society*, 2, 2024. URL https://api.semanticscholar.org/CorpusID:274665949. 3.2, 3.3.3

[480] Daniel Timko and Muhammad Lutfor Rahman. Smishing dataset I: Phishing SMS dataset from smishtank.com. In *Proceedings of the Fourteenth ACM Conference on Data and Application Security and Privacy (CODASPY '24)*, page 289–294, Porto, PRT, 2024. Association for Computing Machinery. ISBN 9798400704215. doi: 10.1145/3626232.3653282. 8.1.3

[481] Daniel Timko, Daniel Hernandez Castillo, and Muhammad Lutfor Rahman. Understanding influences on sms phishing detection: User behavior, demographics, and message attributes. In *Proceedings of the Symposium on Usable Security and Privacy (USEC) 2025*, San Diego, CA, USA, February 2025. The Internet Society. 2.1, 8.1.3, 8.4.2, 8.5, 10.3.1

[482] Tin Tin Ting, Kar Man Cheah, Jie Xin Khiew, Yung Chin Lee, Jun Kit Chaw, and Chong Keat Teoh. Validation of cyber security behaviour among adolescents at malaysia university: Revisiting gender as a role. *International Journal of Innovative Research and Scientific Studies*, 2024. URL https://api.semanticscholar.org/CorpusID:266996322. 3.2

[483] J. Toomey. The age of fraud. *Harvard Journal on Legislation*, 60(1):101–166, 2023. ISSN 0017808X (ISSN). 3.2, 3.3.2, 3.3.2, 3.3.2

[484] Lauri Tuovinen and Juha Röning. Baits and beatings: Vigilante justice in virtual communities. *Proceedings of CEPE*, pages 397–405, 2007. 2.2, 7.1.1

[485] u/ElijahBoumaSims. Research article: Is this a scam??: The nature and quality of reddit discussion about scams [/r/phishing], November 2025. Retrieved from https://www.reddit.com/r/phishing/comments/1p5l83e/research_article_is_this_a_scam_the_nature_and/ on August 3rd, 2026. 7.1.3

[486] u/ElijahBoumaSims. Research article analyzing reddit discussion about scams [r/scams], November 2025. Retrieved from https://www.reddit.com/r/Scams/comments/1p61v0m/research_article_analyzing_reddit_discussion/ on August 3rd, 2026. 7.1.3

[487] u/ElijahBoumaSims. Research article: "is this a scam?": The nature and quality of reddit discussion about scams [r/sextortion], November 2025. Retrieved from https://www.reddit.com/r/Sextortion/comments/1p5m2tp/research_article_is_this_a_scam_the_nature_and/ on August 3rd, 2026. 7.1.3

[488] D. Ueno, M. Arakawa, Y. Fujii, S. Amano, Y. Kato, T. Matsuoka, and J. Narumoto. Psychosocial characteristics of victims of special fraud among japanese older adults: A cross-sectional study using scam vulnerability scale. *Frontiers in Psychology*, 13, 2022. doi: 10.3389/fpsyg.2022.960442. 3.2, 3.2.2, 3.2.3, 3.3.2, 3.3.5

[489] UK Finance. Take five to stop fraud, 2026. Retrieved from https://web.archive.org/web/20260515045858/https://www.takefive-stopfraud.org.uk/ on May 15th, 2026. 3.5

[490] United States Census Bureau. American comunity survey data releases, 2023. https://www.census.gov/programs-surveys/acs/news/data-releases.2022.html. 6.2.2

[491] United States Congress. United States Code, Title 12, Chapter 53, Subchapter V: Bureau of Consumer Financial Protection, 2026. 10.2.2

[492] United States Consumer Financial Protection Bureau. Electronic Fund Transfers (Regulation E). Code of Federal Regulations, 2011. URL https://www.ecfr.gov/current/title-12/chapter-X/part-1005. 12 CFR Part 1005. 10.2.2

[493] United States Department of Agriculture. Child nutrition programs: Income eligibility guidelines. In *Federal Register, Volume 87*. National Archives and Records Administration, 2022. URL https://www.federalregister.gov/documents/2022/02/16/2022-03261/child-nutrition-programs-income-eligibility-guidelines. 6.2.2

[494] United States Federal Bureau of Investigation. Internet crime report 2022. Technical report, FBI Internet Crime Complaint Center (IC3), March 2023. https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf on August 17, 2026. 5, 5.2.4

[495] United States Federal Bureau of Investigation. Sextortion: What kids and caregivers need to know, 2025. Retrieved from https://web.archive.org/web/20250303042405/https://www.fbi.gov/how-we-can-help-you/scams-and-safety/common-frauds-and-scams/sextortion/financially-motivated-sextortion on March 3rd, 2025. 7.2.4

[496] United States Federal Bureau of Investigation. Internet crime report 2024, April 2025. URL https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf. 3.3.2

[497] United States Federal Bureau of Investigation. Internet crime report 2025, April 2026. Retrieved from https://www.ic3.gov/AnnualReport/Reports/2025_IC3Report.pdf on August 17, 2026. 2.1, 2.2, 2.3

[498] United States Federal Trade Commission. Consumer sentinel network data book 2023, Feb. 2024. URL https://www.ftc.gov/reports/consumer-sentinel-network-data-book-2023. 7.2.2, 7.2.2, 7.2.3

[499] United States Federal Trade Commission. Equifax data breach settlement, 2024. Retrieved from https://web.archive.org/web/20260723022847/https://www.ftc.gov/enforcement/refunds/equifax-data-breach-settlement on July 23rd, 2026. 10.2.2

[500] United States Federal Trade Commission. Refund and recovery scams, 2025. Retrieved from https://web.archive.org/web/20250208231804/https://consumer.ftc.gov/articles/refund-and-recovery-scams on February 8th, 2025. 2.4.2, 7.2.4

[501] United States Federal Trade Commission. Consumer sentinel network data book 2024, March 2025. URL https://www.ftc.gov/system/files/ftc_gov/pdf/csn-annual-data-book-2024.pdf. 1, 2.1, 2.2, 2.3, 3.3.2, 3.5

[502] United States Federal Trade Commission. New ftc data show top text message scams of 2024; overall losses to text scams hit $470 million, 2025. Retrieved from https://web.archive.org/web/20250421172343/https://www.ftc.gov/news-events/news/press-releases/2025/04/new-ftc-data-show-top-text-message-scams-2024-overall-losses-text-scams-hit-470-million on November 6th, 2025. 8

[503] United States Federal Trade Commission. Fake prize, sweepstakes, and lottery scams, 2025. Retrieved from https://web.archive.org/web/20250108204624/https://consumer.ftc.gov/articles/fake-prize-sweepstakes-and-lottery-scams on January 8th, 2025. 2.1, 2.4.2

[504] United States Federal Trade Commission. How to avoid a government impersonation scam, 2025. Retrieved from https://web.archive.org/web/20250101230035/https://consumer.ftc.gov/articles/how-avoid-government-impersonation-scam on January 1st, 2025. 2.1

[505] United States Federal Trade Commission. Investment scams, 2025. Retrieved from https://web.archive.org/web/20250209002113/https://consumer.ftc.gov/articles/investment-scams on Feburary 9th, 2025. 2.1, 2.4.2

[506] United States Federal Trade Commission. Job scams, 2025. Retrieved from https://web.archive.org/web/20250101054342/https://consumer.ftc.gov/articles/job-scams on January 1st, 2025. 2.1, 2.4.2

[507] United States Federal Trade Commission. Ftc data show people reported losing $3.5 billion to imposter scams in 2025, 2026. URL https://www.ftc.gov/news-events/news/press-releases/2026/06/ftc-data-show-people-reported-losing-3-point-5-billion-imposter-scams-2025. 1, 2.1, 2.2

[508] United States Federal Trade Commission. Ftc begins enforcing the take it down act, 2026. Retrieved from https://web.archive.org/web/20260622055339/https://www.ftc.gov/news-events/news/press-releases/2026/05/ftc-begins-enforcing-take-it-down-act on June 22nd, 2026. 10.2.1

[509] United States Securities and Exchange Commission. Conflicts of Interest Associated with the Use of Predictive Data Analytics by Broker-Dealers and Investment Advisers. In *17 CFR Parts 240 and 275*. Federal Register, 2023. URL https://www.sec.gov/files/rules/proposed/2023/34-97990.pdf. 5.3

[510] United States Securities and Exchange Commission. SEC Charges Coinbase for Operating as an Unregistered Securities Exchange, Broker, and Clearing Agency, June 2023. Retrieved from https://www.sec.gov/news/press-release/2023-102 on August 17, 2026. 5.3

[511] United States Securities and Exchange Commission. SEC Proposes New Requirements to Address Risks to Investors From Conflicts of Interest Associated With the Use of Predictive Data Analytics by Broker-Dealers and Investment Advisers, July 2023. Retrieved from https://www.sec.gov/news/press-release/2023-140 on August 17, 2026. 5.3

[512] United States Securities and Exchange Commission. SEC Charges Kraken for Operating as an Unregistered Securities Exchange, Broker, Dealer, and Clearing Agency, November 2023. Retrieved from https://www.sec.gov/news/press-release/2023-237 on August 17, 2026. 5.3

[513] United States Securities and Exchange Commission. SEC Charges Crypto Company SafeMoon and its Executive Team for Fraud and Unregistered Offering of Crypto Securities, November 2023. Retrieved from https://www.sec.gov/news/press-release/2023-229 on August 17, 2026. 5.2.3, C.2

[514] U.S. Senate. TAKE IT DOWN Act. 119th Congress, 2025. 10.2.1

[515] U.S. Senate. ReportScams.gov Act. 119th Congress, 2025. 10.2.3

[516] U.S. Senate. Romance Scam Prevention Act. 119th Congress, 2025. 10.2.3

[517] USAGov. Learn where to report a scam, 2026. Retrieved from https://www.usa.gov/where-report-scams on August 17, 2026. 7.3

[518] Iman Vakilinia. Cryptocurrency giveaway scam with youtube live stream. In *2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 0195–0200, New York, NY, Oct. 2022. IEEE. URL https://doi.org/10.1109/UEMCON54665.2022.9965686. 6

[519] Tommy van Steen, Emma Norris, Kirsty Atha, and Adam Joinson. What (if any) behaviour change techniques do government-led cybersecurity awareness campaigns use? *Journal of Cybersecurity*, 6(1):tyaa019, 01 2020. ISSN 2057-2085. doi: 10.1093/cybsec/tyaa019. URL https://doi.org/10.1093/cybsec/tyaa019. 3.5

[520] Anthony Vance, Brock Kirwan, Daniel Bjornn, Jeffrey Jenkins, and Bonnie Brinton Anderson. What do we really know about how habituation to warnings occurs over time? a longitudinal fmri study of habituation and polymorphic warnings. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*, page 2215–2227, Denver, CO, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3025896. 8.3

[521] Yash Vekaria, Nurullah Demir, Konrad Kollnig, and Zubair Shafiq. Understanding data collection, brokerage, and spam in the lead marketing ecosystem. In *2026 IEEE Symposium on Security and Privacy (SP)*, pages 2096–2115. IEEE, 2026. 10.3.2

[522] S.F. Verkijika. "If you know what to do, will you take action to avoid mobile phishing attacks": Self-efficacy, anticipated regret, and gender. *Computers in Human Behavior*, 101:286–296, 2019. doi: 10.1016/j.chb.2019.07.034. 3.2, 3.3.5, 3.4.1

[523] Markel Vigo and Simon Harper. Coping tactics employed by visually disabled users on the web. *International Journal of Human-Computer Studies*, 71(11):1013–1025, 2013. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2013.08.002. URL https://www.sciencedirect.com/science/article/pii/S1071581913001006. 10.3.1

[524] J. Vitak, Y. Liao, M. Subramaniam, and P. Kumar. "i knew it was too good to be true": The challenges economically disadvantaged users face in assessing trustworthiness, avoiding scams, and developing self-efficacy online. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 2018. ISSN 25730142 (ISSN). doi: 10.1145/3274445. 1, 2.3, 3.2, 3.2.2, 3.3.2, 3.3.2, 3.3.3

[525] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. User experiences of torpedo: Tooltip-powered phishing email detection. *Computers & Security*, 71:100–113, 2017. ISSN 0167-4048. doi: https://doi.org/10.1016/j.cose.2017.02.004. URL https://www.sciencedirect.com/science/article/pii/S0167404817300275. 3.4.2

[526] Warren J Von Eschenbach. Transparency and the black box problem: Why we do not trust ai. *Philosophy & technology*, 34(4):1607–1622, 2021. 8

[527] Kurt Wagner. Meta to reduce role of outside content moderators in favor of ai. *Bloomberg*, 2026. Retrieved from https://www.bloomberg.com/news/articles/2026-03-19/meta-to-reduce-role-of-outside-content-moderators-in-favor-of-ai on August 2nd, 2026. 10.2.1

[528] Ari Ezra Waldman. Law, privacy, and online dating:"revenge porn" in gay online communities. *Law & Social Inquiry*, 44(4):987–1018, 2019. 4, 4.2.3

[529] Ari Ezra Waldman. Navigating privacy on gay-oriented mobile dating applications. In *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*, pages 369–381. Emerald Publishing Limited, Bingley, United Kingdom, 2021. 4.1.2, 4.2.3

[530] Chenkai Wang, Zhuofan Jia, Hadjer Benkraouda, Cody Zevnik, Nicholas Heuermann, Roopa Foulger, Jonathan A. Handler, and Gang Wang. VeriSMS: A message verification system for inclusive patient outreach against phishing attacks. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, Honolulu, HI, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642027. 8.1.3

[531] Fangzhou Wang. Breaking the silence: Examining process of cyber sextortion and victims' coping strategies. *International Review of Victimology*, page 02697580241234331, 2024. 2.2, 7

[532] Fangzhou Wang. Victim-offender overlap: the identity transformations experienced by trafficked chinese workers escaping from pig-butchering scam syndicate. *Trends in Organized Crime*, pages 1–32, 2024. 3.5

[533] Jingguo Wang, Yuan Li, and H Raghav Rao. Coping responses in phishing detection: an investigation of antecedents and consequences. *Information Systems Research*, 28(2): 378–396, 2017. URL https://doi.org/10.1287/isre.2016.0680. 2.3

[534] Yizhu Wang, Haoyu Zhai, Chenkai Wang, Qingying Hao, Nick A Cohen, Roopa Foulger, Jonathan A Handler, and Gang Wang. Can you walk me through it? explainable SMS phishing detection using LLM-based agents. In *Proceedings of the Twenty-First Symposium on Usable Privacy and Security (SOUPS '25)*, pages 37–56, Seattle, WA, USA, 2025. USENIX Association. 2.4.1, 8, 8.1.4, 9, 9.2, 9.2.2

[535] Noel Warford, Tara Matthews, Kaitlyn Yang, Omer Akgul, Sunny Consolvo, Patrick Gage Kelley, Nathan Malkin, Michelle L. Mazurek, Manya Sleeper, and Kurt Thomas. Sok: A framework for unifying at-risk user research. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2344–2360, 2022. doi: 10.1109/SP46214.2022.9833643. 1, 2.2, 2.3, 3, 3, 3.1.1, 3.1.3, 3.3, 3.3.2, 3.3.2, 3.3.4, A.1.1, A.1.3

[536] J. Watson, D. Lacey, D. Kerr, P. Salmon, and N. Goode. Understanding the effects of compromise and misuse of personal details on older people. *Australasian Journal of Information Systems*, 23:1–29, 2019. ISSN 14498618 (ISSN). doi: 10.3127/ajis.v23i0.1721. 3.2, 3.2.2

[537] Miranda Wei, Madison Stamos, Sophie Veys, Nathan Reitinger, Justin Goodman, Margot Herman, Dorota Filipczuk, Ben Weinshel, Michelle L. Mazurek, and Blase Ur. What twitter knows: Characterizing ad targeting practices, user perceptions, and ad explanations through users' own twitter data. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 145–162, Berkeley, CA, USA, August 2020. USENIX Association. ISBN 978-1-939133-17-5. URL https://www.usenix.org/conference/usenixsecurity20/presentation/wei. 4.2.2

[538] Miranda Wei, Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. Anti-Privacy and Anti-Security advice on TikTok: Case studies of Technology-Enabled surveillance and control in intimate partner and Parent-Child relationships. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 447–462, Boston, MA, August 2022. USENIX Association. ISBN 978-1-939133-30-4. URL https://www.usenix.org/conference/soups2022/presentation/wei. 6.1.1

[539] Miranda Wei, Sunny Consolvo, Patrick Gage Kelley, Tadayoshi Kohno, Tara Matthews, Sarah Meiklejohn, Franziska Roesner, Renee Shelby, Kurt Thomas, and Rebecca Umbach. Understanding Help-Seeking and Help-Giving on social media for Image-Based sexual abuse. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4391–4408, Philadelphia, PA, August 2024. USENIX Association. ISBN 978-1-939133-44-1. URL https://www.usenix.org/conference/usenixsecurity24/presentation/wei-miranda-understanding. 2.2, 7, 7.2.3, 7.2.4, 10.3.3

[540] Miranda Wei, Jaron Mink, Yael Eiger, Tadayoshi Kohno, Elissa M Redmiles, and Franziska Roesner. Sok (or solk?): On the quantitative study of sociodemographic factors and computer security behaviors. In *33rd USENIX Security Symposium (USENIX Security 24)*, Philadelphia, PA, August 2024. USENIX Association. URL https://www.usenix.org/conference/usenixsecurity24/presentation/wei-miranda-solk. 2.3, 8.1.1, 8.4.1

[541] C.M. White, M. Gummerum, S. Wood, and Y. Hanoch. Internet safety and the silver surfer: The relationship between gist reasoning and adults' risky online behavior. *Journal of Behavioral Decision Making*, 30(4):819–827, 2017. ISSN 08943257 (ISSN). doi: 10.1002/bdm.2003. 3.2, 3.3.5

[542] Monica T Whitty. Anatomy of the online dating romance scam. *Security Journal*, 28: 443–455, 2015. 2.1

[543] Monica T. Whitty. Do you love me? psychological characteristics of romance scam victims. *Cyberpsychology, Behavior, and Social Networking*, 21(2):105–109, 2018. ISSN 21522715. doi: 10.1089/cyber.2016.0729. 3.2, 3.3.1, 3.3.2, 3.3.2, 3.3.5

[544] Monica T Whitty and Tom Buchanan. The online dating romance scam: The psychological impact on victims – both financial and non-financial. *Criminology & Criminal Justice*, 16(2):176–194, 2016. doi: 10.1177/1748895815603773. URL https://doi.org/10.1177/1748895815603773. 2.1, 7.2.3

[545] M.T. Whitty. Is there a scam for everyone? psychologically profiling cyberscam victims. *European Journal on Criminal Policy and Research*, 26(3):399–409, 2020. doi: 10.1007/s10610-020-09458-z. 3.2, 3.3.2, 3.3.2, 3.3.2, 3.3.5

[546] J.W.A. Witsenboer, K. Sijtsma, and F. Scheele. Measuring cyber secure behavior of elementary and high school students in the netherlands. *Computers and Education*, 186, 2022. ISSN 03601315 (ISSN). doi: 10.1016/j.compedu.2022.104536. 3.2

[547] Michael S Wogalter, Vincent C Conzola, and Tonya L Smith-Jackson. Research-based guidelines for warning design and evaluation. *Applied Ergonomics*, 33(3):219–230, May 2002. ISSN 0003-6870. doi: 10.1016/S0003-6870(02)00009-1. 8.1.4

[548] Michael S Wogalter, Christopher B Mayhorn, and Kenneth R Laughery Sr. Warnings and hazard communications. *Handbook of human factors and ergonomics*, pages 644–667, 2021. 2.4.1, 8.5

[549] Janis Wolak, David Finkelhor, Wendy Walsh, and Leah Treitman. Sextortion of minors: Characteristics and dynamics. *Journal of Adolescent Health*, 62(1):72–79, 2018. 4.2.2

[550] Min Wu, Robert C. Miller, and Simson L. Garfinkel. Do security toolbars actually prevent phishing attacks? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 601–610, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933727. doi: 10.1145/1124772.1124863. URL https://doi.org/10.1145/1124772.1124863. 2.4.1

[551] Qunfang Wu, Louisa Kayah Williams, Ellen Simpson, and Bryan Semaan. Conversations About Crime: Re-Enforcing and Fighting Against Platformed Racism on Reddit. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), April 2022. doi: 10.1145/3512901. URL https://doi.org/10.1145/3512901. 5.1.1

[552] H. Xiang, J. Zhou, and B. Xie. Understanding older adults' vulnerability and reactions to telecommunication fraud: The effects of personality and cognition. In *Lect. Notes Comput. Sci.*, volume 12209 LNCS, pages 351–363. Springer, 2020. ISBN 03029743 (ISSN); 978-303050231-7 (ISBN). doi: 10.1007/978-3-030-50232-4_25. 3.2, 3.3.2, 3.3.2, 3.3.5, 3.4.1

[553] Honglian Xiang, Jia Zhou, and Mingjun Liu. Judging online health misinformation: Effects of cyberchondria and age. In *Human Aspects of IT for the Aged Population: 9th International Conference, ITAP 2023, Held as Part of the 25th HCI International Conference, HCII 2023, Copenhagen, Denmark, July 23-28, 2023, Proceedings, Part I*, pages 284–301, Berlin, Heidelberg, July 2023. Springer-Verlag. ISBN 978-3-031-34865-5. doi: 10.1007/978-3-031-34866-2_22. URL https://doi.org/10.1007/978-3-031-34866-2_22. 3.3.2

[554] Aiping Xiong, Robert W. Proctor, Weining Yang, and Ninghui Li. Is domain highlighting actually helpful in identifying phishing web pages? *Human Factors*, 59(4):640–660, 2017. doi: 10.1177/0018720816684064. URL https://doi.org/10.1177/0018720816684064. PMID: 28060529. 2.4.1

[555] Akira Yamada, Shoma Tanaka, Yukiko Sawaya, Ayumu Kubota, So Matsuda, Reo Matsumura, Shun Umemoto, N Christin, J Nakajima, K Crichton, et al. Mobile security behavior observatory: Long-term monitoring of mobile user behavior. In *Proceedings of the Sixteenth Symposium on Usable Privacy and Security*, volume 20, 2020. 3.5

[556] Akira Yamada, Kyle Crichton, Yukiko Sawaya, Jin-Dong Dong, Sarah Pearman, Ayumu Kubota, and Nicolas Christin. On recruiting and retaining users for security-sensitive longitudinal measurement panels. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 347–366, Boston, MA, August 2022. USENIX Association. ISBN 978-1-939133-30-4. 3.5

[557] Yang Yang, Yuhong Xu, Yizhou Sun, Yuxiao Dong, Fei Wu, and Yueting Zhuang. Mining fraudsters and fraudulent strategies in large-scale mobile social networks. *IEEE Transac-*

tions on Knowledge and Data Engineering, 33(1):169–179, January 2021. ISSN 1558-2191. doi: 10.1109/TKDE.2019.2924431. 3.2, 3.2.2, 3.3.1

[558] R.E. Yoro, F.O. Aghware, M.I. Akazue, A.E. Ibor, and A.A. Ojugo. Evidence of personality traits on phishing attack menace among selected university undergraduates in nigerian. *International Journal of Electrical and Computer Engineering*, 13(2):1943–1953, 2023. ISSN 20888708 (ISSN). doi: 10.11591/ijece.v13i2.pp1943-1953. 3.2, 3.3.2

[559] R.E. Yoro, F.O. Aghware, B.O. Malasowe, O. Nwankwo, and A.A. Ojugo. Assessing contributor features to phishing susceptibility amongst students of petroleum resources varsity in nigeria. *International Journal of Electrical and Computer Engineering*, 13(2): 1922–1931, 2023. doi: 10.11591/ijece.v13i2.pp1922-1931. 3.2, 3.2.2

[560] YouTube. How does youtube manage harmful content?, 2024. Retrieved from `https://web.archive.org/web/20241128131412/https://www.youtube.com/howyoutubeworks/our-commitments/managing-harmful-content/` on November 28th, 2024. 6.4.2

[561] L. Yu, G. Mottola, L.L. Barnes, O. Valdes, R.S. Wilson, D.A. Bennett, and P.A. Boyle. Financial fragility and scam susceptibility in community dwelling older adults. *Journal of Elder Abuse and Neglect*, 34(2):93–108, 2022. ISSN 08946566 (ISSN). doi: 10.1080/08946566.2022.2070568. 3.2, 3.2.2, 3.3.2, 3.3.2, 3.4.1

[562] Yaman Yu, Saidivya Ashok, Smirity Kaushik, Yang Wang, and Gang Wang. Design and evaluation of inclusive email security indicators for people with visual impairments. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2885–2902, San Francisco, CA, USA, 2023. doi: 10.1109/SP46215.2023.10179407. 1, 2.3, 3.1.1, 3.2, 3.3.2, 9.5.3

[563] Narges Zare, Cori Faklaris, Sarah Tabassum, and Heather Richter Lipford. Improving mobile security with visual trust indicators for smishing detection. In *Proceedings of the 2025 IEEE World AI IoT Congress (AIIoT '25)*, pages 0212–0221, Seattle, WA, USA, 2025. IEEE. 8.5

[564] Zelle. Fraud & scams overview, 2025. Retrieved from `https://web.archive.org/web/20250201174250/https://www.zellepay.com/safety-education/fraud-scams-overview` on Feburary 1st, 2025. 7.2.2, 10.2.2

[565] Saijun Zhang, Meirong Liu, Yeefay Li, and Jae Eun Chung. Teens' social media engagement during the covid-19 pandemic: a time series examination of posting and emotion on reddit. *International Journal of Environmental Research and Public Health*, 18(19): 10079, 2021. 5.1.1

[566] Yining Elena Zhang, Lila Berger, and Jing Chen. Examining the effectiveness of speech and earcon alerts for aiding phishing email detection. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 68:209 – 214, 2024. URL `https://api.semanticscholar.org/CorpusID:271964094`. 3.2, 3.4.2

[567] Jiamin Zheng, Yue Deng, Jessica Chen, Shujun Li, Yixin Zou, and Jingjie Li. Characterizing scam-driven human trafficking across chinese borders and online community responses on rednote. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Com-

puting Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3791786. URL https://doi.org/10.1145/3772318.3791786. 3.5

[568] Sarah Zheng and Ingolf Becker. Presenting suspicious details in User-Facing e-mail headers does not improve phishing detection. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 253–271, Boston, MA, August 2022. USENIX Association. ISBN 978-1-939133-30-4. URL https://www.usenix.org/conference/soups2022/presentation/zheng. 2.4.1

[569] Sarah Y. Zheng and Ingolf Becker. Checking, nudging or scoring? evaluating e-mail user security tools. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS '23)*, pages 57–76, Anaheim, CA, USA, August 2023. USENIX Association. ISBN 978-1-939133-36-6. 8.1.1, 10.3.1

[570] Sijie Zhuo, Robert Biddle, Yun Sing Koh, Danielle Lottridge, and Giovanni Russello. Sok: Human-centered phishing susceptibility. *ACM Trans. Priv. Secur.*, 26(3), April 2023. ISSN 2471-2566. doi: 10.1145/3575797. 2.1, 2.3, 3.3.2, 8.1.1, 9.3.5