Against Online Abuse and Toward Sociotechnical Security & Privacy

Miranda Wei

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Commitee:

Tadayoshi Kohno, Chair

Franziska Roesner, Chair

Richard Anderson

Elissa M. Redmiles

Program Authorized to Offer Degree:
Paul G. Allen School of Computer Science & Engineering

© Copyright 2025

Miranda Wei

University of Washington

Abstract

Against Online Abuse and Toward Sociotechnical Security & Privacy

Miranda Wei

Co-Chairs of the Supervisory Committee:
Tadayoshi Kohno
Franziska Roesner
Paul G. Allen School of Computer Science & Engineering

Technology-facilitated abuse is an escalating challenge—from toxic content on social media to image-based sexual abuse—with a significant impact on the well-being of people globally. This dissertation advocates for a sociotechnical approach to computer security and privacy to understand and combat technology-facilitated abuse. Sociotechnical approaches investigate systems of people and technology and require studying the complexities of their interconnections. In the first part of this dissertation, I evaluated existing mitigations for online abuse, finding that they place an undue burden on individuals and have limits in addressing the systemic challenges of online abuse. In the second part, I next analyzed human-centered security and privacy research and identified opportunities for the increased consideration of social factors, such as through operationalizing sociodemographic factors or gender stereotypes. Finally, in the third part, I applied sociotechnical threat modeling to characterize two emerging forms of online abuse: interpersonal surveillance and control and synthetic nonconsensual explicit imagery (also called "deepfake nudes"). Throughout this dissertation, I also study how gender and interpersonal relationships contribute to disparate experiences of safety. Theoretically, this dissertation connects computer

science inquiry with feminist science and technology studies (feminist STS) and sociology; this enables supplementing inductive descriptions of online abuse with social theories of how and why online abuse happens. Practically, this dissertation integrates people's lived experiences and social perspectives into security and privacy research, informing researchers, policymakers, and industry practitioners on more sociotechnically informed mitigations for online abuse. This dissertation underscores the transformative potential of a sociotechnical approach to security and privacy not only for mitigating online abuse but also for security, privacy, and safety at large.

Contents

List of Figures

List of Tables

Acknowledgements

At my (first) Security Lab PhD Visit Days event, we played Jackbox Games and I was thrilled to *not* get the inside jokes. People seemed to genuinely like each other and I understood that the lab and its (then two) leaders were unique. I am immensely fortunate to have had the mentorship of two extraordinarily kind, dedicated, and open-minded scholars: my co-advisors, Franzi Roesner and Yoshi Kohno. Their support, intellectual collaboration, and Slack memes have made possible work that I never would have imagined. I am continually inspired by Franzi's astute critical thinking, Yoshi's creativity for out-of-the-box approaches, and enthusiasm from both (my writing has picked up some exclamation points along the way!!). During our first meetings, they told me that their goal as professors is to help PhD students figure out and get to wherever they want to go. I attest to their success in this regard and I aspire to pass on such support.

They say, "shoot for the moon: even if you miss, you'll end up among the stars." I misheard and thought, "start with amazing mentors: even if there's a pandemic and other unprecedented crises, maybe I can still do a PhD." I have had incredible mentors who guided me through immense challenges, both before and during my PhD. I am indebted to Blase Ur for funding me to go to SOUPS 2017 even though my poster was rejected; attending that conference changed my life. I continue to benefit from his generosity in connecting me within his academic network. In just two post-bac/pre-doc years, Blase and Michelle Mazurek trained me in research methods, academic writing, time management, and other critical academic skills. I thank Maximilian Golla

for patiently fixing the countless LaTeXor Git mistakes I've made over the years, particularly in the bibliography. I will never get tired of the memes and maybe one day my German will finally be good enough to read them all. My love for user studies was in large part surfaced by Elissa Redmiles' meticulous proficiency with survey methodology among other methods; Elissa is one of a select few who paved the way for me to rest assured that qualitative and mixed methods belong in computer science. From these scholars, I had incredible role models (and thus developed unrealistically high expectations for myself) before I even started my PhD. Throughout my time at UW, I learned from Richard Anderson's candidly shared wisdom from a rich career in ICTD and computer science. Thank you for always listening to my existential crises and providing an essential outside perspective on my work.

My academic achievements would not have been possible without an inspiring set of collaborators, including Grace Brigham, Sunny Consolvo, Kovila Coopamootoo, Yael Eiger, Pardis Emami-Naeini, Matthias Fassl, Kurt Hugenberg, Apu Kapadia, Patrick Gage Kelley, Yoshi Kohno, Tara Matthews, Jessica McClearn, Sarah Meiklejohn, Jaron Mink, Collins Munyendo, Chris Page, Lucy Qin, Franzi Roesner, Renee Shelby, Sophie Stephenson, Kurt Thomas, Emily Tseng, Rebecca Umbach, Tina Yeung, and Eric Zeng. I thank the many physical and mental health professionals who have supported my wellness during six grueling years, especially Coll and Dante. Through academic conferences, I have cultivated an international network of friends that make me feel like a Real Academic™ and that I will always say "omg, heeey!" to, including Verena Distler, Lea Gröber, Carolyn Guthoff, Catherine Han, Sharon Heung, Deepak Kumar, Victor Le Pochat, Frank Li, Cindy Lin, Alexandra Nisenoff, Lea Schönherr, Emily Tseng, Yixin Zou, and many more. Thank you also to Nassim Parvin for teaching Feminist STS Studio, and countless other contemporary scholars who create new spaces to critique and reimagine scholarship.

Thank you to Grant Ho for giving me one of the best pieces of advice about how to pick where I would go for my PhD: pick the place where you're excited to go in every day... whose

faces do you want to see when your paper is rejected? Kaiming Cheng and Kentrell Owens, I will cherish our five years of hilarious hijinks through some of the toughest parts of PhD and life. I'm still ecstatic that I bamboozled Eric Zeng into writing a paper about TikTok with me, and hope someday we might write another, or at least do some skiing. To Maddie Burbage, Yael Eiger, and Tina Yeung: you all are the heroes we don't deserve but desperately need. Your incredible commitments to community care and advocacy give me hope for the future. Thanks to David Kohlbrenner and Nirvan Tyagi for excellent advice during my academic job search and best of luck when the parents aren't home. Thank you to the many other current and former members of the UW Security and Privacy Lab who have and continue to make it a welcoming, diverse, and intellectually stimulating environment, including Christine Chen, Inyoung Cheong, Ivan Evtimov, Aarushi Dubey, Michael Flanders, Chris Geeng, Theo Gregersen, Gregor Haas, Taylor Hansen, Rachel Hong, Karl Koscher, Umar Iqbal, Evan Lam, Rachel McAmis, Alexandra Michael, Peter Ney, Basia Radka, Mattea Sim, Lucy Simko, Anna Kornfeld Simpson, and Henry Wong. And thank you to Pelicana for being a sure bet for happy hour and usually not running out of pickled radish.

I thank the excellent UW staff whose labor make so many tasks seem invisible: Joe, Elise, Les, Christopher, and other grad advising staff; Stephanie, Bill, and other financial specialists; Zaid, Amber, Sean, and other facilities staff. I express my gratitude to and solidarity with UAW4121 for tenaciously supporting academic student employees at the University of Washington.

In the sage words of Han and Dom: "*who you choose to be around you lets you know who you are*" and "*I don't have friends, I've got family.*" I will dearly miss the UW CSE PhD student community, including Willie Agnew, Chris Geeng, Madeleine Grunde-McLaughlin, Rachel Hong, Jason Hoffman, Tae Jones, Innocent Obi, Sudheesh Singanamalla, Anna Spiro, Priyal Suneja, Galen Weld, Ryan Zambrotta, Matt Ziegler, and many more. No matter where I am in the world, I've relied on the lifelong kinship of Ahmed, Andrew, Eujene, Dylan, Gadiel, Ivan, Shiv, Sunny, and Val. Finally, thank you to Eno, Matthias, and Mura for showing me what home means.

To those who are silent.

Chapter 1

Introduction

Reports from across academia, industry, and government show that abuse facilitated by technology is pervasive and on the rise. Nearly one in four people in the US know someone who was doxxed[1] on social media [529]. The DOJ estimated that over 900,000 Americans were cyberstalked in 2019 [398]. During the 2023-2024 US school year, deepfakes of someone at school were reported by 40% of students [326]. Though this dissertation focuses on the US due to the location of its author, prevalence and proliferation of abuse is global. One in seven adults across ten countries have experienced threats to distribute intimate content [264]. UK police reported a 317% increase in cyberstalking cases, such as Apple Airtags, from 2018 to 2023 [621], while industry reported an 84% increase in deepfake videos between 2018 and 2019 [156], and then a 550% increase by 2023 [269]. These deepfakes have targeted diverse groups, from South Korean musicians to politicians in Gabon, Malaysia, Ukraine [156, 519], and beyond.

Despite significant research and a vibrant landscape of advocacy efforts, the full severity of abuse facilitated by technology remains unknowable. Abuse that targets public figures, such as

[1]Doxxing refers to someone's private information, such as home address or other "documents" (docs, or dox), being publicly shared without consent.

politicians, journalists, or entertainers, is likely more widely known, but other abuse happens to private individuals in their homes, schools, and workplaces. Stigma impairs the reporting of many forms of abuse [20, 19], as people who have been targeted may be reluctant to disclose or seek institutional support. However, known personal accounts of various forms of TFA frequently describe it being the worst experiences of their lives, with vast and dramatic impacts on people's finances, physical health, emotional and sexual well-being, and more [37, 55, 302, 495, 279, 381].

Addressing the vast scope of technology-facilitated abuse will require the contributions of a wide range of stakeholders. Contributions may be simultaneously bottom-up and top-down, coming from grassroots efforts and social movements, as well as from experts in specialized domains, such as in law and policy, social work, and computer science. Technology-facilitated abuse is a deeply sociotechnical challenge, requiring people across social and technical domains to work together to imagine and implement mitigations for abuse. To facilitate collaboration between people from across domains, a shared definition is critical.

Defining technology-facilitated abuse

From the above examples, an intuitive definition of technology-facilitated abuse (TFA) is:

> **TFA, def.(a):** *An individual experience of severe physical, emotional, financial, sexual, or other harm inflicted by another person, that happens online or is mediated by technology.*

In 2021, Thomas et al. [565] conducted a review of over 150 research papers to create a taxonomy of online hate and harassment. This systemization of knowledge (SoK) provided a more detailed picture of online abuse, inclusive of online hate and harassment, than could be previously found in the computer security and privacy literature. The SoK describes online hate and harassment through seven categories of attacks, including specific attacks like cyberbullying, sextortion, doxing, stalking, and more. Bailey et al.'s *Emerald International Handbook of Technology-Facilitated*

Violence and Abuse also introduces technology-facilitated violence and abuse (TFVA) as an umbrella term describing numerous abusive behaviors, including but not limited to image-based sexual abuse, cyberstalking, and hate speech [32], while Koukopoulos et al. used three rounds of surveys with 316 experts to conceptualize and define TFA [315]. These works come from different disciplines, but share an inductive approach—whether through prior work, notable examples, or established experts—to defining technology-facilitated abuse and threats to online safety.

Another definition of TFA could be derived by analogy from sociologists' conceptualization of abuse in professional settings. The American Sociological Association (ASA)'s Code of Ethics states that "harassment consists of a single intense and severe act or of multiple persistent or pervasive acts which are demeaning, abusive, offensive, or create a hostile professional or workplace environment" [30]. In a 2016 statement, the ASA further clarified the relationship between abuse, harassment, and power:

> Sociologists understand that social organizations are stratified along a number of dimensions and that means that power is unequally distributed among the members of any organization. Some power differentials are brought into organizations in the very persons and social identities of the people who make them up. Others are a function of the positions different people hold within organizations... Harassment is, above all, an abuse of power. It is a form of abuse which demeans and disempowers persons with whom one has a professional relationship such that they are unable to do their best work and have themselves and that work seen in the most favorable light possible, and to grow and thrive as human beings. [31]

Substituting the professional environment with an online environment and digital technologies, another definition of TFA could be:

> **TFA, def.(b):** *Severe or persistent harm facilitated by technology that is demeaning,*

> *offensive, or creates a hostile environment that otherwise impairs people's ability to thrive as human beings, particularly as enabled by socially constructed power differentials.*

In contrast to systematically inductive approaches, this approach defines abuse through a systemic lens, drawing on decades of sociological theory.

Both definitions have their merits and limitations. Particularly for computer security and privacy researchers, **def.(a)** provides a more technically grounded and approachable definition. Yet, in relying on "I know it when I see it" reasoning, this definition risks arbitrary disagreement between individuals and lacks a generalized theory of its causes. In the eyes of human-computer interaction (HCI) researchers or social scientists, **def.(b)** provides a more societally informed approach. However, the immense scale and severity of TFA are due to its technology-facilitated nature, which requires the interdisciplinary expertise of both social scientists and computer scientists to jointly combat. I seek to draw on both definitions and corresponding bodies of literature to outline an approach that is both technically grounded and societally informed. Merging both definitions, I define TFA as:

> **Technology-facilitated abuse (TFA):** *Severely or persistently harmful acts facilitated by technology that are demeaning, offensive, or create a hostile environment that otherwise impairs people's ability to thrive as human beings, particularly as enabled by socially constructed power differentials. Such harm may be but is not limited to physical, emotional, financial, or sexual harm. Examples include online hate and harassment, interpersonal violence and coercive control, online scams, and more.*

This definition centers the social factors and conditions which bring about TFA, while also offering identifying characteristics and key examples that aid in designing technical solutions. Further, this exercise in definition demonstrates the key thesis of this work: that working against technology-facilitated abuse informs and requires a sociotechnical approach to computer security and privacy.

Structure and summary of this dissertation

This dissertation interrogates the predominant paradigm in human-centered computer security and privacy that regards one experience of harm as an individual experience (see **Chapter 2**, subsection *Marginalized, Vulnerable, and At-Risk Users in Computer Security and Privacy*). Rather, I advance a sociotechnical approach to computer security and privacy (S&P), grounded in my work aimed to research and combat technology-facilitated abuse (TFA). Specifically, I focus on online abuse, a large subcategory of TFA that is characterized by the (ab)use of online technologies, i.e., social media or other internet-connected devices and platforms. Today, a majority of technologies function by connecting to the internet, making most TFA also online abuse. For example, many instances of stalking, harassment, or coercion occur through social media or online messaging platforms. Online abuse can be an extremely severe and happens to people of all levels of security and privacy, but it accounts for a small portion of the computer science literature.

Part one: Limitations in current online abuse mitigations. In the first part of this dissertation, I evaluate two existing mitigations for online abuse: through online advice and through help-seeking on Reddit. Security and privacy advice is often pointed to by well-meaning advocates as a way to proactively address concerns; through interviews with experts on online hate and harassment, I evaluated a breadth of advice available online for addressing hate and harassment concerns (**Chapter 3**). We develop a ranked list of advice suitable for general audiences about how to mitigate or protect themselves from online abuse, as well as surface the underlying criteria that experts use to evaluate advice for addressing online hate and harassment. I next evaluated help-seeking posts on Reddit for image-based sexual abuse (**Chapter 4**). We find that users post for a range of image-based sexual abuse—from nonconsensual sharing to synthetic image creation, and from sextortion scams to recorded sexual assault—and uniquely characterize the wide range of this abuse with first-person reports not found in prior work. We identify patterns

in the genders and relationships of perpetrators and victim-survivors across multiple types of image-based sexual abuse, and summarize the critical types of help needed; this help is sometimes provided by commenters, but other times commenters fall short.

Together, this first part identifies helpful components of existing mitigations, while also documenting limitations related to their dependence on individual action. Individual approaches can provide immediate relief against online abuse and thus may be necessary, but are insufficient on their own. Distilling insights from these two studies, I find that these existing mitigations risk amplifying the idea that individuals must bear the majority of the burden to stop online abuse. The online guides direct advice at individuals, which could imply that individuals who do not take the proper precautions shoulder responsibility in not stopping their own abuse. Reddit forums similarly require individual initiative to post, but not all individuals will think to seek help online. No individual can fully prevent being subject to online abuse if it is enabled by vast, historical social conditions. Therefore, researchers must also look beyond the individual level.

Part two: Mapping societal factors as a new approach. In the second part of this dissertation, I step outside online abuse to look broadly at the state of human-centered security and privacy research. Beyond individuals, researchers have used sociodemographic factors to measure differences in security and privacy behaviors. I conducted a literature review of prior quantitative studies of security and privacy behaviors and how they vary between people of different genders, races, cultures, or other sociodemographic factors (**Chapter 5**). We find evidence of some trends for some sociodemographic factors (e.g., gender, age), but vast unknowns still remain. Comparatively little quantitative research has explored race or income, and further, most work surfaced in this literature review did not provide causal explanations for differences. Therefore, we develop guidelines for applying sociodemographic factors in future human-centered security and privacy research.

To better understand the nuanced relationship between sociodemographic factors and security and privacy, I focused on gender and posed a specific mechanism: if people have different expectations of their own or others' abilities because of gender stereotypes, these beliefs could partially explain differences in behaviors. I surveyed people in the US and indeed found evidence of stereotypes that women were more likely to be emotional and gullible, and to take poor security and privacy actions, while stereotypes of men assumed they were more likely to be engaged with such topics and take protective actions (**Chapter 6**). Establishing this mechanism corroborates the view that the field of S&P can benefit from a sociotechnical approach, which connects studies of technical behaviors with sociological concepts and mechanisms.

By mapping societal factors and how they relate to technical security and privacy, researchers can build on individual experiences to advance the study of online safety and harm. The field of computer security and privacy has and continues to tackle complex issues involving technical and social systems, such as supply chain vulnerabilities, mass surveillance, or data privacy concerns from advertising; the expertise of computer security and privacy researchers have made meaningful strides in the usability of security notifications [486, 10, 616, 225] and other end-user security and privacy tools [603, 335]. Researchers have also pioneered investigations into social and technical aspects of security and privacy, such as differing conceptions of security and human values [412], or security and privacy in domains of unequal power relations such as healthcare [349], families [553], immigration [431], intimate partner violence [575], and much more. Yet, explicitly naming this work as sociotechnical is a relatively new and uncommon occurrence by security and privacy researchers.[2] In this dissertation, I draw on histories of social and sociotechnical research

[2]For a partial review of security and privacy research that uses the term *sociotechnical*, see the following works, all published in the last fifteen years. With respect to software engineering, Bruce Schneier defined sociotechnical debt as "the long-term costs that result from avoiding or not fully addressing social needs in the present [in cultural and social infrastructure]" [514] and Ross Anderson gave lectures about the dependability of sociotechnical infrastructure [24, 23]. Philip Rogaway provided a sociotechnical critique of techno-optimism in computer science [492]. Degeling et al. outlined sociotechnical design for privacy [157], Goerzen et al. sketched sociotechnical security as a framework [223], and McGregor et al. provided a case study of a successful sociotechnical security system [388].

from outside computing (see Section 2.2) to operationalize sociodemographic factors and gender stereotypes as one way to do explicitly sociotechnical security and privacy research. Further, I position online abuse as a crucial challenge for security and privacy researchers to both center and address the sociological considerations of power and oppression in security and privacy.

Part three: Using sociotechnical threat modeling to characterize emerging online abuse threats. In the third part of this dissertation, I return to the topic of online abuse to demonstrate how researchers can draw on a sociotechnical approach to characterize threat models for emerging forms of online abuse. Doing so incorporates prior social theory and also identifies promising avenues for lasting mitigation of online abuse.

Applying sociotechnical threat modeling to selected TikTok content, I find that people share advice on how to harm their partners' and children's S&P through interpersonal surveillance and control (**Chapter 7**). Within relationship and parenting subcommunities on TikTok, creators make videos to share many techniques, such as for surreptitiously tracking the location of or reading messages of people in close relationships. The perpetrators motivations were deeply social—to manage information and exert control in interpersonal relationships—and shaped by online discourse about gender, with implications for the design of technologies. By surfacing these techniques, we enabled designers and practitioners to implement mitigations. Further, we demonstrated how TikTok is a valuable source of information for tracking new techniques for online abuse.

Applying sociotechnical threat modeling to schools, I interviewed middle and high teachers to understand the emerging threat of students generating synthetic nonconsensual explicit imagery (**Chapter 8**). As experts on child development and student behavior, teachers were attuned to potential motivations for this form of image-based sexual abuse perpetration as well as the potential interventions possible within the school context. Teachers highlighted that motivations

would include gender and sexual abuse, showing how interpersonal violence and trauma manifest in technologies ostensibly with non-abusive intentions.

Conclusion. I conclude this dissertation by reflecting on this sociotechnical approach and the potential of emancipation and liberation for security.

Contributions

This dissertation makes multifaceted contributions that are practical and theoretical to the fields of S&P and HCI. The papers in this work have been published in leading security, privacy, and HCI venues: USENIX Security, CHI, IEEE S&P, and SOUPS. Further, this work has informed policymakers,[3] industry practice, and my co-organization of three academic workshops.[4] Above all, this dissertation draws connections between disparate fields, methods, and perspectives to create a meeting point for future work against technology-facilitated abuse and toward sociotechnical security and privacy.

Practical. S&P practitioners have long depended on threat models to clearly and precisely characterize potential harm to software or hardware systems. By applying threat modeling in a sociotechnical way to online abuse, specifically for interpersonal surveillance and control (Chapter 7) and synthetic nonconsensual explicit imagery (Chapter 8), this dissertation identifies perpetrators of harm, as well as their motivations and techniques for doing so in specific contexts. The development of mitigations for online abuse is only possible with accurate threat models. Further, in evaluating existing mitigations for online abuse, Chapter 3 and Chapter 4 identify how individual action (through advice for proactively managing online hate and harassment, or seeking help on Reddit for image-based sexual abuse) can be help, but ultimately needs industry-backed,

[3] Comments to the FTC re: Commercial Surveillance ANPR, R111004 and presentations to the Commission Nationale de l'Informatique et des Libertés (CNIL, France's data protection authority) in 2024 and 2025.

[4] At SOUPS 2024 and 2025, Gender, Online Safety, and Sexuality (GOSS); at CSCW 2025, Reflexivity & Reflection (R&R) for Sociotechnical Safety, and Co-Constructing the Future of Digital Intimacy.

advocacy organization-led, or other forms of institutional support. Overall, the work described in these chapters have and will continue to inform practical mitigations for widespread online abuse.

Theoretical. Inductive definitions of TFA, while practical and accessible, can be limited in their generalizability. By connecting studies of online abuse, particularly involving computer science methods and perspectives, to sociotechnical and sociological fields of inquiry, researchers can theorize about the underlying mechanisms and factors. Theorizing is powerful for its capacity to not only describe but also explain. This enables researchers to conceptualize nuanced factors, such as sociodemographic factors or societal discourse, and to develop richer perspectives, including causal sociotechnical explanations for phenomena. This dissertation has theorized about the role of gender (Chapter 6), interpersonal relationships (Chapters 4 and 7), and other identity characteristics (Chapter 5) in online abuse and sociotechnical S&P more broadly. Rather than viewing online abuse as random and unpredictable, this dissertation generally theorizes that systemic oppression enables abuse and harm. The conclusion (Chapter 9) discusses possibilities for further leveraging sociotechnical thought to imagine and apply transformative justice approaches to technology-facilitated abuse.

For Security & Privacy. Though TFA and online abuse have existed for as long as technology and the internet have been ubiquitous, only in the last few years has the S&P research community mentioned online harassment or abuse in their Call for Papers.[5] This dissertation builds on the initial recognition to build momentum for more widespread focus on various forms of TFA. Researchers and practitioners can use the specific examples and techniques identified in this work to develop new mitigations, with a better understanding of the constraints and people's needs. Further, this dissertation augments technical perspectives of S&P and computer science with people's lived experiences of safety—including children, parents, adults in intimate relationships, social

[5]"Online abuse and harassment" first appeared in the Call for Papers of USENIX Security in 2021, and "Hate, Harassment, and Online Abuse" first appeared in the Call for Papers of IEEE Security & Privacy in 2022.

media users, and more—equipping S&P practitioners to tackle some of the greatest technological challenges of our times.

For Human-Computer Interaction. Given that TFA consists inherently of interactions between humans and computers, the aforementioned practical and theoretical contributions of this dissertation are also relevant for HCI. This subsection will more specifically enumerate the methodological and empirical contributions [613] for non-S&P HCI.

This dissertation contains two novel case studies for collecting and analyzing social media data: an LLM and human coder pipeline for surfacing nuanced sociotechnical concepts in Reddit data (Chapter 4) and a process for using TikTok data to monitor emerging forms of online harm (Chapter 7). Follow-on work is being conducted based on the former and has already been published based on the latter [553, 219]. While similar methods are more mainstream in HCI, these case studies demonstrate the wide applicability of HCI methods to various areas of computing, i.e., S&P; non-S&P HCI researchers can continue to adapt such methods for other topics.

S&P's conception of threat models are somewhat analogous to HCI's conception of user personas. Thus, Chapter 7 develops user personas of perpetrators of interpersonal surveillance and control, while Chapter 8 develops user personas of students who generate synthetic nonconsensual explicit imagery. In studying TFA and online abuse, this dissertation enriches the HCI community's empirical understanding of negative experiences that arise when people and technology interact.

HCI researchers and practitioners have been concerned with social computing and sociodemographic factors for far longer than S&P. This dissertation provides further empirical evidence for how social factors shape behavior, such as how social media platforms enable help-seeking (Chapter 4), or how gender stereotypes translate between technical domains (Chapter 6). Further, HCI is inherently sociotechnical, as evident by the sociotechnical gap, which Ackerman identifies as one of the central problems for HCI: "the divide between what we know we *must* support

socially and what we *can* support technically" [5]. The pervasiveness of online abuse demonstrates the ongoing limitations of existing technologies to ensure the safety of all people. This dissertation contributes to the bridging of this gap by bringing S&P, HCI, and other social perspectives to bear on the immense challenges of online abuse.

Chapter 2

Background

This dissertation draws on knowledge from many different academic disciplines in an effort to advance sociotechnical safety (people's safety within the interplay of social systems and technical systems), which includes challenges in online abuse research as well as security and privacy research. It is impossible to fully disambiguate technical and social perspectives, but for the purposes of this background chapter, I characterize each separately. Where relevant, this background also forward references areas of work that are presented in later chapters' related work sections.

2.1 Technical Perspectives Related to Sociotechnical Safety

Within technical perspectives to sociotechnical safety, computer science researchers have primarily focused on engineering, i.e., improving the design of existing systems or inventing new systems. For decades, computer security and privacy has explored the implementation and vulnerabilities of computing systems from all levels of the software stack. Most relevant to online abuse is research on human-centered security and privacy. Much work about how people interact with

security and privacy systems can be traced to the subfield of usable security and privacy, which is primarily concerned with designing and evaluating the usability of computer systems with respect to security and privacy. Such work is often published in the Symposium On Usable Privacy and Security (SOUPS), as well as other leading technical computer science conference venues: USENIX Security Symposium, the IEEE Symposium on Security and Privacy (IEEE S&P), the ACM Conference on Computer and Communications Security (CCS), and the ISOC Network and Distributed System Security Symposium (NDSS). Research focused on people's interactions with security and privacy technologies is also published in leading human-computer interaction or social computing venues, such as the conference on Human Factors in Computing Systems (CHI) or the Conference on Computer-Supported Cooperative Work and Social Computing (CSCW), among others.

Usable Security and Privacy. Often attributed to either Saltzer and Schroeder [499] in 1975 or Adams and Sasse [6] in 1999, technical security and privacy research turned towards the role of the user as the agent interacting with computer systems. Early questions in the field concerned the usability of computing systems as the primary barrier towards security and privacy, e.g., the usability of PGP, an email encryption program [603]. In the following decades, research exploded on usability topics, including security and privacy practices, e.g., selecting passwords and authenticating [581, 224], managing online privacy [298], reacting to security warnings and notifications [11, 178], as well as security and privacy tools, e.g., password managers [110], privacy dashboards [185]. For more specific related work on security advice, see Part III, Section 7.2.2. The focus on users also invited research on different stakeholders in the security and privacy process, e.g., software developers [4, 229], corporate employees [239, 127].

Marginalized, Vulnerable, and At-Risk Users in Computer Security and Privacy. Within and adjacent to usable security and privacy, researchers are also identifying that some people

disproportionately experience security and privacy risks and harms. Warford et al. identified ten contextual factors that might increase security and privacy risk, e.g., societal oppression, relationship with an attacker, or public prominence [592]. People might experience additional risks because of their identity, e.g., LGBTQ individuals [212], South Asian women [503] (see more related work on gender in S&P research in Part II, Section 6.2.1), because of their employment, e.g., journalists [386], sex workers [379], activists [145], or because of life events, e.g., undocumented migration [234], surviving intimate partner violence [203, 573], seeking refuge [533]. These risks include being targeted more for online hate and harassment (see more related work in Part I, Section 3.2), having greater difficulties identifying scams and phishing [593], being targeted by nation state actors [361, 360], or otherwise facing harms to their security and privacy. For more specific related work on at-risk populations in S&P research, including qualitative studies and meta-analyses of user studies, see Part II, Section 5.8. This part of the literature begins to address critiques from privacy and surveillance scholar Alice Marwick, who found that some privacy research made issues of power invisible [365].

Technical Perspectives on Online Abuse. Users who are at-risk of facing online abuse face unique challenges. Many forms of online abuse are perpetuated by someone who is known to the target, in contrast to traditional adversaries studied in security and privacy, like governments or companies. Security and privacy researchers call these perpetrators "interpersonal adversaries," and interpersonal adversaries often have more intimate and private knowledge about their target than other kinds of adversaries, so they can have a more violating impact on someone's lived experience of security and privacy. For more specific related work on interpersonal safety in S&P research, specifically between intimate partners or parents and children, see Part III, Section 7.2.1.

Security and privacy researchers also often use large datasets, such as from publicly available data gathered from social media or the internet, in order to empirically measure online abuse at

scale [26, 199, 16]. These measurements can then inform the design of systems. For more specific related work on technical mitigations against online hate and harassment in the form of online platform safety features, see Part I, Section 3.2. Large-scale measurements may also be used by policymakers to inform new law and policy. For more specific related work on youth online safety as well as policy and regulation about child sexual abuse material (CSAM), see Part III, Section 8.2.2. However, over-reliance on quantitative empirical methods can be problematic, not only for privileging one method to the exclusion of other valuable methods. As recounted by Arvind Narayanan, a distinguished quantitative researcher in computer science, the current practice of quantitative methods often signifies a willingness to see the status quo as unproblematic and justify this status quo, despite vast inequalities in reality [406].

Adjacent to security and privacy research, social computing researchers also study online abuse. These perspectives might come from analyzing social media platforms (e.g., see early TikTok research in Part III, Section 7.2.3), but also specific forms of harm, such as nonconsensual intimate media [465, 464]. As a field long focused on the implications of networked social computing platforms, this field has also explored types of online interactions related to online abuse, such as help-seeking (see Part I, Section 4.2.2).

There are complex social, political, and economic forces that generate the conditions that researchers use to define these populations of marginalized, vulnerable, and at-risk users. For example, women are marginalized in a patriarchal society that affords men more social, political, and economic privileges [42]; sex workers are only vulnerable in a whorephobic society that fears sex workers [120] and criminalizes sex work [35, 13]. The prevailing trend in technical perspectives, especially within security and privacy research, is to focus on well-defined user populations in specific contexts. Categorization systems are widespread in many domains and often regarded as effective or beneficial, but ultimately serve some points of view at the expense of others [64]. Further, this fragmentation of user populations risks establishing rigid silos of knowledge without

shared information. How might technical researchers interested in supporting marginalized, vulnerable, and at-risk users proceed from here in ways that bypass the fragmentation problem?

2.2 Social Perspectives Related to Sociotechnical Safety

Social perspectives to sociotechnical safety can be found in nearly all corners of social science and the humanities because safety and harm have always been foundational issues in society. As an interdisciplinary scholar, I have relied on ideas and theory [493] with origins in sociology, feminist and gender studies, and STS, but also from psychology, political science, and criminology, to guide my thinking and approach. Though a comprehensive summary of these fields is beyond the scope of this work, I attempt to trace key ideas that informed my research to their origins outside of computing. Doing so is imperative for ethical research [470] and for citational justice [321]. For the present purposes, I begin by providing some context on feminist, science and technology studies (STS), and critical studies of social systems. Then, I review perspectives on abuse and harm in the social sciences.

Feminist, STS, and Critical Studies. A key idea from science and technology studies (STS) is that the social and technical aspects of a sociotechnical system are interdependent, such that understanding each alone is impossible [103]. Studying both aspects together is multiplicative rather than additive: this approach raises lines of inquiry about how broader societal forces inform technical design and engineering, how what is technologically possible informs social experiences, and many other interdisciplinary questions. Sociotechnical approaches have been applied in many contexts since the term was coined by British sociologists in the 1950s, such as business and management, human-computer interaction, and AI policy [572, 103]. This dissertation takes a sociotechnical approach to security and privacy by considering people's online safety, i.e., experiences of online abuse, computer security and privacy attitudes and behaviors, as co-produced

by both existing technologies and societal factors.

In interpreting societal factors, I have been immensely inspired by the work of feminist and feminist STS scholars, as well as their contributions to and within HCI. This work analyzes the influence of interlocking systems of oppression—such as through gender but also race, class, and many others—on society. Using an analytical framework that acknowledges social inequality is created by multiple axes of social division is often credited to Kimberlé Crenshaw's coining of intersectionality and her pivotal work on Black women's experiences in the US [136, 137], but had also been employed in various ways before or in different contexts [470]. Broadly, feminism interrogates patriarchal institutions, and intersectional feminism contends that sexist institutions form a matrix of domination in combination with white supremacist, colonial, ableist, and other oppressive institutions [137, 42, 166]. This structural standpoint has been transformative in shifting my focus from beyond solely individuals to the broader social systems that perpetuate differences, inequities, and ultimately threats to sociotechnical safety.

Another key area of influence from feminist studies on my research is epistemological, stemming from Sandra Harding and other feminist scholars' investigation of what is "objective" and whose knowledge is privileged [242]. Put succinctly by feminist writer Andrea Dworkin: "While gossip among women is universally ridiculed as low and trivial, gossip among men, especially if it is about women, is called theory, or idea, or fact" [172]. Feminist standpoint theory, as developed by key thinkers also including Sandra Harding, Donna Haraway, and Patricia Hill Collins, contends that all knowledge is situated and that standpoints from the margins will be especially informative [242, 240, 121], e.g., Black women will have different perspectives on the patriarchy and white supremacy than Black men or white women as a result of their unique experiences. Together, these feminist epistemological concerns have prompted me to position online abuse as a vital and pressing challenge for security and privacy research, rather than dismissed as a societal ill with no possibility of (socio)technical intervention. As a specific suggestion for researchers,

standpoint theory encourages reflexivity in considering their positionality when conducting research [270, 274], including disclosing that positionality in research reports if relevant and safe to do so [343]. Where possible, positionality statements are included in the individual chapters of this dissertation.

Many feminist perspectives are similar to critical studies as well as science and technology studies (STS) in highlighting the hegemonic influence of institutions. An influential account from STS is that technology shapes and is shaped by the politics of the designers and the broader societal environment in which it was created and is used, whether explicitly or implicitly [609, 130].[1] Over 40 years ago, Winner critiqued the assumption that technologies are neutral [609], reflecting the practice of questioning normative views of what is commonly observed [89]; critical studies, which is not restricted to only STS or computing, also emphasizes using reason to reflectively understand and systematically critique the organization and aims of society [166]. This dissertation is situated where feminist, STS, and critical studies converge, endorsing that "critical STS places a greater emphasis on power and identity [than STS] and tends to evince a more visible commitment to specific values such as feminism, socialism, and anti-racism" [89].

The focus on social systems and institutions from feminist, STS, and critical studies have broadly shaped my approach to sociotechnical security and privacy, in prompting a deeper consideration of social norms that backdrop research decisions and interpretations. For example, a deeper consideration of patriarchy in the context of everyday security and privacy topics led to an exploration of gender stereotypes (Chapter 6; see related work on gender stereotypes in Section 6.2.). Also relevant is the influence critical theory had and continues to have on the field of demography [277] – such that there now exists the subfield of critical demography [429] – which is interested in the reflexive study of human social, economic, health-related, or other outcomes

[1]For a recent summary on science and technology studies (STS), see Section I in "The Scale and the Reactor" by Ryan Calo for law and technology scholars. [89]

based on demographics (see more related work in Part II, Section 5.2). Researchers in security and privacy are also interested in demographic analyses of security and privacy behaviors, e.g., [129, 507], which thus prompted a reconsideration based on critical demography (Chapter 5). The depth of feminist, STS, and critical studies is vast. I believe many of the concepts, frameworks, and theories from these fields will continue to be immensely inspirational for advancing security and privacy research.

Abuse and Trauma. Through decades of research and advocacy about addressing domestic violence and intimate partner violence, perspectives on interpersonal abuse have transformed it from being a private matter with little recourse for individuals to a significant public health issue [95]. Journals such as *Trauma, Violence, and Abuse*, *Journal of Interpersonal Violence*, and *Violence Against Women* are dedicated to publishing knowledge about abuse between people, commonly but not always as perpetrated by men against women. In 1982, Ellen Pence, Coral McDonnell, and Michael Paymar developed the Power and Control Wheel (see Figure 2.1), a conceptual tool that helped formalize interpersonal abuse as an ongoing pattern of violence, coercion, intimidation and abuse, i.e., of one person abusively wielding power and control against another [146]. This understanding of abuse predates the ubiquitous age of the internet, but all tactics described in the power and control wheel can now be facilitated through technology. Broadly, knowledge about (offline) abuse from criminology, gender studies, sociology, and psychology help inform an understanding of the causes, manifestations, and potential mitigations for technology-facilitated abuse. For more specific related work on image-based sexual abuse (IBSA), including the continuum of types, the role of technology, and the resultant harms, see Part I, Section 4.2.1; Part III, Section 8.2.1 also summarizes related work on synthetic IBSA specifically.

Research on abuse often aligns with core feminist beliefs, namely that abuse is a product of social systems:

Figure 2.1: The Power and Control Wheel, a visual representation of the primary tactics used in abuse in intimate relationships. The wheel was developed by social workers in conjunction women who had experienced abuse through support groups in Duluth, Minnesota [146].

> While a broad spectrum of people have experienced TFVA [technology-facilitated violence and abuse] across different categories of age, sex, race, ethnicity, ability, sexuality, or socioeconomic status, TFVA is not simply a collection of random acts of hostility and animosity. It is a product of existing intersecting layers of structural and systemic inequalities (Southern & Harmer, 2019), such as misogyny (Henry, Flynn, & Powell, 2020), homophobia (Green, 2019), transphobia (Colliver, Coyle, & Silvestri, 2019), racism (Kerrigan, 2019), colonialism (Carlson, 2019), and ableism (Hall, 2019), with some forms disproportionately affecting children and young people (Powell & Henry, 2019; Quayle & Koukopoulos, 2018). [32]

In this way, social perspectives present one possible resolution to the fragmentation problem: through a thorough and nuanced analysis of the relationship between social systems for security, privacy, and abuse. Indeed, studying the harm facing marginalized, vulnerable, and at-risk users would be incomplete without an account of how they came to be marginalized, vulnerable, or put at-risk, i.e., from those very social systems.

If abuse is the product of social systems, then so too are its effects. Social workers and social science researchers have also advanced the study of trauma, i.e., the result of harmful events that have "lasting adverse effects on the individual's functioning and mental, physical, social, emotional, or spiritual well-being" [504]. To address these severe impacts, social workers have developed trauma-informed approaches defined by six key principles: safety, peer support, trustworthiness and transparency, mutual collaboration, historical issues, and empowerment [505]. Trauma-informed approaches have been adapted for computing [106] and design [626], as well as social media [516] and HCI research with youth [474]. These socially informed insights about the origins and possible mitigations for trauma can run in opposition to rigid approaches in security and privacy that prioritize identifying and eliminating vulnerabilities in software or hardware.

For challenges like online abuse that remain pervasive in society, a nuanced understanding of trauma holds significant promise for addressing its sociotechnical complexities.

Part I

Evaluating Existing Support for Online Abuse

Chapter 3

"There's so much responsibility on users right now:" Expert Advice for Staying Safer From Hate and Harassment

Online hate and harassment poses a threat to the digital safety of people globally. In light of this risk, there is a need to equip as many people as possible with advice to stay safer online. We interviewed 24 experts to understand what threats and advice internet users should prioritize to prevent or mitigate harm. As part of this, we asked experts to evaluate 45 pieces of existing hate-and-harassment-specific digital-safety advice to understand why they felt advice was viable or not. We find that experts frequently had competing perspectives for which threats and advice they would prioritize. We synthesize sources of disagreement, while also highlighting the primary threats and advice where experts concurred. Our results inform immediate efforts to protect users from online hate and harassment, as well as more expansive socio-technical efforts to establish enduring safety.

This chapter originally appeared as the paper "'There's so much responsibility on users right

now:' Expert Advice for Staying Safer From Hate and Harassment" at the CHI Conference on Human Factors in Computing Systems in 2023 [597]. 'We' in this chapter refers to me and the co-authors: Sunny Consolvo, Patrick Gage Kelley, Tadayoshi Kohno, Franziska Roesner, and Kurt Thomas.

3.1 Introduction

Online hate and harassment is a threat with pernicious reach, negatively impacting the safety—e.g., emotional, sexual, or physical safety—of over 48% of internet users around the world [565]. While certain populations are at higher risk of experiencing targeted attacks—such as creators [566], journalists [105], gamers [316], survivors of intimate partner abuse [373, 204], and people with marginalized identities [509, 286, 113, 7, 170]—*anyone* can become a target of online hate and harassment. Going online today necessitates that internet users navigate a complex array of technology-mediated hate and harassment, such as toxic content, brigading (coordinated abusive behavior online), non-consensual sharing of intimate imagery, or device-enabled location surveillance [565]. As such, there is a need to prepare as many people as possible with appropriate knowledge and best practices for staying safer.

Advocates have published a wealth of resources to educate potential targets about protections for online hate and harassment. Non-governmental organizations (NGOs) like PEN America's *Online Harassment Field Manual* helps journalists and others in "navigating online abuse and tightening digital safety" [18]. Feminist Frequency's *Speak Up & Stay Safe(r): A Guide to Protecting Yourself From Online Harassment* is "designed for women, people of color, trans and genderqueer people, and everyone else whose existing oppressions are made worse by digital violence" [206]. Platforms also publish resources, such as YouTube's *Creator Safety Center*, which helps creators "make a plan to stay safe online" [622].

Advice and its framing ranges from general (e.g., broadly applicable) to tailored (e.g., highly specialized). Existing online advice for staying safer from hate and harassment tends to be tailored, such as for marginalized populations that are commonly targeted, or for common potential threats. Tailored advice is invaluable for populations that experience disproportionate risks, yet there is also an immense challenge to create and maintain unique advice for numerous disparate groups. There is comparatively little general advice for staying safer from hate and harassment, though general advice will be increasingly beneficial as more people experience hate and harassment. Such advice is a valuable addition to—not a replacement for—tailored advice. Particularly because many targets of hate and harassment may not predict being targeted or seek out advice, general advice establishes a consistent message for advice-givers to repeat at scale, hopefully reaching people before they experience attacks.

In this work, we explore developing general advice to stay safer from online hate and harassment, that is, advice that is broadly applicable and can be given without additional context about the user. We engage leading scholars and advocates to synthesize and evaluate the existing landscape of advice, including to identify frequently repeated advice that is not achievable, and to understand what experts believe would make such advice easier to adopt. We first gathered 219 disparate pieces of advice from existing guides, deduplicated them into 45 protective practices, and further categorized each by the threat it is intended to address. We focus on safety advice that can be implemented before hate and harassment occurs—i.e., prevention or mitigation—and scope advice narrowly to *proactive practices*. We then conducted interviews with 24 subject matter experts (based primarily in Western countries) who work with people experiencing online hate and harassment to assess three research questions: **RQ1: Informing user threat models.** Which online hate and harassment threats do experts believe most internet users should prioritize taking action to prevent or mitigate, and why?

RQ2: Prioritizing existing advice. For specific hate and harassment threats, how do experts prioritize existing advice for internet users who might experience them, and why?

RQ3: Recommending overall safety strategies. Assuming they do not have details about users' unique situations and there is no known ongoing attack, what are experts' top recommendations for internet users to stay safer from online hate and harassment?

Overall, experts felt that most internet users should focus their safety efforts on three of the seven categories of threats [565] we asked about: toxic content, content leakage, and surveillance. For some threats—such as account lockout and control, which didn't make the top three—there was a clear prioritization of advice: use two-factor authentication (2FA), use strong passwords, and to a lesser extent, use a password manager. Conversely, expert perspectives on how to mitigate content leakage or surveillance were far more discordant. Advice such as keep your camera covered, use anti-virus to detect spyware, or never share your location information with apps drew a range of perspectives. Towards overall safety strategies for minimizing harm, we find that experts recommended a mindset of data minimization, staying abreast of classic security advice, being self-aware and self-determined online, as well as participating in and fostering healthier online communities.

Our findings underscore a reality echoed by nearly every expert we spoke with: safety from online hate and harassment currently falls predominantly on users to enact. Experts judged that alleviating this burden would require pro-social, community-building approaches to increase safety for all. For advocates designing education materials, our work exposes the current state of generally applicable advice as well as multiple competing priorities that need to be considered when creating and delivering advice. For platform developers, our analysis surfaces gaps in protections and limitations of existing safety tools that lead to experts not recommending their use. And finally for users, our research provides a ranking of the most impactful existing advice

that can be enacted today.

3.2 Related Work

Experiences of Hate and Harassment. Hundreds of millions of people globally experience online hate and harassment [565, 391, 446], enduring serious physical, emotional, professional, relational, and financial harms [113, 509]. Prior research into online hate and harassment and protective practices is expansive. We rely on a taxonomy of experiences from Thomas et al. [565], which synthesizes the literature into seven categories of threats: toxic content (e.g., bullying, hate speech, trolling), content leakage (e.g., doxxing, non-consensual intimate images), overloading (e.g., brigading, dogpiling, denial of service), surveillance (e.g., stalking), false reporting, impersonation, and lockout and control (e.g., account takeover).

Online hate and harassment often builds on other axes of oppression. Harm tends to be disproportionately experienced by marginalized people, e.g., transgender people [509], women [113, 293, 502, 587, 286, 285], and Black and other marginalized racial or ethnic groups [293, 170, 286]. Attacks are more likely to be perpetuated by privileged groups such as men with a greater social dominance orientation [560]. Attacks may also narrowly target at-risk users in an attempt to silence voices—such as journalists, gamers, and creators [316, 105, 566, 542]—or coerce and control individuals as in intimate partner abuse [204, 205, 373]. The broad reach of online hate and harassment, and the reality that many individuals are unaware of the risks until they experience an attack, underscores the need to provide generally applicable advice for staying safe as a precursor to tailored advice.

Providing General Security Advice. Security advice should be effective, actionable, and understandable [485], as well as consistent and concise [46]. Unfortunately, the collective state

of security advice (not just for online hate and harassment) is far from concise, with experts offering hundreds of pieces of advice [487, 485]. Fragmentation means that users learn advice from different sources [482]—including stories [466, 447] or social "triggers" [150]—depending on skill levels and socioeconomic status [481], age [411], or other factors. Claims that advice is helpful are easy to make, but empirically impossible to refute [267], leading many researchers to call for prioritization [487, 46, 266, 289]. Security advice is often perceived to offer a poor cost-benefit tradeoff—high cost, low benefit—so motivation to follow advice is weak [265, 181]. To aid adoption, the delivery of advice should help people understand why the advice would benefit them [46, 266, 267]. We explore themes related to prioritization, cost tradeoffs, and delivery as part of our analysis of advice for staying safer from online hate and harassment.

Tailoring Security Advice. Significant research has also explored how to tailor support and security advice to at-risk groups, such as civil rights protesters [66, 588], employees [147, 148], human trafficking survivors [104], journalists [46, 386, 387, 385], older adults [411], politicians [126], queer individuals [212], refugees [533], and sex workers [379]. In tailoring advice for specific populations, these studies lie on the opposite end of a spectrum from the studies of general security advice described earlier. Advice could also be tailored by specific hate and harassment threats, but little academic research seems to have used that lens.

Though specialization enables more targeted support to groups that have been historically overlooked, it also enshrines criteria for additional support, i.e., group membership. For some groups, membership is evident or persistent (e.g., by identity, career), but may not be for the ever-increasing set of people who experience online hate and harassment. Some potential targets may not seek out tailored advice or even realize that they are at risk until after an attack is underway. Further, as the number of groups increases, creating and maintaining unique tailored advice becomes progressively difficult. To grapple with such difficulties, this work explores developing

general advice, absent specific user information.

Platform Safety Affordances. Almost all major platforms that allow user-generated content now explicitly prohibit hate and harassment [438], and they are continually building features to combat online hate and harassment. Automated features to reduce online hate and harassment include automated moderation of content [99, 537, 292, 312, 473, 82] or accounts [284, 496, 489]. In terms of manual efforts, platforms allow individuals [134] or authorized reporters [372] to report offending content (although the subsequent decisions can be seen as unfair or opaque [434]) or to implement crowdsourced blocklists [294, 213]. In particular situations, users or communities that are determined to be harmful have been deplatformed entirely [98, 276, 15]. Other efforts aim to provide peer support for users experiencing hate and harassment (e.g., Squadbox for email [359] and the Heartmob support community [2, 52]). In our work, we investigate experts' opinions of the current state of safety online, noting when they support advice recommending certain affordances, or when none exist to protect against certain attacks.

3.3 Methods

We interviewed 24 hate and harassment subject matter experts in July and August 2022 to discuss what advice might be generally applicable, that is, they would give to "general internet users" to stay safer from online hate and harassment. We use to this term throughout the remainder of this chapter to capture most internet users, irrespective of their risk level, as anyone can be targeted by online hate and harassment. As part of this, we also explored the complexities of providing safety advice in a general manner (i.e., not targeted to particular groups) and how to prioritize a large body of safety advice for users with limited time and resources.

3.3.1 Recruiting & Participants

We recruited subject matter experts—hereafter referred to as *experts*—who had a background in providing support to people experiencing online hate and harassment. Towards developing advice that would be general and widely applicable, we aimed to recruit participants who represented a diverse set of roles, populations assisted, and geographies. We made sure to recruit experts who had experience supporting marginalized populations. We identified 55 experts and organizations involved in the development of the advice guides we gathered (see Section 3.3.2), had publications related to hate and harassment safety practices, or were professional contacts. We directly solicited their participation via email; 24 participated in our study. Our 24 participants were academics[1] (n=12), NGO employees (7), and industry professionals (6).[2] Their specializations included social media (7), gaming (6), journalism (4), intimate partner abuse (3), online content creators (2), youth (1), activists (1), and attacker coordination (1). Participants' had two to 40 years of experience (average: 10 years, total: 237 years) in roles related to hate and harassment. Participants primarily operated in the U.S. (20), but also the U.K. (2), Australia (1), and Turkey (1), additionally speaking about France (1) and the Caribbean (1). We caution that no set of experts can comprehensively cover all people who experience online hate and harassment (e.g., all demographics, all occupations). We discuss this limitation further in Section 3.3.6.

3.3.2 Gathering Advice

Prior to conducting the interviews, we aggregated existing digital-safety advice related to hate and harassment. We gathered the advice from online searches, preliminary discussions with experts, and the domain knowledge of the authors of this work. We collected 49 online support

[1]Participants' academic departments included Computer Science, Journalism, Information Sciences, Public Policy, Criminology, and Human-Computer Interaction.

[2]Totals do not add up to 24 due to multiple roles.

resources, then filtered out those that did not address proactive practices (27), did not provide actionable advice (8), or only incidentally addressed hate and harassment (6).[3] Resources targeted audiences such as general internet users (e.g., OnlineSOS, Consumer Reports), social media users (e.g., Heartmob), journalists (e.g., PEN America), youth (e.g., Planned Parenthood), and more. Of the final set of 15 resources, five were tailored to specific at-risk populations, five to specific threats, and three to specific at-risk populations facing specific threats. Only two were not tailored (i.e., for anyone who might face hate and harassment online).

Across the support resources were 219 pieces of non-unique advice. Two researchers engaged in affinity diagramming to deduplicate advice and identify which of the seven categories of hate and harassment the advice best helped prevent or mitigate [565]. This effort resulted in 45 unique pieces of advice. As part of this process, we omitted advice about ongoing attacks (e.g., "deactivate accounts if you are being doxxed") or recovery, as our focus was on proactive practices.

As part of our interview protocol, we asked participants whether there was any additional advice they felt was missing. After applying the same scoping criteria as before and deduplicating advice, participants identified six "new" pieces of advice in total, demonstrating our approach achieved sufficient coverage of most advice. Of those six pieces, only one was mentioned by more than two experts. We discuss new advice in Section 3.4.2.

3.3.3 Study Procedures & Data Collected

Our semi-structured interview protocol consisted of four phases that were completed in a single, remote session with each participant.[4] First, we asked participants about their background in helping to protect people from online hate and harassment, as well as any specific populations they assisted.

[3]The complete list of advice guides that informed our work is included in the supplementary material.
[4]Our interview script is included in the supplementary material.

Second, we asked participants to rank which of the seven categories of hate and harassment threats general internet users should prioritize preventing or mitigating [565]. Given prior work emphasizing the need for minimalism and prioritization [485, 126], we developed this activity to require a discrete ordering. We asked experts to "think aloud" [100] while ranking to capture their underlying thought processes and opinions on each threat category.

Third, participants engaged in a card sorting activity, continuing to "think aloud," where they categorized the 45 pieces of advice into "High," "Medium," or "Low" priority, or advice they "Don't recommend." Rather than sorting all 45 pieces at once, this phase was broken into five parts, based on the seven categories of threats that each piece of advice was best positioned to prevent or mitigate.[5] The 5 parts were:

1. Lockout & Control – *9 pieces of advice to sort*,
2. Content Leakage – *13 pieces*,
3. Surveillance – *11 pieces*,
4. Toxic Content – *6 pieces*, and
5. Impersonation, Overloading, & False Reporting – *6 pieces*.

We decided on this approach during pilot testing. We found that it helped participants avoid over-indexing on the threat (which we captured in the second phase), and instead focus on the task of ranking individual pieces of advice. This partitioning also reduced the cognitive load of comparing 45 pieces of advice at once. After participants had sorted all advice in one threat category and if it had not yet been mentioned, we asked participants what, if any, advice was missing for that threat.

Lastly, we asked participants to enumerate the top three overall recommendations they would give to a general internet user to stay safer from online hate and harassment (which could be

[5] In the event an expert felt a piece of advice spanned multiple threats, we discussed with experts what implications that had for the advice and its priority to capture any missed nuance.

independent of the advice they ranked). We then engaged in an open discussion about the challenges of delivering advice; what, if any, existing advice guides they thought were effective; and ecosystem changes that might help shift the burden of staying safer from online hate and harassment away from users.

All interviews were led by the same researcher. They lasted from 63 to 97 minutes (average: 88 minutes). Each participant received a $100 USD gift card (or equivalent local currency) as a thank you. The amount was set by our institution for studies involving experts.

3.3.4 Analysis Approach

We used a mixed quantitative and qualitative approach to analyze our data, informed by how our knowledge and expertise is situated [240]. Our team's primary lens is security and privacy, with additional expertise in social media, online safety, and human-computer interaction. Our research and analysis focused on technical advice that users could follow to stay safer from online hate and harassment, which is only one of many approaches to digital safety.

From our semi-structured interviews, we gathered ordinal and count data about how experts ranked threats and pieces of advice. We quantitatively analyzed this data to produce an average ranking of the threats (RQ1) and proportions of how experts prioritized the advice (RQ2), as well as to inform the order of results subsections. To add qualitative depth, we applied thematic analysis to experts' open-ended responses to understand the factors that informed their threat prioritization (RQ1) and advice evaluation (RQ2), as well as generate themes from experts' top safety strategies (RQ3). We use thematic analysis [73], both inductively and deductively, because of its flexibility with respect to theory or goal, and its emphasis of researcher subjectivity as "analytic resource" for interpretation [71]. With a deductive approach, we referred to our own domain knowledge, as well as prior work, to direct our analysis of which factors informing threat

prioritization and advice evaluation we thought might be relevant (e.g., severity and agency [510], effectiveness and actionability [485]). To analyze factors that experts talked about as important, we used an inductive approach [564], and focused on the semantic (i.e., reflecting what experts explicitly said) as opposed to latent (i.e., experts' underlying assumptions) [73].

During interviews, a researcher who was not leading the interview took notes, focusing on capturing content. For analysis, notes were reformatted from per-interview to per-research question, i.e., threat ranking, advice prioritization, and overall top advice. One researcher read and re-read all responses, and developed a list of rationales (i.e., themes) that participants used to prioritize threats and evaluate advice, as well as categories of participants' top advice. We reviewed our ideas by revisiting the data, writing reflective memos, regularly meeting with members of the team, and iteratively updating the themes until we felt we had reached meaning sufficiency [73]. In the results, we report quotes (transcribed from interview recordings) to illuminate (a) instances where experts largely agreed, and/or (b) nuances on which experts disagreed, but were novel and insightful.

3.3.5 Ethics

Our study plan was reviewed by experts at our institution[6] in domains including ethics, human subjects research, policy, legal, security, privacy, and anti-abuse. We note that our institution does not require IRB approval, though we adhere to similarly strict standards. Prior to any data collection, all participants signed a consent form, which included agreement to record their session. At the start of each session, we re-confirmed consent (two participants requested that their sessions not be recorded, so they turned off their cameras and we only recorded audio and screens for the card sorting with their permission). We also reminded participants that their engagement

[6]This study was conducted at Google.

was entirely voluntary; they could pause, skip activities, or stop the session at any time and still receive the full thank you gift.

We protected our study data—including videos, audio, notes, and transcripts—by encrypting all records at-rest, restricting access to only the core research team (and institutional administrators), and requiring two-factor authentication with a physical security key to access the information. Video recordings, audio recordings, and transcripts were set to auto-delete after 6 months, though we kept some anonymized notes to be used in the publication process. Finally, we asked each participant whether they would like to be recognized in any acknowledgements or materials produced as part of the research. As a best practice, we attribute quotes only to a participant ID; we specifically omit unique details, phrases, or words from quotes to mitigate identification of participants.

3.3.6 Limitations

Given the breadth of digital-safety experiences, our evaluation of advice is non-exhaustive and limited to the 45 pieces of advice we identified prior to our study, and the 6 additional pieces of advice mentioned by participants. Our de-duplication of hundreds of pieces of similar advice may have resulted in omitting nuanced language that some experts viewed as important to the delivery. Many participants viewed advice through the lens of the populations they help protect (e.g., gamers, journalists, etc.), as well as through their geographic biases, highlighting the challenges of generalized safety advice in the absence of additional information about the person seeking help. Nevertheless, we reached meaning sufficiency [73] on the themes for how experts prioritized threats and evaluated advice before concluding our final interview.

General advice, compared to tailored advice, is unavoidably less accurate and thus might consider the wrong threats for some individuals. General advice might have limited benefits for

those experiencing extreme instances of hate and harassment and unnecessary costs for those who do not experience any. We were interested in exploring this limitation of general advice, so we asked experts how they would rank potential threats for a general audience. We report their rankings and thought processes in our results.

Relatedly, our use of the term "general internet user" in interviews may have introduced biases; most of our experts were in the U.S. where white men are assumed to be the default persona [399]. To combat these biases, we recruited experts with a range of perspectives and backgrounds, and also asked experts to explain who they imagined advice would or would not serve.

3.4 Results

Most experts agreed on three categories of hate and harassment threats that general users should prioritize taking action to prevent or mitigate: toxic content, content leakage, and surveillance. Experts commonly used three dimensions—severity, prevalence, and agency—as ranking criteria for evaluating the seven categories of threats (Section 3.4.1). Of the 45 pieces of advice experts were asked to rank, they most highly prioritized enabling two-factor authentication (Section 3.4.2). When ranking individual pieces of advice, experts weighed factors such as efficacy, ease of implementation, and effect on online participation. Experts' top overall advice recommended minimizing personal data online and developing an awareness of the unique threats that one might be targeted by, as well as taking pro-social actions to build safer online communities (Section 3.4.3). In this section, we discuss each of these findings in further detail.

3.4.1 Ranking Potential Threats

As part of the study, experts ranked which, if any, of seven categories of hate & harassment-related threats internet users should prioritize protecting themselves from, and why. In this section, we

Table 3.1: Ranking of hate and harassment threats. This includes overall average ranking (highest = 1, lowest = 7), the number of experts who ranked a threat as a top priority (maximum of 24), and the number of times experts ranked a threat as one of their top three priorities (maximum of 24).

Threat category	Average ranking	Top threat	Top 3 threats
Toxic Content	2.88	8	16
Content Leakage	2.92	7	14
Surveillance	3.33	5	12
Lockout & Control	3.96	3	12
Impersonation	4.25	1	8
False Reporting	4.96	0	7
Overloading	5.71	0	3

describe the criteria experts used to rank the categories, then review results for each category.

Ranking criteria. As shown in Table 3.1, experts were split on the foremost category of threat they thought internet users should prioritize. This was, in part, due to differences in the criteria 22 of our 24 experts used while ranking (two did not mention any criteria). Their ranking criteria included the *severity* of (potential) harms that might result from a threat, the *prevalence* of the threat (i.e., the likelihood of an attack occurring), and the *agency* of users to mitigate the threat.

For 10 experts, *severity* of (potential) harms was their primary criterion when ranking threats, and particularly threats to "physical safety, their bodily integrity, [as well as] to their mental health" (P22), echoing Scheuerman et al.'s *Framework of Severity* [509]. One expert favored this strategy because it allocated attention to those most in need of help:

> *"People who are targeted by the most severe forms of online hate and harassment are in marginalized communities and they need additional protections."* – P21

Nine experts relied on *prevalence* as their primary criterion for ranking threats. Experts expressed that this meant any guidance would better resonate with internet users, as it reflected attacks they were more likely to encounter. As P18 explained: "What is the most prevalent

problem right now... that people need to be aware of?" For other participants, prevalence reflected a disciplinary norm that stemmed from limited time and resources:

> *"In computer security, you want to educate people about attacks or threats they are likely to encounter. There are some attacks that are only relevant to government agencies, or high-profile organizations and so on."* – P1

Three experts used *agency* as their primary criterion for ranking. These experts remarked on the importance of building on user self-efficacy: "What is the lightest lift for a user?" (P23). These experts focused on which threats had the most meaningful existing protections, or where "a well-timed warning or educational intervention" (P20) might be effective.

The differences across our experts in the primary criterion—and even secondary and tertiary criteria—they used for ranking emphasize a challenge for protecting internet users from hate and harassment: there is no consensus yet for which problems to prioritize, or even *how* to prioritize them. While rankings may meaningfully differ for at-risk groups, many members of those groups may be unaware they are at-risk, or an event may suddenly put them at-risk [592]. General awareness of certain hate and harassment threats can thus provide critical, early protection before they are targeted. In this light, we explore which threats stood out more than others for experts, and where opinions diverged.

Toxic content. On average, toxic content—which includes bullying, hate speech, and sexual harassment—ranked as the highest priority threat across experts, often because of its prevalence. P15 noted that it was "the number one type of harassment that I see." Others added that toxic content could incur emotional harm and have "significant long-term repercussions" (P16), and that some users "might not even know that they are [experiencing it]" (P6), contributing to a greater need for users to prioritize learning what constitutes toxic content and taking proactive measures to prevent it.

Some experts ranked toxic content with lower priority, as—though it can cause harm—it "usually doesn't get to physical, severe harm" (P13) and because prevention is better handled at the community-level: "toxic content normalizes certain types of behavior, so it's a greater danger as a community norm than towards an individual" (P19). Others ranked it lower priority, saying that users had more agency:

> *"You can remove yourself from those situations either by logging out or by initiating or installing all of the protection features that a lot of online platforms have. It really sucks... [sending toxic content] is not okay—no one should do that—but you can remove yourself from those situations."* – P3

Content leakage. Content leakage—which includes doxxing and non-consensual sharing of intimate images—was ranked the second highest threat on average. Experts pointed to how common this threat is—"people send sexts all the time" (P10)—though often underestimated the risks, because people "really cannot imagine what it's like to be doxxed" (P21). The severity of content leakage, experts judged, arose because leakage is irreversible and attacks could easily spill over into users' "real lives, their experience of life outside" (P3) such as by facilitating stalking. Conversely, other experts rated content leakage a lower priority because it is less prevalent—"requires more work from the trolls" (P4)—or because users have less agency to prevent it:

> *"I can't think of any particular platform that really does an effective job of full control of [content leakage]... A lot of people have to escalate. So it's not just primarily relying on tools in the online space, but looking at resources that could help them seek justice offline."* – P24

Surveillance. Just five experts ranked surveillance—which includes stalking and monitoring accounts or devices—as the foremost threat in the context of hate and harassment, though it featured in 12 experts' top three. In general, experts felt surveillance was unlikely to be prevalent

and was "more context dependent" (P19). Though experts noted that it had the potential to cause severe harm (e.g., it can be a "high risk to physical safety"), P22 thought that people had more agency to prevent it (i.e., people "generally have more control and can find technical solutions").

Experts emphasized three contexts where this prioritization changed. The first was individuals experiencing intimate partner abuse, as surveillance "often begins before people realize they're in an abusive relationship" (P12), preceding the phases of abuse as identified in Matthews et al. [373]. The second was for people in civil society targeted by government-backed harassment and trolls: "one of the biggest digital issues [for journalists], [it] leads to physical threats and imprisonment, or assassination" (P4), and third, for prominent individuals [592] as attacks were "more relevant for popular accounts for people of a certain reputation" (P1). Experts broadly commented that incidents with surveillance could be exceptionally severe for targets:

> *"It's one of those thing where if it happens to you, it's going to have a significant impact emotionally and for your physical safety. In terms of long term consequences, it impacts how you interact in online spaces."* – P24

Lockout and control. Experts disagreed on how prevalent lockout and control—manipulating devices, being maliciously locked out of one's account—would be for an internet user specifically in the context of online hate and harassment.

However, many felt this was a more general security threat due to the prevalence of phishing and data breaches. For example, P8 noted that the "prevalence is high if you're vulnerable to a credential stuffing attack" while P17 ranked this threat the lowest because it is "not a primary way perpetrators attack people in the context of hate and harassment."

Regardless of the prevalence of this threat, experts remarked that being locked out of accounts and devices could facilitate other threats. Experts emphasized that targets "have to lock down [their] accounts and personal information first" (P14) in order to prevent down-stream harms,

such as content leakage or surveillance. In this way, experts prioritized account security as a locus of agency:

> *"[Lockout and control] strikes me as the most invasive. So anything where somebody feels like they don't have control over their own content to me, is the number one [priority]."* – P3

Impersonation. Only one expert ranked impersonation—fake profiles or communication posing as the target—as their foremost threat, commenting that it poses a "very immediate threat to personal information, devices, and can have a very large effect on someone's life" (P14). In terms of severity, experts agreed about the potential for impersonation to affect an individual's emotional well-being and reputation, as well as "collective harm on people in your network" (P24). Similar to surveillance, experts noted the low prevalence for most internet users, though it could be higher priority for prominent figures.

Impersonation was seen as harder to prepare for, or even not preventable at all. One expert pointed out the precarity of people who have begun to gain public followings, but may not have all the resources of more prominent public figures:

> *"The place I see impersonation happen a lot is with low-level influencers... they're less likely to know it; they won't have a [support] team."* – P21

Some experts spoke to the challenges of recovering from impersonation: that marginalized people are harmed the most because there are "not a lot of tools or legal protections" (P19) for them, and that it was a "pain in the butt to get platforms to respond to impersonation reports and get them taken down" (P23). One expert with personal experience assisting targets of harassment seemed more optimistic about recovery, saying that in their experience, it "usually turns out more alright than other situations" (P10).

False reporting. No expert in our study ranked false reporting—such as swatting or false abusive

account reporting—as the top threat for internet users, though seven put it in their top three. Experts viewed false reporting as a very rare occurrence, though they noted that it was more common on gaming platforms and among "big armies of trolls" used by "authoritarian regimes" (P4).

Experts noted the severity of harms stemming from false reporting could be extremely divergent or unpredictable. P6 shared that false reporting was a "standard bullying tactic" employed by kids—one that might not lead to consequences for those employing it or to those targeted by it (though it would slow triaging legitimate complaints). On the other hand, P20 spoke about how swatting could cause extremely severe harm, including being fatal. The viability of false reporting as a tactic, and thus agency of users to act, largely fell to the review process of the emergency service or platform contacted, which could be complicated by limited resources:

> *"The claim is usually that the content they have, the video they've shared, or the post is of a 'sexual nature.' And it doesn't contain any of it. But because it's in a foreign language that isn't supported by the platform, it's taken down immediately."* – P15

Overloading. Just three experts ranked overloading—including brigading, notification bombing, or denial of service attacks—in their top three threats; similar to false reporting, none ranked it as the top threat. Most experts commented that while overloading could be frustrating, it has a low prevalence of occurring for most internet users (notable exceptions are those with high profile accounts or websites). For notification-based or network-based attacks, experts felt such attacks were low severity: "it's not necessarily going to affect your psyche or your personal well-being" (P4) and "annoying but not as important" (P5). Experts expressed that overwhelming volumes of potentially toxic comments could be far more severe:

> *"For an individual to get piled on... that was one of the primary tools that Gamergate used to harm their targets. It was very harmful, the scale of the harm, in addition to the*

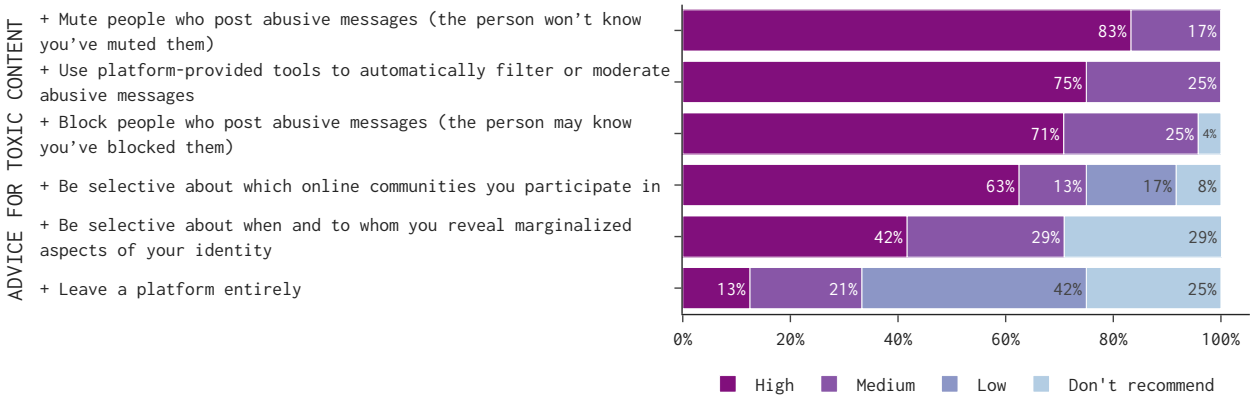

Figure 3.1: Ranking of advice that users could employ to help prevent toxic content. Experts favored all forms of platform-provided moderation tools over advice that curtailed online participation.

toxicity." – P19

3.4.2 Prioritizing Current Advice

Experts ranked each of the 45 pieces of advice we collected as "high," "medium," or "low" priority, or advice they "don't recommend." In reasoning aloud, experts weighed factors such as efficacy, ease of implementation (and the existence of appropriate tooling), and whether advice curtailed a user's participation online. In this section, we review advice for staying safer from each threat, ordered by the average ranking of each threat from the prior section. We highlight only the advice that experts ranked highly, or where experts felt challenges persist or alternative solutions are needed. The complete set of advice is shown in Figures 3.1–3.7.[7]

Preventing toxic content: Agreement about muting and blocking, but challenges around curtailing personal expression. To combat toxic content, experts favored platform-assisted moderation, with 83% highly prioritizing *mute people who post abusive messages* and 71% *block people who post abusive messages* (Figure 3.1). Experts prioritized muting over blocking because blocking is more visible to attackers, who might escalate attacks when they find out they have

[7]A unified, ranked list of all advice is included in the supplementary material.

been blocked. Additionally, blocking impedes potential targets from monitoring their attackers:

> *"[Targets] don't want to read misogynist or racist comments, but they need to know that certain conversations exist, or whether they face threats. So they want to mute."* – P4

Muting allows a target to quietly filter offensive users they encounter online (e.g., community members), whereas "blocking sends a signal you no longer want to interact" (P24). As such, experts noted that being aware of and being quick to use these features could curb future harm, in addition to their conventional use when there is an active attacker.

When asked if any advice to help prevent toxic content was missing, 13 experts said that reporting hate and harassment should be included,[8] grouping it with blocking or muting as a standard best practice. Experts recommended reporting to the platform as well as to civil society organizations that can organize multiple reports, noting that reporting was a primary mechanism for platforms to find new issues and make improvements. At the same time, experts lamented that "reporting doesn't have an immediate impact" (P16) and could be detrimental emotionally if the platform ultimately determined the reported attack did not cross a policy line:

> *"It's more harmful for the person [who submitted the report] to get a message that this wasn't even [determined to be] harmful."* – P24

While experts broadly agreed on the high prioritization of advice for mitigating toxic content, advice that required a user to limit their participation online was far more contentious, even when it was considered to be effective at preventing an attack. Of experts, 63% highly prioritized *be selective about which online communities you participate in* and just 42% *be selective about when and to whom you reveal marginalized aspects of your identity*, while 29% of experts did not recommend the latter at all. Among experts who rated either highly, a common refrain was being aware of

[8]During our advice gathering, we came across reporting, but at the time, we regarded it as not being proactive and thus out of scope for this study. However, we include it here because so many experts mentioned the importance of being aware of this feature. Additionally, reporting, like blocking and muting, are features general internet users should be aware of in advance, so they are prepared if or when attacks occur.

unsafe communities and what you share as part of dealing with the realities of hate and harassment today:

> *"As a user, you should be able to decide... where you feel comfortable the most. If you don't feel comfortable on say, [platform], because a) you're not sharing that much and b) you're getting a lot of information pollution, or you don't find it useful at all, it makes sense to be selective."* – P15

> *"Heartbreaking. The whole idea of not being able to bring your whole self to an experience... Sadly I would always give that advice for today. I hope it's not advice I need to give in the future."* – P20

Experts who were opposed expressed concerns that such advice required more nuance than was possible for a general guide. Others felt such recommendations gave up the ability to participate freely:

> *"I understand the practical reasons behind it, but philosophically it's not right to expect people to do that... I've been doing stuff with [platform type], and there's this general philosophy we're trying to disrupt: 'If you don't like it you can go somewhere else.' I don't like that sensibility being recommended from the top down."* – P10

The most contentious advice for combating toxic content was *leave a platform entirely*. Only 13% of experts ranked it highly, while 67% put it as low priority or not recommended. Experts in support highlighted it could be appropriate as a last resort:

> *"It's always a tradeoff between having fun and not receiving too much harm... It's not the first thing you should do to deal with harm, you should try other things first. But if the harm is too pervasive and this is the only way to prevent it, they should."* – P13

However, most experts opposed this advice due to losing voices of people targeted by hate and harassment, or the quality of life for following it:

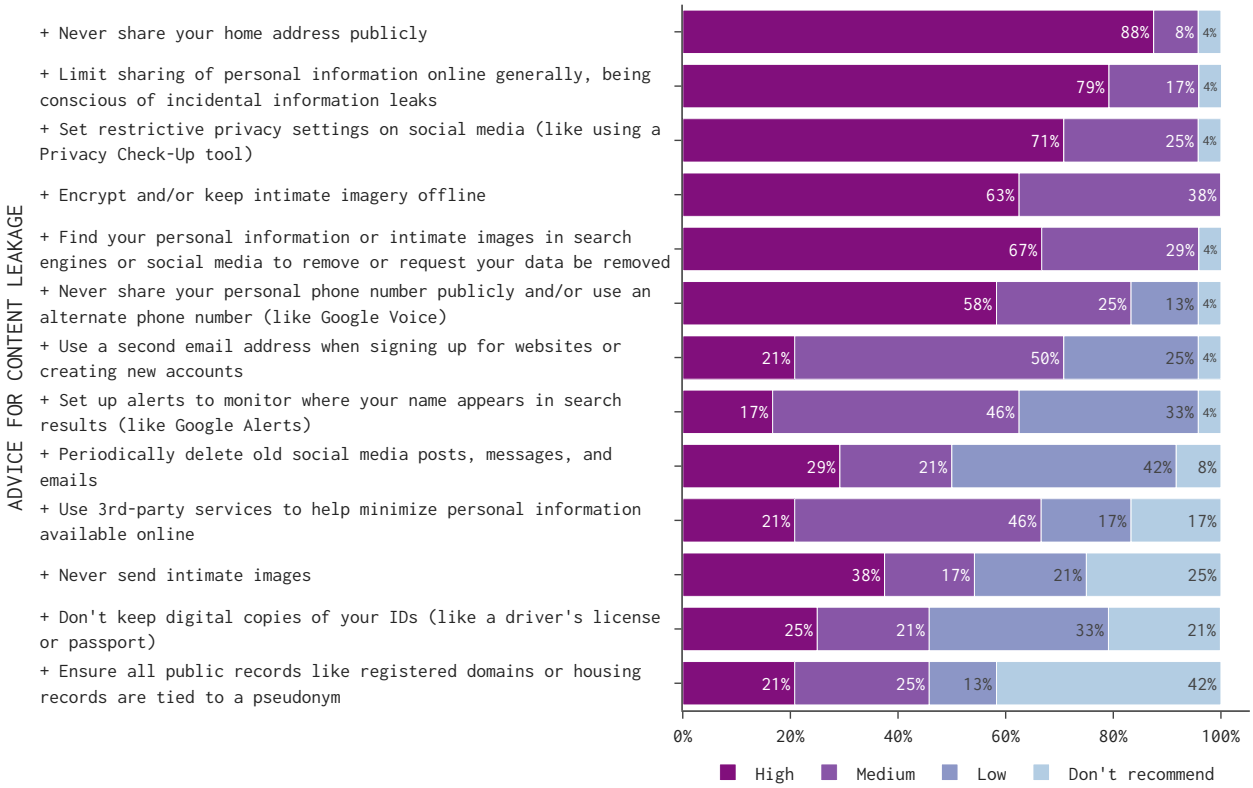

Figure 3.2: Ranking of advice users could employ to help prevent content leakage. Experts prioritized advice involving data minimization involving one's address, phone numbers, and personal information.

> *"Just imagining the life of a perfectly secure user is really depressing. Is that really a life at all?"* – P10

Experts recommended an alternative: taking a break or turning off notifications in order to disconnect. Broadly, advice for combating toxic content was more sparse compared to other threats we discuss. However, it was also one of the few threats with protections built-in to most platforms today.

Preventing content leakage: Agreement about the need to restrict information that's publicly available, but challenges with the ease of implementation and curtailing personal expression. To combat content leakage, experts recommended that individuals focus on restricting what information they share (Figure 3.2). 88% of experts highly prioritized *never share*

your home address publicly and 79% highly prioritized *limit sharing of personal information online generally, being conscious of incidental information leaks*, reasoning that "the more information that's out there, the more potential for leakage" (P11). For other highly recommended advice, such as *set restrictive privacy settings on social media (like using a Privacy Check-Up tool)*, experts believed user awareness to be low: P3 commented that "most people don't know they can change their settings."

Though restricting information sharing was perceived as effective, experts discussed challenges with a cluster of advice that would be effortful to implement. For example, 58% of experts highly prioritized not sharing personal phone numbers, but P6 noted that people might do so accidentally—"maybe you didn't intend to share it publicly but it's attached to a review or something." Similarly, only 25% of experts reported that not keeping digital copies of IDs was a high priority, because digital copies of IDs are becoming very common and sometimes obligatory (e.g., vaccination records to help manage the COVID-19 pandemic). Other pieces of advice that experts thought could be helpful but would require excessive effort for a general internet user included using a second email address for accounts, using third party services to remove information online (e.g., DeleteMe), or ensuring that public records like domain name registration or housing records are tied to a pseudonym.

Experts were very divided whether *never send intimate images* should be recommended to prevent content leakage: 38% prioritized it highly, 38% prioritized it as medium or low, and 25% would not recommend it. Some experts noted that never sharing would be highly effective—"that's one of the easy ones" (P12)—while other experts considered the advice to be victim blaming:

> *"If people want to share intimate images, technology should support their ability to do so."*
> – P8

To sidestep issues of personal digital expression, experts were in greater agreement that people

should *encrypt and/or keep intimate imagery offline*, as 63% highly prioritized doing so. Experts emphasized the offline part most—"don't use cloud storage" (P7), "prefer offline to encrypted" (P3)—but mentioned "there are a lot of tools now to keep these under lock and key" (P24). Experts also recommended other tips for sending intimate images more safely, such as only sending them to highly trusted people, or ensuring the images do not include identifying details such as one's face or tattoos.

Another challenge that experts noted for preventing content leakage was that certain pieces of advice would be relevant only for a subset of users. Only 17% of experts advised general users to *set up alerts to monitor where your name appears in search results (like Google Alerts)*:

> *"Only if you have some higher risk factor. Are you a streamer, or do you work in an industry where you deal with the public in a way that you are more likely to encounter harassment? Working at [a high profile company], this was a huge concern of mine."* – P20

Other experts added that alerts were also only useful for people with unique names, and cautioned that alerts would lead to frequent false alarms for people with common names.

Similarly, experts judged that reviewing old content was only worth the effort for certain groups:

> *"People will go after you if you are a journalist and write about sensitive topics like politics or extremism. So they will search for what you wrote as a student from 10 years ago, which you may have forgotten about."* – P4

67% of experts considered *find your personal information or intimate images in search engines or social media sites to remove or request your data be removed* high priority to do once in a while, though P6 cautioned that overemphasizing this advice "can make people really paranoid" and "only gives this advice if there is a reason, like someone saw a picture of you online or you have

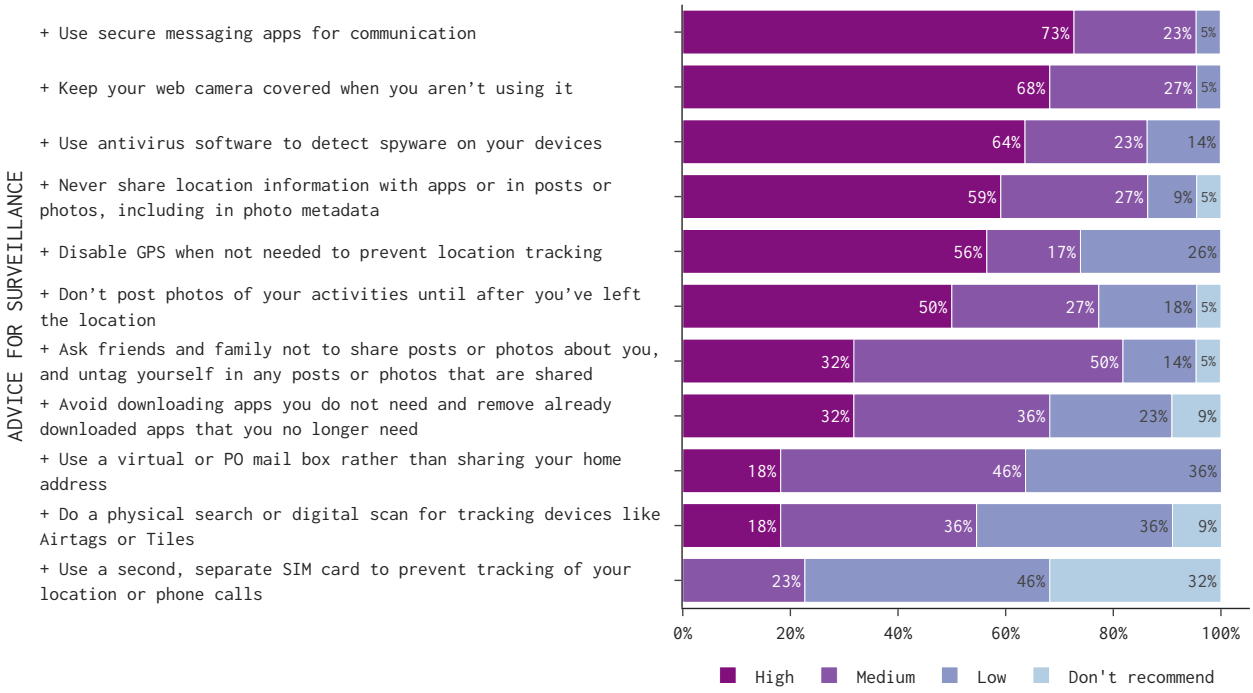

Figure 3.3: Ranking of advice users can employ to help protect themselves from surveillance. Experts prioritized making use of privacy tooling and limiting usage of certain application features.

an abusive ex."

Preventing surveillance: Agreement about the usage of privacy tools, but challenges around effectiveness and ease of implementation. High priority advice for surveillance focused primarily on using strong privacy tools, or limiting certain application features that might leak one's location or identity (Figure 3.3). However, experts' evaluation of advice surfaced challenges about whether advice would be effective in mitigating a surveillance threat such as stalking.

73% of experts highly prioritized *use secure messaging apps for communication*, but multiple experts viewed secure messaging more through a lens of general security threats, rather than hate and harassment. For example, P16, who ranked the advice as high priority, explained: "I do recommend [secure messaging] to people, maybe not in this [hate and harassment] context,

but I generally do." Other highly ranked advice for mitigating surveillance via compromised devices was also more protective against general threats, and less aligned to surveillance for hate and harassment. Advice such as *keep your web camera covered when you aren't using it* and *use antivirus software to detect spyware on your devices* were highly prioritized by 68% and 64% of experts respectively, as they were seen as supporting user agency—they are simple steps that could provide some protection: "no harm in doing it, but I wouldn't say you need to go home tonight and cover every web camera" (P14). Yet, P8 clarified that cameras were only a superficial concern for surveillance and ranked this as low priority:

> *"[You're] not dealing with the root cause. If you're worried about your web camera, [you] should be worried about bad software in general on your device."* – P8

Thus, despite experts finding some advice in this section high priority, there remains room for new advice and protections that would more effectively protective against surveillance.

Experts were generally not in favor of other more strict physical access measures such as *use a virtual or PO mail box rather than sharing your home address*, *do a physical search or digital scan for tracking devices like Airtags or Tiles*, or *use a second, separate SIM card to prevent tracking of your location or phone calls* due to the substantial effort of implementing the advice. Experts felt this advice "really depends on your threat model" (P9) and expressed that they were "not sure creating an atmosphere of anxiety is needed" (P20) for general internet users. However, experts noted that in some contexts, these practices became critical:

> *"If you are running from an abusive spouse, then absolutely... But I wouldn't recommend everyone in the world do this. "* – P11

Experts also warned of the challenges of enacting this advice successfully. Searching for physical tracking devices is "really difficult to do... people don't know how to do a digital scan" (P12) and "may not be possible for people who aren't well versed" (P15), echoing Gallardo et al.'s findings

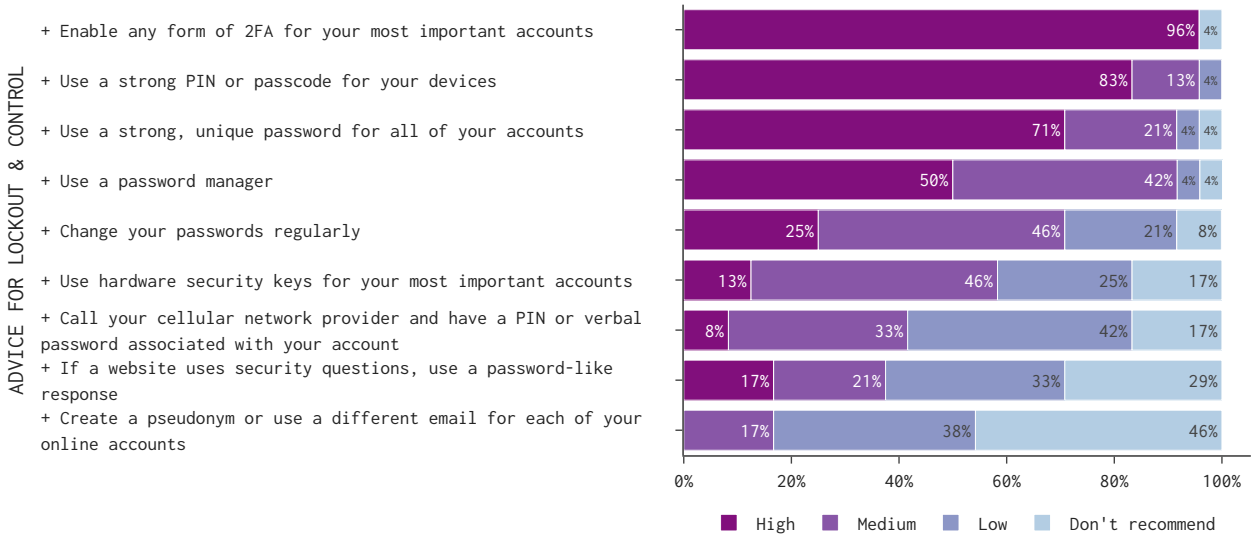

Figure 3.4: Ranking of advice users could employ to help prevent lockout and control. Experts limited their advice to proven account security best practices.

that detecting surveillance issues is difficult [207]. Likewise, "it's a lot of work to get a P.O. box for all deliveries. It's inconvenient for real life" (P12). As a whole, experts felt this advice was best suited to people who knew they were in a surveillance situation, but not something that general internet users needed to be concerned about.

Preventing lockout and control: Agreement about establishing account hygiene, but challenges with the ease of implementation. To protect against account-based threats, experts overwhelmingly favored protections they considered to be basic account hygiene (Figure 3.4). 96% of experts highly prioritized *enable any form of 2FA for your most important accounts*, as did 83% *use a strong PIN or passcode for your devices*, and 74% *use a strong, unique password for all of your accounts*. As P16 explained regarding 2FA:

> *"If you are actually worried about people hacking [your account], a password isn't enough."*
>
> – P16

Experts also discussed how 2FA alleviates the need for users to change passwords regularly, noting the reality that many users do not use strong *or* unique passwords. Experts also noted that users

are becoming more familiar with it and finding it "less horrible" [123] than they expected. Only one expert did not recommend 2FA because "people get locked out of basic services often" (P12).

Favorability of 2FA stopped short of hardware keys (as opposed to SMS or on-device prompts), with just 13% of experts stating hardware keys were a high priority, mainly because it was unnecessarily burdensome for general users. P16 felt this level of security was only needed "if you have the nuclear codes" while others stated this was more important if you had business secrets or professional accounts that might be targeted.

The effort necessary to protect against attackers exploiting weak security questions or having multiple accounts to avoid a single source of failure was also viewed as too onerous. Of experts, 62% rated *if a website uses security questions ... use a password-like response* and 84% rated *create a pseudonym or use a different email for each of your online accounts* as low priority or not recommended. For hardening security responses, experts were concerned primarily with users forgetting responses. For managing multiple accounts, experts felt the credentials would be too much to remember:

> *"How are you going to keep track? ...we've all got at least 10 or 20 different accounts."*
> – P11

When asked about any missing advice, experts added four pieces for helping prevent lockout and control: keeping account recovery vectors up-to-date (mentioned by 2 experts), checking whether passwords have been exposed by a breach (2), never sharing passwords (1), and keeping an eye out for notifications of suspicious account logins (1).

Preventing impersonation: Lack of effective advice. Across experts, there was no existing advice—nor any advice they could provide—that a consensus felt was high priority to help prevent impersonation (Figure 3.5). Advice such as *ask friends, family, and colleagues to help keep an eye out for impersonation* were ranked as both high and low priority by 35% of experts. As a proactive

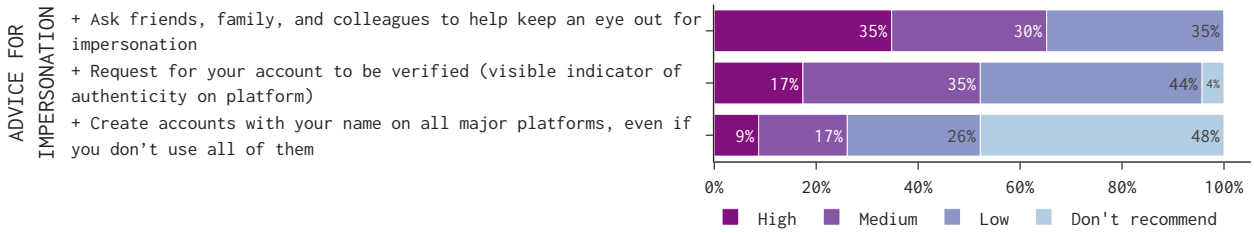

Figure 3.5: Ranking of advice users could employ to help prevent impersonation, none of which experts felt was effective for general internet users.

practice, most experts viewed this as too "paranoid," particularly in light of the low prevalence of impersonation in their experiences. Similarly, experts raised concerns about feasibility. As P4 put it:

> *"Do you really think your friends and family and colleagues will spend the time to look out for impersonation for you? They don't care. They have so many things to do."* – P4

Experts felt this advice was more pertinent when responding to an active or previous attack (i.e., if someone has been or is being impersonated):

> *"If you were being targeted, you should do this. But not if you didn't have reason to believe you were being targeted."* – P14

Experts also deemed other forms of bolstering one's digital identity as infeasible or ineffective: 48% ranked *request for your account to be verified* as low priority or not recommended, while the same was true for 74% of experts when ranking *create accounts with your name on all major platforms*. Verification (e.g., a visual indicator of trust available on many social media platforms) was perceived as restricted by platforms to celebrity-like individuals who had a sufficiently large audience, and thus beyond the capabilities of most internet users.[9] Likewise, managing multiple accounts that a user wasn't planning to actively use was viewed as burdensome and potentially even harmful due to compounding account security risks (e.g., the reality that many users would

[9]Our interviews were conducted several months before the December 2022 roll out of *Twitter Blue*.

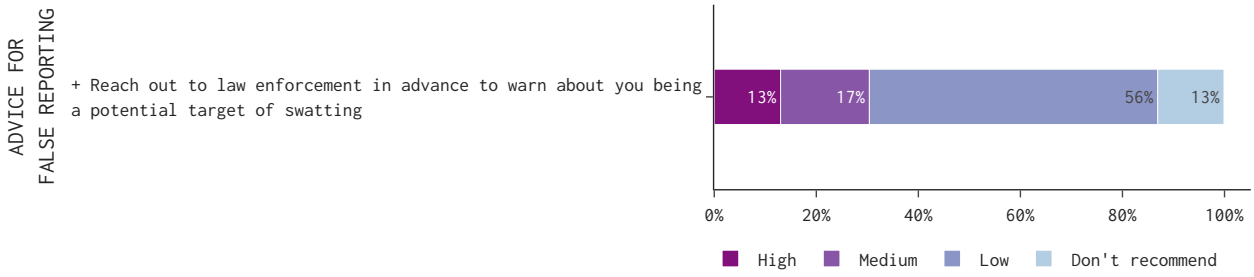

Figure 3.6: Ranking of advice users could employ to help prevent false reporting. Experts viewed swatting as outside the scope of general internet user threat models. Likewise, law enforcement might not be equipped to handle warnings.

likely use weak passwords).

> *"I don't recommend that at all. That's basically saying you need to sign up for everything... If you don't have good password hygiene and use the same password on all of them, you can be compromised faster."* – P9

The lack of advice for impersonation stems, in part, from the challenge that attacks frequently occur without a target's knowledge, and often on platforms where the target is not a participant (e.g., fake dating profiles, fake social media accounts).

Preventing false reporting: Lack of effective advice. When gathering existing advice, the only advice we found to combat false reporting was to *reach out to law enforcement in advance to warn about you being a potential target of swatting* (Figure 3.6). A majority of experts—69%—ranked this as either low priority or not recommended, most commonly because of the low prevalence of swatting on general internet users:

> *"If you're likely to get swatted, then it's a high priority. If you're just a regular person and you did this, the police would think you're crazy ... In the general case, you shouldn't even think about [being swatted]."* – P1

Other concerns focused on the perceived indifference of law enforcement, a lack of law enforcement training on how to handle such warnings, or a general distrust of law enforcement (particularly

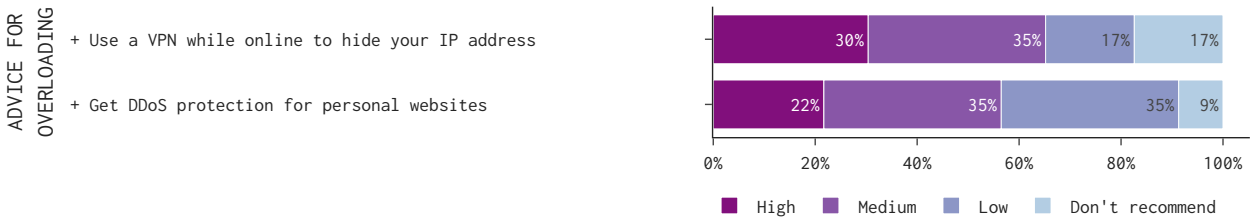

Figure 3.7: Ranking of advice users could employ to prevent overloading. While VPN and DDoS protection services exist, experts felt they were too cumbersome or out-of-scope for most hate and harassment that general internet users would experience.

in authoritarian regions):

> *"This one is complicated. A lot of times law enforcement isn't well set up to do anything with this information. Maybe a good idea, but it's contingent on where you are in the world."* – P20

While swatting is the most severe form of false reporting in terms of physical harm, there remains a lack of helpful advice for attacks that attempt to silence a target by having their account terminated. Such attacks depend entirely on the procedures and practices of third-party platforms, which targets can only partially navigate by choosing where they participate.

Preventing overloading: Lack of effective advice. While overloading encompasses multiple threats—such as notification bombing, brigading, or dogpiling—existing online advice we found was limited solely to network security (Figure 3.7). For *use a VPN while online to hide your IP address*, there was a large spread of prioritization among experts. For P15, this was a "general thing that everyone should be doing," whereas for P8, this advice was "pretty in the weeds and not relevant to most, but if you're targeted, could be reasonable." Other concerns included barriers to access, usability concerns around proper configuration, and misconceptions about what protections VPNs provide (as recent work has also explored [49, 468, 9]).

Similarly, *get DDoS protection for personal websites* was prioritized as either medium or low by 70% of experts. P22 felt it was a "no brainer, but not easy," whereas most experts felt this

advice should be restricted to people who had personal websites with a higher likelihood of being targeted.

The lack of guidance for brigaiding or dogpiling—such as when a person goes viral outside their intended audience—exposes a critical gap in advice today for general internet users. This is particularly problematic as these attacks occur spontaneously, limiting the window for a target to react, or to control the spread of their content once its shared beyond spheres where they have platform-provided privacy controls.

3.4.3 Overall Safety Strategies

When we asked experts to describe their personal top three recommendations for general internet users with respect to online hate and harassment, we received responses that varied greatly in specificity. Some experts named discrete actions, such as pieces of advice from Section 3.4.2, while others spoke broadly about things users should keep in mind. We synthesize the 65 top recommendations of the experts we interviewed below.[10]

Data Minimization (recommended 24 times). Across all experts, the most common top recommendation was to minimize sharing personal information. Experts spoke about the importance of reducing the amount of personal information that is available online, both by being mindful of what a user shares, as well as deleting existing data that is already online. However, experts were also cautious about recommending that people limit what they share online noting that it "may not eliminate the potential for things to happen" (P23). Going further, P23 explained that data minimization is not a sustainable solution:

"Putting limits on self-expression may keep you safe in the short term but it's not good

[10]Most, but not all, experts gave top recommendations. One expert passed on giving any top recommendations, explaining that one-size-fits-all advice did not exist. Some experts combined multiple recommendations, so counts do not sum to 65.

for the health of online spaces overall." – P23

Echoing this concern, P8 reasoned that the framing of the advice would be crucial:

> *"Being careful about what you put online is always a reasonable thing to suggest to people. It is a little victim-blaming at the end of the day, right? So it has to be worded appropriately, but certainly good advice."* – P8

In addition to limiting sharing, experts favored auditing security and privacy settings, especially for social media accounts or location tracking. P24 noted that it was important to consider how information is presented online, and making sure that users know who content is visible to. Privacy and security settings, similar to limiting information available, were seen by experts as actions where users had agency, which may be why they were the most common pieces of top advice. Further, these recommendations align with our finding that content leakage was, on average, the second most important hate and harassment threat that experts thought general users should be concerned with (see Section 3.4.1).

Account Security (recommended 18 times). Experts frequently recommended general account security practices, including using 2FA, creating strong and/or unique passwords, and using a password manager. P3 described these tips as putting yourself on the path of least resistance:

> *"You don't have to set up the most complicated security system you can think of. Do things that will slightly deter you from having a bad experience online compared to the general public."* – P3

Self-Determination and Awareness (recommended 17 times). Experts believed that users should determine for themselves *where* they choose to engage online:

> *"Consider the community you're engaging in and its culture... if you're going to be on 4chan, you're going to get hateful content... so it's better to start off in more protected, smaller, or closed communities with better norms."* – P2

By being more aware of the community norms, as well as the potential protections afforded by certain platforms, experts reasoned that users could better avoid harm. Experts also recommended that users pay attention to *how long* to engage online, or in P2's words, "decide for yourself how much bullying or harassment you're willing to endure." By determining how much abuse an individual is willing to tolerate, experts reasoned that users could decide when to "leave the platform, especially if it's continuous and targeted – the platform isn't for you" (P11) or at least temporarily "remove yourself from any situation from which you feel unsafe" (P20).

In a similar vein, experts recommended that users stay aware of how they might be threatened, and what existing tools could help. Searching for yourself online was seen as a good way to "be aware in general of your digital footprint or online presence" (P15). Given that threat modeling is a standard practice in security for enumerating threats, two experts explicitly recommended it, and one expert implicitly: "Think deeply about who has access to your devices and how you keep those secure" (P24).

Safer Through Community (recommended 9 times). The final strategies recommended by experts were communally-focused. Experts recommended reporting hateful or harassing content—"my favorite is still: block aggressively" (P7)—not only for immediate individual relief, but also because doing so would ultimately help foster safer online communities.

> *"Don't be a silent bystander... we're not going to create a better world by being silent about it. Use the tools you've got. If you can report, report. If you can stand up for folks, stand up for folks... So it's not just about protecting yourself, it's about being a good digital citizen. It's important because if you're waiting for others to change, there won't be change."* – P18

Other experts further supported the need for pro-social behaviors that would improve broader online communities by proactively looking out for others, as well as sharing the responsibility for

creating healthier online environments. If users do experience harm, one expert recommended reaching out for help from trusted parties. P13 hoped people who have been targeted would understand that:

> *"It's not your fault. As long as we expose ourselves online, there are dangers that we face. Many times, survivors blame themselves for it. They aren't sure whether it's harm or if they're overreacting. Or they think that they did something wrong so they should be blamed for receiving harassment. The internet environment can be toxic sometimes, and platforms may have given you limited tools to address the harassment, so you feel like you have less agency, but it's not your fault. We should acknowledge that others have responsibility to protect them."* – P13

3.5 Discussion

In this work, we sought to find generally applicable advice that would contribute to individual safety from online hate and harassment without additional context about the user. From an interview study with subject matter experts, we outlined a cluster of top threats they believe users should prioritize and advice users can employ to help prevent those threats, as well as overall safety strategies. We now step back to discuss tensions our work surfaces for efforts to help people stay safer from online hate and harassment.

Our work illustrates the complementary roles of general and tailored advice. Though our aim was to explore general advice, the current landscape of online hate and harassment makes both general and tailored advice valuable, given the unique benefits and limitations of each.

Most prior hate and harassment safety advice—including the advice we collected for our work—takes a tailored approach. Tailored advice centers marginalized populations that are at disproportionate risk for online hate and harassment, providing invaluable support to those who

may need it the most. Yet, tailored advice is extremely challenging to create and maintain. Experts in our study who served as advocates for specific populations expressed that existing resources were insufficient, despite not even serving all groups that need support. Further, groups needing tailored support may not know such resources exist or how to find them. Therefore, tailored advice is best for users who understand that they are at a disproportionate risk and helps them focus their effort where it will be most effective.

Contrasting tailored approaches, prior work on traditional security and privacy advice has called to "identify the smallest and most easily actionable set of behaviors to provide the maximum user protection" [485]. In some contexts, such as when advice-givers do not have more detailed information about users' situations or when users do not wish to reveal sensitive information about their situation, general advice is the only viable option. General advice empowers individuals to adopt effective safety practices with lasting consequences even before they are at risk or become aware of tailored advice for their situation. Users are also more likely to follow general advice that multiple sources consistently repeat, though such advice approximates an average threat level and can under- or over-prepare potential targets. Therefore, general advice is best viewed as a baseline of protection for a wide range of users, and as a stepping stone towards tailored advice.

Throughout this work, we grappled with the need for advice that would be relevant for an ever-increasing proportion of internet users who will face online hate and harassment and the heterogeneous experiences that each user will have. Both general and tailored advice can have a valuable role in supporting potential targets. Our study further shows that general advice rarely contradicts tailored advice; instead, general advice is best for when less information about users is available, and tailored for when more information is.

The lack of consensus on top threats poses a challenge for which education and safety tools advocates should focus limited resources towards developing. In the absence of

contextual information about a person's unique needs, experts only loosely agreed on which threats general internet users should prioritize preventing or mitigating. Part of this complexity stemmed from the three competing dimensions that experts used to rank threats: severity of harm from the threat, prevalence of the threat, and agency that users have to combat the threat. For example, some experts who had experience supporting targets of intimate partner abuse were especially attuned to the *severity* of threats posed by targets' intimate partners, ranking lockout and control as well as impersonation threats higher than other experts. But some experts who supported journalists or content creators whose jobs necessitated they have a prominent online presence were particularly attuned to *prevalent* forms of online hate and harassment, tending to rank toxic content higher. Our interviews did not indicate a clear path for resolving these tensions, or if such a path even exists. Experts also leaned on their considerable, deep experience for the particular populations they served, which do not represent all people who experience hate and harassment. A remaining question for future work is: how might research and practice deliver relevant advice to people's unique risk profiles *at scale*, especially if particular at-risk groups are not yet understood?

While better empirical measurement may assist arriving at a consensus and thus how to best allocate resources, some applications might necessitate prioritizing one dimension over others. Companies with broad user bases might focus on prevalence, acknowledging that severity and agency fall to other actors. Specialized support providers, such as for survivors of intimate partner abuse, might center their efforts on high-severity threats. Taken together, these efforts would aim to communally balance the needs of specific groups that are at heightened risk for specific types of hate and harassment, while also considering some other users may never face such risks. The multiplicitous approach also addresses a caution from prior work "against using worst-case scenarios when average-case is what users care about" [266]. The average case of hate and harassment is not yet known and could very well change over time. Further, the nature of hate

and harassment incidents does not allow for clean distinctions between "average" and "worst."

Effective advice requires letting a user make their own decision, at the right moment. Many experts emphasized that how and when advice is offered is challenging, if not more so, than developing the advice itself. Our evaluation of advice centered which practices would be most helpful, and was less concerned with the particular phrasing, given different platform features (e.g., restricting vs. muting accounts). Further, many experts criticized the wording of advice that was prescriptive, explaining that starting advice with "never" (e.g., never share intimate images) could be a non-starter. Instead, P8 described that allowing users to decide for themselves whether to adopt such advice would improve adoption, by ensuring they fully understood the protections and trade-offs of a given piece of advice. This sentiment echoes prior work on security behaviors broadly: "that the benefit [of following security advice] is greater than the cost must be shown, not assumed or asserted" [266]. This further embodies the principle of *enablement* from trauma-informed computing (which builds on the premise that accounting for trauma's effects is widely beneficial for all users, traumatized or not): computing should enable users to make informed decisions for themselves [106].

As with other security advice, experts pointed to times when people might be more receptive to enacting advice, such as after personal experiences with hate and harassment, or after hearing about others' experiences. However, delaying the adoption of advice until after an attack occurs may expose the target to irreversible harms (e.g., content leakage). Such complexities reiterate the need for proactive advice that is generally applicable in the absence of knowing which threat might occur, complementing crisis resources to provide redress after a harm has occurred.

The (apparent) effectiveness of some advice is at tension with the tendency for such advice to further perpetuate and entrench marginalization. Expert opinion was divided on advice seen as effective that also significantly curtailed personal expression (e.g., *never send*

intimate images, be selective about which online communities you participate in, be selective about when and to whom you reveal marginalized aspects of your identity). Some experts judged this advice to be up to personal decision, so users have the final say on what they are comfortable with. However, other experts highlighted how certain advice might seem effective now, but also systematically problematic. For example, never sending intimate images could make content leakage less likely, but it may be interpreted as implying that those who initially send intimate images are at fault and not the perpetrators who actually leak (i.e., nonconsensually share) such content. P8 commented that such advice was victim-blaming because technology should support users in how they choose to express themselves online. Further, self-limiting advice entrenches the marginalization that certain populations already endure. Experts described that some gamers who are women and/or Black avoid harassment by not joining voice channels with strangers, at the expense of their own enjoyment of the games.

Experts discussed the ways that the burden of avoiding harassment online is inequitably distributed, with marginalized populations having already accepted limitations to self-expression in order to exist online. Yet, when experts described how advice for at-risk populations—such as journalists, survivors of intimate partner abuse, or content creators—might differ from general internet users, there was a tendency to strictly recommend more advice, in addition to other high priority advice for all. This poses an untenable burden for marginalized groups to enact tens of pieces of advice for each type of threat. As prior work has stressed, "spending more time on security is not an inherent good" [266].

The status quo places greatest responsibility on individuals to keep themselves safe, necessitating new solutions. Many experts remarked that a majority, if not all, of the burden for staying safer online currently fell to users, reiterating a prior observation that "we [the HCI & security communities] have used user effort as a first resort, not last" [266]. In order to reduce

the need for individual responsibility, many experts commented on the larger need for building communities with norms against hate and harassment. Additionally, experts pointed to the benefits of social support networks in coping (e.g., identifying friends who can provide emotional support) if online hate and harassment occurs. In one expert's estimation, reassurance was a large portion of support:

> *"More than anything, people need comforting, someone to tell them that they're okay."*
> – P15

Social support might be especially valuable for threats where individual agency is low, and thus advice is sparse. For example, there was more advice for content leakage where privacy controls were a central defense, versus overloading or false reporting where attacks depended heavily on attacker capabilities and third-party practices.

These directions work in tandem with producing general and tailored advice. Advice serves as a critical, interim protection during the process of systemic change. Through both individual and communal effort, we hope to create a safer internet for all.

3.6 Conclusion

In this work, we conducted interviews with 24 subject matter experts to understand which pieces of advice can broadly and immediately help most internet users stay safer from online hate and harassment. We used a lens of security and privacy to tackle the broad online hazard of hate and harassment, decomposing it into a set of technology-mediated threats to develop pragmatic guidance for anyone who might be a potential target. Experts weighed different criteria to determine which threats should be prioritized, i.e., prevalence or (potential) severity of the threat, as well as individual agency. This resulted in an overall ranking of toxic content, content leakage, and surveillance as the top three hate and harassment threats most internet users should take

action to prevent or mitigate. Further, we note the factors experts used to evaluate existing pieces of advice—efficacy, ease of implementation, and effect on online participation—and find a select few pieces of advice experts agreed were broadly applicable, while many other threats lacked suitable advice for users to implement. Overall, our work identifies technical and design directions to support users in staying safer from online hate and harassment, while surfacing tensions and challenges on the notion of individual responsibility to do so at all.

Acknowledgements

We are deeply grateful for and recognize the contributions of our expert participants: Eve Crevoshay, Molly Dragiewicz, Jennifer Golbeck, Arzu Geybulla, Weszt Hart, Laura Higgins, Caroline Humer, Rachel Kowert, Liz Lee, Kat Lo, Thomas Ristenpart, Linda Steiner, Gianluca Stringhini, Leonie Tanczer, Elodie Vialle, Viktorya Vilk, Jessica Vitak, Kimberly Voll, Daricia Wilkinson, Sijia Xiao, and our anonymous experts. We thank our reviewers for their valuable suggestions in improving our paper, Anna Turner and Stephan Somogyi for piloting our study, and Tara Matthews for reviewing our methods. This work was supported in part by the U.S. National Science Foundation under award CNS-2205171 and a gift from Google.

Chapter 4

Understanding Help-Seeking and Help-Giving on Social Media for Image-Based Sexual Abuse

Image-based sexual abuse (IBSA), like other forms of technology-facilitated abuse, is a growing threat to people's digital safety. Attacks include unwanted solicitations for sexually explicit images, extorting people under threat of leaking their images, or purposefully leaking images to enact revenge or exert control. In this chapter, we explore how people seek and receive help for IBSA on social media. Specifically, we identify over 100,000 Reddit posts that engage relationship and advice communities for help related to IBSA. We draw on a stratified sample of 261 posts to qualitatively examine how various types of IBSA unfold, including the mapping of gender, relationship dynamics, and technology involvement to different types of IBSA. We also explore the support needs of victim-survivors experiencing IBSA and how communities help victim-survivors navigate their abuse through technical, emotional, and relationship advice. Finally, we highlight sociotechnical gaps in connecting victim-survivors with important care, regardless of whom they

turn to for help.

This chapter originally appeared as the paper "Understanding Help-Seeking and Help-Giving on Social Media for Image-Based Sexual Abuse" at the USENIX Security Symposium in 2024 [596]. 'We' in this chapter refers to me and the co-authors: Sunny Consolvo, Patrick Gage Kelley, Tadayoshi Kohno, Tara Matthews, Sarah Meiklejohn, Franziska Roesner, Renee Shelby, Kurt Thomas, and Rebecca Umbach.

Warning: This chapter includes descriptions and quotes about image-based sexual abuse.

4.1 Introduction

People increasingly share sexually explicit images with consent in intimate relationships [245], as cultural norms change and image sharing capabilities increase. However, this trend has coincided with a rise in *image-based sexual abuse* (IBSA): a continuum of harassment and scams involving the receipt, generation, and distribution of sexually explicit images [382, 383]. Examples include unwanted solicitations for sexually explicit images [490, 548], sextortion [138, 424], and nonconsensually sharing sexually explicit images [383]. In terms of scale, one in ten women under the age of 30 in the US has been threatened with or experienced the nonconsensual sharing of their nude images [334], and roughly one in twenty adult men in the US has experienced sextortion [177].

As with other forms of technology-facilitated abuse—including intimate partner abuse [573], stalkerware [205], and interpersonal surveillance [600]—the support needs of people experiencing IBSA (victim-survivors) are complex. Perpetrators can be intimate partners, peers, or strangers. Even against perpetrators with basic technical capabilities, preventing the distribution of sexually explicit images can be daunting. Support available today includes image fingerprint databases

used by platforms to take down imagery [407, 554] and institutional guides on how to respond to IBSA [439]. However, victim-survivors may be unaware of these resources or find them ineffective, leading them to seek alternative support.

In this chapter, we explore how adults seek and receive help for IBSA on Reddit, a popular social media platform for threaded dialogue. Given the scarcity of victim-survivors who turn to law enforcement or platforms for help [91, 495], social media provides an important avenue for disclosure and support. Expanding on knowledge from prior work studying help-seeking on Reddit for sexual abuse (e.g., rape) [20, 396, 426], we focus on help-seeking on Reddit across the continuum of *image-based sexual abuse*. Specifically we investigate:

1. **IBSA types**. What types of IBSA do people seek help for on Reddit? What gender and relationship dynamics between perpetrators and victim-survivors do they disclose? How might differences between IBSA types influence the development of supportive solutions?
2. **Help-seeking**. What help do they seek? How do their needs vary across IBSA types? What actions do they share that they have already taken?
3. **Help-giving**. What help do they receive? How supportive is it? What gaps does this help fill compared to other interventions? What gaps remain?

To answer these questions, we used a mixed methods approach to sift through 5.7 million English-language Reddit posts shared on relationship and advice subcommunities over the last 3 years. Leveraging a novel large language model (LLM) data processing pipeline and extensive manual review, we identified more than 100,000 queries for help related to IBSA—roughly 2% of posts on the subcommunities. This method allowed us to analyze a much larger sample than previous qualitative Reddit work. We drew on a stratified sample of these posts to qualitatively explore the continuum of IBSA including financial and nonfinancial sextortion, nonconsensual *synthetic* explicit imagery, pressurized sexting, cyberflashing, nonconsensual explicit imagery, and

recorded sexual assault.

We found that although IBSA covered a wide range of contexts, perpetrators, and harms, victim-survivors nevertheless shared similar needs: to be heard and supported through life-changing experiences of abuse. Across types of IBSA, victim-survivors sought information about their options (technical, legal, or otherwise), advice for coping with distressing emotions, and suggestions for managing relationships. Timely support was crucial: in half of the cases we analyzed, victim-survivors were seeking immediate help for active IBSA. Though the help provided by the Reddit community was sometimes oversimplified and made limited use of institutional support resources, it was also largely empathetic and validated victim-survivors' experiences, helping to partially address their needs.

Our work characterizes IBSA to chart additional directions to support victim-survivors, in ways that complement the distinct avenues for help that exist today. We reflect on the role of technology in facilitating IBSA and identify opportunities for technologists and platforms to help prevent or mitigate IBSA, while also discussing how our insights can inform advocates in providing support for victim-survivors.

4.2 Related Work

4.2.1 Image-based sexual abuse (IBSA)

IBSA is an umbrella term referring to "taking, distributing, and/or making threats to distribute, nude or sexual image[ry][1] without a person's consent" [454, p. 1] (see also: [159, 382, 383, 452, 453]). For the purposes of our study, we expand this definition to include the unwanted receipt of and

[1]As with prior work [454], we use "imagery" to include photos and video.

synthetic generation of sexually explicit[2] images, that is, images portraying nudity or sexual conduct. IBSA is a type of technology-facilitated abuse in which a person leverages technology to exert control over another [56, 159, 259]. We refer to the person enacting IBSA as the *perpetrator* and the person experiencing IBSA as the *victim-survivor*.

Continuum of IBSA. IBSA exists on a continuum, occurring in diverse relational contexts, with distinct motivations, and rapidly evolving threats [383]. The types of IBSA that inform our investigation include:

1. *Sextortion*: a perpetrator makes threats to distribute sexually explicit images of a person unless they comply with the perpetrator's demands [138, 198, 261, 424].
2. *Nonconsensual synthetic explicit imagery (NCSEI)*: a perpetrator uses software (e.g., photo editing or generative AI tools) to create sexualized depictions of a person (e.g., "deepfakes," "cheapfakes") [197, 263, 378, 420, 610].
3. *Pressurized sexting (PS)*: a person experiences unwanted solicitation for explicit images (e.g., "coerced sexting") [490, 548].
4. *Cyberflashing (CF)*: a person receives an unwanted explicit image [380].
5. *Nonconsensual explicit imagery (NCEI)*: a perpetrator uses explicit imagery for revenge or to enact control [383] (e.g., "revenge porn"); or otherwise nonconsensually creates, retains, or distributes explicit imagery (e.g., "upskirting," "downblousing") [41, 339].
6. *Recorded sexual assault (RSA)*: a perpetrator records or distributes imagery from a sexual assault [258].

Role of technology. Technology often plays a role in both the generation and distribution of IBSA. Outside of consensual sharing or nonconsensual recording, methods of obtainment

[2]Sexually explicit imagery can be distinguished from intimate imagery, which also includes images of people in private or sensitive contexts (e.g., sleeping, in states of intoxication, or without religious coverings).

include hacking [200], photoshopping [384], and using generative AI tools to create synthetic imagery [378, 578, 610]. Distribution channels used by perpetrators include—but are not limited to— websites [115, 257, 353, 576], social media platforms [200, 378, 518], and mobile apps [280, 381].

Harms. Harms from IBSA—as with other forms of sexualized violence—are not uniformly experienced, but are nonetheless serious and consequential [37, 55, 302, 495], carrying emotional, physical, financial, and social impacts [279]. When images are distributed, victim-survivors become visually recognizable to friends, family, co-workers, or others [279], and imagery is often shared with other personally identifiable information (e.g., names, social media handles, phone numbers, and/or addresses [114, 200, 555]). The often permanent nature of online material intensifies IBSA harms, as imagery can be repeatedly downloaded, saved, and shared, making "complete" removal challenging, or even impossible [34]. As IBSA often co-occurs with other forms of gender-based violence (e.g., intimate partner abuse, stalking, and sexual harassment) [381], harms may compound with other polyvictimizations.

4.2.2 IBSA help-seeking and help-giving

Overall, rates of IBSA help-seeking through peers and family, institutional support, and platforms are low [91, 495]. One potential reason is that in contrast to other forms of sexualized violence, targets of IBSA are inherently not granted anonymity, as they are facially (or otherwise) identifiable [279]. Another challenge is that help-givers may hold victim-blaming attitudes, which have been found among the general public [88, 196, 631], law enforcement [59, 632], and victim-survivors' friends and family [547].

Online help-seekers for topics such as health often look for specific advice or information, acknowledgement, or sympathy [21, 231, 410]. Prior work also studies online help-seeking for sexual abuse (e.g., rape); by contrast, our work considers online help-seeking for *image-based*

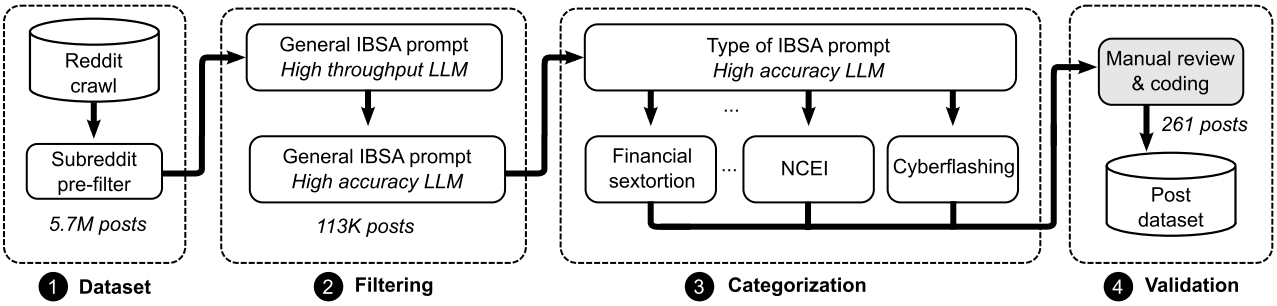

Figure 4.1: Data processing to identify help-seeking posts on Reddit. We relied on recent advances in LLMs to automatically sift through 5.7M Reddit posts to identify 113K posts likely related to help-seeking for IBSA. We automatically labelled each post with one of seven types of IBSA. Coders randomly pulled a sample of these posts to manually validate the LLM-generated metadata and reach a minimum threshold of 40 posts for each of the seven IBSA types. As a post could mention multiple types of IBSA, our *Post dataset* includes a total of 261 posts.

sexual abuse. When disclosing online about sexual abuse (that was not image-based), victim-survivors perceived online fora as safer spaces to disclose stigmatized experiences and connect with others, with reduced interpersonal risk [19, 20, 303, 396, 426]. As we show, social media—specifically Reddit—has at least hundreds of thousands of posts related to IBSA help-seeking. Thus it is crucial to understand this support system and how best to enmesh it with other social, institutional, and platform resources.

4.3 Methodology

We conducted a mixed-methods study of IBSA help-seeking and help-giving experiences on Reddit. We discuss the process we followed to identify IBSA posts, our qualitative coding practices, and the ethics and limitations of our approach.

4.3.1 Data processing

We built a data processing pipeline that automatically identified help-seeking for several types of IBSA discussed on Reddit. We outline the full pipeline in Figure 4.1 and discuss each component

Title: [TITLE]
Article: [ARTICLE]

Question: Does the title or article discuss any form of image-based sexual abuse, such as sending or receiving unwanted nude images, unwanted flashing, being coerced into sharing nude images, fear of having nsfw or intimate images leaked or hacked, sextortion, nonconsensual image sharing, or revenge porn? Provide both an answer and a summary explaining your reasoning.
Answer:

Figure 4.2: Initial prompt designed to identify posts that contained any discussion of IBSA.

below.

Dataset. The first step in our data processing pipeline was to identify Reddit posts and subreddits relevant to our research (Figure 4.1, ❶). Our Reddit dataset originates from a continuous Internet-wide crawl of public URLs in a way that respects robots.txt and other rules for crawlers. In order to reduce the search space for IBSA posts, we focused on 43 popular subreddits related to scams, advice, relationships, dating, and sex—including r/advice, r/askwomen, r/askmen, and r/sextortion—that we identified from an initial manual exploration (see Appendix A.1.1 for the full list). All of these subreddits had at least 1,000 members. In total, this dataset includes 5.7M English-language posts from 2.9M unique users (i.e., "original posters" on Reddit, posters in this work) published between April 1, 2020 and September 12, 2023.

Filtering. We analyzed each Reddit post using a suite of LLMs and prompts to identify those likely to be about help-seeking for IBSA (Figure 4.1, ❷). While we initially explored a keyword-based search, the terminology and context proved too nuanced to capture without a prohibitive number of false positives to manually review. Instead, we queried Google Cloud's lighter-weight Vertex AI text-bison LLM [227] with a prompt to identify posts generally discussing IBSA (Figure 4.2), then refined the search by repeating the prompt with a more accurate (but expensive) text-unicorn LLM [227].

We validated this approach using a manually curated test set of 80 posts that discussed IBSA

Table 4.1: Our qualitative dataset consists of seven distinct types within the continuum of IBSA, as defined in Section 4.2.1. Prior work largely collapses FS and NFS into the single category *sextortion*. For our purposes, the differences were meaningful, so we split them into two distinct categories. For each type, we include the total number of stratified posts and comment threads that we analyzed. A single post could describe experiences with more than one IBSA type.

Type	**Key**	**Posts**	**Threads**
Financial sextortion	FS	50	22
Nonfinancial sextortion	NFS	46	26
Nonconsensual *synthetic* explicit imagery	NCSEI	51	32
Pressurized sexting	PS	45	34
Cyberflashing	CF	41	33
Nonconsensual explicit imagery	NCEI	45	35
Recorded sexual assault	RSA	40	20
Total (Any type of IBSA)	–	261	160

(generated from our initial manual exploration and when exploring the feasibility of keyword matching), and 197 non-IBSA posts. Our chained prompt correctly identified 79 of the 80 IBSA posts (98.8% recall) and 194 of the 197 non-IBSA posts (98.4% specificity). While the final dataset of posts we use is manually validated to remove all errors, this high degree of recall and specificity gave us confidence that we did not incorrectly omit some IBSA concepts from our study, while at the same time reducing the toil of manual validation. When applied at scale, our filtering identified 113K Reddit posts likely discussing IBSA—2% of posts on the subreddits we analyzed.

Categorization. To ensure that our study captured distinct experiences in the continuum of IBSA, we queried text-unicorn to categorize each post by the type(s) of IBSA involved in order to support stratified sampling (Figure 4.1, ❸). We derived these types—listed in Table 4.1—based on our preliminary analysis of prior literature (Sections 4.2.1 & 4.3.2). We found that the LLM struggled with some concepts more than others. For example, *sextortion* was easy for the LLM to categorize, while *recorded sexual assault* was more error prone (and also more rare). To counteract this and account for posts that discuss multiple types of IBSA, we also created (more computationally

expensive) per-type prompts with multi-stage reasoning to distinguish overlapping concepts (e.g., financial and nonfinancial sextortion). See Appendix A.1.2 for the general categorization and specialized prompts. We rely on manual review to validate these labels, discussed next. Given the potential for error with automated categorization (absent manual review), we avoid providing a relative breakdown of the types of IBSA at-scale, and focus only on our stratified sample.

Validation. During validation, we selected from the outputs of each categorization prompt and manually vetted the labels (Figure 4.1, ❹). Here, we corrected for any missed or erroneous labels and discarded any posts that were not related to help-seeking for IBSA. We also removed posts where the victim-survivor appeared to be under the age of 18.[3] We repeated this until we achieved a minimum of 40 posts for each of the seven types of IBSA,[4] resulting in a final dataset of 261 posts. We sampled 40 posts per IBSA type to achieve meaning sufficiency [73], balanced with reasonable researcher effort. Some posts involved multiple types of IBSA, thus requiring we label fewer than 280 posts to meet our minimum sample size. Each post in our sample on average contained 285 words (min: 24, max: 2,363).

For each post, we re-crawled the live site for the current state of discussion (i.e., all threads) on December 11, 2023. This yielded 2,159 threads consisting of 4,225 individual comments from 2,298 unique accounts. To allow for a robust qualitative analysis, we filtered this set down to the top three upvoted threads per post, and then randomly sampled 160 of these popular threads. We use *posts* to understand the types of IBSA for which people were seeking help and what help they were seeking, and *threads* to understand what help they received. We summarize our qualitative dataset in Table 4.1.

[3]Some posts may originate from minors, but we cannot confirm. The advice we analyzed appeared targeted towards adults. The nuances of help-seeking and giving are fundamentally different for child sexual abuse (e.g., lack of agency, mandatory reporting) and thus beyond the scope of this work.

[4]We sampled independently per IBSA type. Some posts discussed multiple IBSA types, which is why there are more than 40 posts for some types.

4.3.2 Qualitative analysis

We relied on a rigorous, qualitative analysis to describe and identify themes in our 261 posts and 160 threads. We used codebook thematic analysis (TA) that combined inductive and deductive approaches, which aligned to our mixed-methods approach [191]. We employed a five-stage codebook TA [491]: (1) sourcing initial codes; (2) developing initial codes; (3) codebook design; (4) codebook application; and (5) interpretation. For this study, use of a codebook enabled a refined and focused analysis of qualitative data [132].

Sourcing initial codes. As we developed our research questions, we conducted an initial analysis of IBSA literature and preliminary scan of IBSA help-seeking posts on Reddit – to summarize and identify potential codes [132] relevant to help-seeking needs, help-giving behaviors, and other characteristics of the data that applied to our research questions. Existing literature on IBSA types and risk factors informed our initial deductive codes; codes describing IBSA help-seeking needs and help-giving behaviors were developed inductively, as these are more emergent areas in the literature.

Developing initial codes. Next, one researcher familiarized themselves with the raw data, cumulatively reading thousands of Reddit posts about IBSA during the development of our data processing pipeline (Figure 4.1). This researcher then applied the initial codes to a random set of 100 posts stratified across IBSA types and then generated new codes from the raw data to identify more useful attributes of the data.

Codebook design. Next we developed the codebook, finalizing labels, definitions, and exclusions. All members of the research team reviewed the codebook; we took note of initial disagreements to iteratively update our codebook to account for these nuances. We segmented our codebook into three sections. The first section focused on the *nature of the IBSA*: the type of IBSA, the platforms involved, details about the perpetrator and victim-survivor, the origin of the IBSA imagery, and

method of distribution. The second section focused on *help-seeking*: when help-seeking occurred, strategies already attempted, and what help was sought. The third section focused on *help-giving*: the type of support or advice offered and the interactivity of help-giving. See Appendix A.1.3 for details on individual codes and definitions.

Codebook application and reliability. Four coders applied the codebook to our posts dataset and three applied the codebook to our threads dataset, with two independently coding each post or thread. For reliability, we used *consensus coding* [92] for consistent judgment [65]. All coders iteratively discussed disagreements via meetings or online chat. Between discussions, one coder reviewed remaining disagreements, resolved obvious or already-discussed issues, and noted where discussion was still needed. We chose to ensure coding reliability through consensus coding and discussion because our codes were not mutually exclusive and the data were nuanced.

Interpretation. Lastly, we iteratively collated various codes, reviewed the data and memos, and discussed themes. Alongside our thematic analysis, we offer descriptive statistics. We report counts in the results with the notation *X of Y*, where Y is the total number of posts, threads, or people about which that code is specified, that is, excluding unspecified. In sections where Y is constant, only X is reported for brevity.

4.3.3 Ethics

Similar to prior work [43, 573, 600], we rely on data shared publicly by users on social media. We excluded all posts where it was clear that the poster/victim-survivor was under the age of 18. Our work does not directly recruit participants, but our study plan was reviewed by experts at our institution in domains including ethics, human subjects research, policy, legal, security, privacy, and anti-abuse. While the institution of the authors who conducted the data analysis does not require IRB approval, we adhere to similarly strict standards.

To mitigate potential harms that may come from victim-survivors' data being exposed to unexpected audiences, we rephrased all quotes to preserve meaning but obscure the original source. Strategies for disguising the source of content are increasingly common in research fields investigating social media data [173, 457], as well as recommended by digital-safety researchers [44], particularly when obtaining informed consent is not feasible. Scholars emphasize that disguise is an ethical practice for protecting participants' privacy [363]; additionally, contacting the original posters of the content we studied could unnecessarily re-traumatize victim-survivors. After rephrasing quotes, we searched each rephrasing to ensure the original source was not identifiable in the returned results. To balance protecting posters' privacy with data integrity, another researcher compared the rephrasing to the original quote to confirm the semantic meaning of the quote was not changed. This rephrasing was post-hoc and did not affect our analysis.

As this project involved sustained engagement with traumatic content, researchers involved in analysis took different measures to support their well-being, including: weekly individual and group check-ins, reading about secondary trauma [346], meeting with trauma-informed experts, having access to therapists, taking breaks (e.g., playing Tetris, which is being explored as a tool for reducing traumatic flashbacks [291, 534]), and restricting reading of traumatic posts to shared or designated workspaces. Our use of LLMs also reduced the amount of manual review required.

4.3.4 Limitations

As with all research, ours has limitations. Our crawl of Reddit may be incomplete, and our LLM-based search for IBSA help-seeking may miss some concepts. We attempted to minimize biases by validating our prompts on an independent sample of posts and by focusing on qualitative results rather than comprehensive quantitative findings. Additionally, our US-based research team—whose domain expertise includes computer security and privacy, human-computer interaction,

criminology, gender-based violence, and social computing—apply our own interpretations to the stories shared on Reddit.

Our visibility into IBSA help-seeking and help-giving is also limited to what people mention when posting to Reddit in English. Reddit users are predominantly in the US, but also in the UK, India, and Canada [614]. Additionally, comments studied in Section 4.6 do not include those removed by Reddit mods, whose invisible labor contributes valuable content moderation [341]. Percentages are reported in the results for reader ease but should not be interpreted as generalizable to all cases of IBSA, given the limitations of our data collection.

How we report gender. Because IBSA can be a form of gender-based violence, we coded gendered terminology about the victim-survivor(s) and perpetrator(s) when specified.[5] One challenge inherent in analyzing social media data is that posts do not reliably or consistently provide gender disclosures. We inferred gender in five ways: Reddit conventions to self-identify (e.g., "21F" meaning a 21-year-old female), gendered nouns (e.g., woman, boyfriend, girl), gendered pronouns (e.g., she, he), body parts (e.g., dick pic, breasts), or other (e.g., posting to a gendered subreddit, asking for opinions of "other women"). We most often inferred victim-survivor and perpetrator gender through gendered nouns (24%; 60%) or self-identifications (56%; 23%), but also inferred based on solely pronouns (8%; 17%) and body parts (26%; 1%). To avoid piecemeal reporting, we collapsed sex and gender, such that "man" includes all masculine terms, including gendered nouns and pronouns, and "woman" likewise includes all feminine terms. These codes should be interpreted as researcher inferences based on gendered terms in Reddit's broadly cisheteronormative context, which may or may not align with the gender identities of the individuals involved.

[5]Gender was unspecified for 26% of perpetrators and 42% of victim-survivors.

4.4 Characterizing IBSA Experiences

We begin by characterizing the IBSA experiences about which posters sought help. While prior work has identified the types of IBSA covered here, we contribute a description of all seven types from the same dataset, highlighting common, co-occurring, and distinguishing patterns across types. This approach provides an expanded understanding of a range of IBSA experiences and sets the stage for how to broadly support help-seeking needs.

In most cases, the poster identified as the victim-survivor (in 237 of 261 posts; 91%); in others (30; 11%), the poster was seeking help on behalf of the victim-survivor, e.g., a friend or intimate partner.[6] We synthesize these experiences, examining how abuse unfolded, perpetrators' apparent motivations, and gendered patterns between perpetrators and victim-survivors.

4.4.1 Financial sextortion (FS)

Financial sextortion occurred when a perpetrator threatened to expose explicit images of a victim-survivor unless the perpetrator was paid money.

Clear methods to obtain images. To initiate contact for financial sextortion, a perpetrator typically connected with a victim-survivor via a dating or social media app before moving to direct messaging or a communication app, engaging in a conversation with the victim-survivor within minutes to days. The perpetrator then coercively obtained, made claims about, or created explicit images following one or a combination of the following methods: they sent (inauthentic) explicit images to encourage the victim-survivor to reciprocate; they claimed to have an explicit image of the victim-survivor (with or without evidence); and/or they created an explicit image (e.g., by attaching the victim-survivor's face—such as taken from a profile picture—to someone

[6]Counts do not sum to 261 because some posters were both a victim-survivor and seeking help on behalf of other victim-survivors of the same perpetrator(s).

Table 4.2: Relationship of perpetrator(s) to victim-survivor, as described in posts.

Perpetrator	**FS**	**NFS**	**NCSEI**	**PS**	**CF**	**NCEI**	**RSA**
Stranger	48	18	31	12	26	6	8
Intimate partner	0	10	8	21	3	15	4
Friend	1	0	5	5	4	7	8
Ex-intimate partner	0	10	1	2	1	8	8
Colleague	0	0	1	2	4	1	4
Family member	0	0	1	1	1	2	0
Other	1	0	0	0	2	0	0
Unspecified	0	8	5	3	0	6	10

Table 4.3: Gender of the perpetrator(s), as described in posts. *Ambiguous* refers to posts where the poster initially described the perpetrator with feminine gendered terms but shifted to masculine gendered terms when a false identity was revealed.

Perpetrator	**FS**	**NFS**	**NCSEI**	**PS**	**CF**	**NCEI**	**RSA**
Woman	21	8	15	2	6	1	3
Man	3	20	16	39	27	31	31
Ambiguous	4	2	3	0	0	0	0
Unspecified	22	15	16	1	8	13	6

else's body; such synthetic images are further discussed in Section 4.4.3).

Once the perpetrator had established or asserted that they had an explicit image, they issued a threat: either the victim-survivor had to pay or the explicit image would be distributed (e.g., *"Send me money or I'll share your nudes"*). Payment demands varied, but were typically the local equivalent of $100–1500 USD (specified in 28 of 50 posts; 56%).

Perpetrators were usually strangers, often women; victim-survivors men. Of the 50 posts, 48 identified the perpetrator as a stranger (Table 4.2). Of the 28 posts that specified the perpetrator's gender, 21 were believed[7] to be women (Table 4.3). Of the 27 posts that specified the

[7]Some posters, in line with prior work [198], acknowledged that perpetrators may be using this gender identity as

Table 4.4: Gender of the victim-survivor(s).

Victim-survivor	FS	NFS	NCSEI	PS	CF	NCEI	RSA
Woman	1	13	11	25	18	27	26
Man	26	9	22	3	2	3	9
Unspecified	23	24	19	18	25	18	8

victim-survivor's gender, 26 were men (Table 4.4).

Fear & self-blame. Perpetrators tried to exploit the victim-survivor's fear of embarrassment. Posters often emphasized the threatened scale of distribution: *"I'll send these masturbation videos to everyone you know, including your friends, colleagues, and relatives."* Victim-survivors often appeared to be in a state of panic when they posted, and they sometimes partially blamed themselves, claiming that they *"messed up"* or *"feel so stupid."*

What they've done; what else could they do. Posters sought help to determine if the threat was credible, prepare for what to expect next, and ask for strategies to cope with their fear. When seeking help, victim-survivors often mentioned technical protections they had already taken, such as blocking the perpetrator, reporting the perpetrator to the platform, or bolstering their social media privacy and security settings. They then asked the community what else they could do.

4.4.2 Nonfinancial sextortion (NFS)

At a high level, nonfinancial sextortion was similar to financial sextortion except the threats were not financially motivated.

Personalized demands for more images or control. Nonfinancial sextortion tended to be customized to the particular victim-survivor and their situation. Perpetrators—who were often

a false persona.

intimate partners—used the threat of exposing an explicit image to demand new images and/or to exert emotional, physical, or sexual control over the victim-survivor: *"he was obviously getting mad and eventually exploded and told me he'd send out my nude pics if I didn't have sex with him."*

Perpetrators often were intimate partners, usually men; victim-survivors women. Nonfinancial sextortion was usually perpetrated by a current (10 of 38 specified; 26%) or former intimate partner (10; 26%) (Table 4.2).[8] Victim-survivors often noted that the image the perpetrator threatened to expose was obtained with consent in a different context (8 of 43 specified), or nonconsensual secret recording (4). When specified, perpetrators tended to be men (20 of 30; 67%), victim-survivors women (13 of 22; 59%) (Tables 4.3 & 4.4).

Fear & helplessness. Like financial sextortion, perpetrators of nonfinancial sextortion attempted to control victim-survivors with fear of embarrassment. However, when compared to financial sextortion, posts about nonfinancial sextortion tended to include a more pronounced tone of helplessness: *"I feel suffocated with him and I don't see any type of future with him, but he has the power to ruin my life."*

Relationship & emotional advice. Victim-survivors expressed uncertainty about how to navigate their relationship with the perpetrator: *"should I cave to his demands so he doesn't dump me?"* They also asked for help coping with the emotional trauma associated with this type of IBSA.

4.4.3 Nonconsensual *synthetic* explicit imagery (NCSEI)

Nonconsensual synthetic explicit imagery involved digitally manipulated explicit images. A majority of cases co-occurred with financial sextortion (27 of 51 posts, 53%; discussed in Sec-

[8]We coded posts about sextortion that did not mention a demand for payment as NFS. Such posts where the perpetrator was a stranger resembled (and may have been) FS, so we omit them from discussion here.

tion 4.4.1). Among other cases, a main theme appeared to be about the perpetrator exploring fantasies, especially when the victim-survivor and perpetrator were socially connected.

Simple manipulations. Posters explained that basic editing techniques were used to generate explicit images, using terms like *"photoshopped"* or *"filters with my face."* In cases of financial sextortion, perpetrators often sent victim-survivors a "collage" of explicit images and other media—a list of the victim-survivor's social media connections, screenshots of messages between the victim-survivor and perpetrator, or calls-to-action from the perpetrator to incriminate the victim-survivor (e.g., calling them a "pedophile")—to scare the victim-survivor into paying.

In other cases, posters found or learned about explicit images on the perpetrator's—usually the poster's intimate partner's—device, presumably for personal use such as the perpetrator's sexual gratification or exploration. In these images, victim-survivors' or other relations' faces were superimposed into explicit images: *"I recently found out [my intimate partner] had tons of pictures with my friends and family where their faces had been photoshopped onto sexually explicit photos to simulate porn scenes."*

The emergence of sophisticated manipulations. The (suspected) use of generative AI to create nonconsensual synthetic explicit imagery was mentioned in 3 of 51 posts (6%). In those cases, victim-survivors found the imagery's photorealism terrifying: *"Someone used a nudifying app to make naked photos of me from my social media pictures. I'm terrified."*

Nonconsent, fear, & disgust. The quality of the synthetic imagery was not a focus of posters; instead, they discussed the harm that even simple manipulations could create. In financial sextortion cases, victim-survivors shared that synthetic imagery might deceive family or peers: *"The nudes are fake, but it's not like my parents can tell that."* In other cases, posters expressed disgust because the synthetic imagery was nonconsensually created by people the poster knew, and the images portrayed people in sexually explicit ways that the poster found inappropriate or offensive.

For example, one poster found synthetic images of a friends' family member, which left the poster *"in tears... I'm utterly disgusted and distraught by the situation."* Another poster discovered that their partner had been creating images of their friends and family for years, and despite knowing they were fake, viewed the creation of the images as a grave breach of trust: *"The files... were created a few years ago and I cannot believe how blind I was. I trusted him and now I am broken."*

Navigating relationships & feelings. Posters sought help for how to deal with perpetrators (including asking for relationship advice) and to make sense of how they felt about the discovery of nonconsensual synthetic explicit imagery.

4.4.4 Pressurized sexting (PS)

Pressurized sexting occurred when a perpetrator coerced or demanded that a victim-survivor send explicit imagery. In a majority of these posts (29 of 45; 64%), victim-survivors claimed they had not sent the images; rather, they asked for help mitigating coercion.

Navigating relationship expectations. Pressurized sexting occurred primarily in relationships, that is, by current intimate partners of the victim-survivor (21 of 42 specified; 50%), or between strangers who were prospective partners via online dating or social media apps (12; 29%).

New vs. established relationships. When pressurized sexting occurred in the early stages of a (potential) relationship, posters expressed feelings of being manipulated or "used for nudes": *"I can't believe someone asks for nude pictures, and their only motivation in talking to me is trying to get nudes."*

When in established relationships, victim-survivors often attempted to resolve conflicts between what they were comfortable with and their intimate partners' expectations. For example, some asked if pressurized sexting warranted ending a relationship: *"My boyfriend asks for nudes and if I refuse he ignores me until I give in. Can someone tell me if we can fix this, or if I should just*

get therapy to help me break up with him?" This was especially true if the pressurized sexting was perceived as the sole problem in a relationship. *"This guy is great except he asks me to send nudes... I don't want to break up but I hate when he does this. What should I do?"*

Perpetrators were usually men, victim-survivors women. Most perpetrators were presumed to be men (40 of 43 specified; 93%); victim-survivors tended to be women (25 of 27 specified; 93%) (Tables 4.3 & 4.4).

4.4.5 Cyberflashing (CF)

Cyberflashing occurred when a victim-survivor received unwanted sexually explicit imagery.

Online dating. Cyberflashing commonly occurred in the process of seeking romantic relationships online. Victim-survivors frequently discussed whether they were out-of-step with shifting norms around sending or receiving explicit images: *"Can someone please explain why people send unsolicited nudes?"* In some cases, cyberflashing occurred as part of pressurized sexting (Section 4.4.4), where a perpetrator would demand reciprocity after sending unwanted explicit images: *"I told him I didn't want any, but he sent dick pics anyways...he asked for a picture of my breasts but I said no cuz I never send naked pictures of myself."* The tone of these posts was often that of frustration, disgust, and disillusion.

Platonic relationships. Less often, cyberflashing occurred in platonic relationships (e.g., between coworkers or friends). In those cases, the poster was typically concerned about the ongoing relationship and if/how they should address the abuse.

Perpetrators were usually men, victim-survivors women. Most perpetrators were men (27 of 33 specified; 82%), while victim-survivors were women (18 of 20 specified; 90%).

4.4.6 Nonconsensual explicit imagery (NCEI)

Nonconsensual explicit imagery involved perpetrators producing or distributing explicit images without the victim-survivor's consent. It often involved abuse by an intimate partner, ranging from a one-time incident to ongoing abuse.

Perpetrators were usually men; often partners, friends, or peers. Victim-survivors were usually women. Victim-survivors were usually women (27 of 30 specified; 90%), while perpetrators were often men (31 of 32 specified; 97%). Most perpetrators were known to the victim-survivor (33 of 39 specified; 85%) rather than strangers (6; 15%).

Images were often recorded without consent. The trusted status many perpetrators had with victim-survivors enabled secret recordings (20 of 38 specified; 53%). Victim-survivors shared how perpetrators recorded them: *"took nude photos of me while sleeping"* and *"recorded me [naked] in my room."* Less frequently, images were recorded through coercion (6; 16%) or shared with consent in a different context (6; 16%).

Navigating relationships; fear of possession & distribution. Nearly half of the posts focused on how to navigate the relationship (e.g., whether and how to confront or break up with the perpetrator). Victim-survivors were concerned about the perpetrator's possession of the explicit imagery (18 of 41 specified; 44%) or its distribution, e.g., via messaging (14; 34%), social media (9; 22%), or public websites (3; 7%).

Victim-survivors concerned mainly with *possession* primarily sought help on how to negotiate with the perpetrator, delete any explicit imagery, know if the imagery had been shared, or know how many images existed: *"How can I find and delete as many as possible of the naked photos he has of me?"* Conversely, victim-survivors contending with *distribution* focused their help-seeking on what recourse, if any, existed: *I dunno what to do. He sent my mom my nudes because I was breaking up with him... Any advice would be appreciated.*

4.4.7 Recorded sexual assault (RSA)

Recorded sexual assault involved the creation or distribution of images of sexual assault. It frequently resulted in victim-survivors experiencing ongoing sexual trauma in addition to trauma from images being created or distributed.

Perpetrators were often known men; victim-survivors tended to be women. Perpetrators were often men (31 of 34 specified; 91%) that victim-survivors identified as former intimate partners (8 of 32 specified; 25%), friends (8; 25%), colleagues (4; 13%), or current intimate partners (4; 13%), but also strangers (8; 25%). Victim-survivors were often women (26 of 35 specified; 74%), but sometimes men (9; 26%).

Co-occurred with other abuse. Victim-survivors shared that their assault was often secretly recorded (23 of 40; 58%), such as when they were unconscious, sleeping, drugged, or under the influence of a substance. Recorded sexual assault often co-occurred with other abuse (13 of 39; 33%)—sometimes prolonged intimate partner abuse—leaving victim-survivors in a state of severe trauma. Most victim-survivors shared that perpetrators retained recordings of the assault (26 of 40; 65%); some perpetrators distributed the recordings via messaging (7; 18%), social media (4; 10%), or websites (4; 10%).

Trauma, processing, & coping. All of these posts were deeply troubling and resulted in trauma to the victim-survivor and typically others who were exposed to the recording. Most victim-survivors asked how to cope with or understand what they had experienced, as a precursor to processing their emotions or taking a step towards remediation: *"Was this sexual assault?," "Did I just make poor decisions or was this wrong?," "I don't know how to feel,"* and *"I'm wondering if any part of this sexual assault was normal."*

4.4.8 Co-occurring IBSA types

54 posts described more than one IBSA type. Most prominently, of 51 posts on nonconsensual synthetic explicit imagery, 27 co-occurred with financial sextortion and 7 with nonfinancial sextortion, indicating that perpetrators commonly leveraged synthetic images to sextort. Cyberflashing and pressurized sexting co-occurred in 9 posts, indicating that perpetrators sometimes both sent and pressured the victim-survivor for explicit images. Nonconsensual explicit imagery co-occurred with multiple other IBSA types in 9 posts (out of 45 total); these posts mostly described multiple abusive events. One post described three types of IBSA (NCEI, cyberflashing, pressurized sexting), and one described four (NCEI, recorded sexual assault, pressurized sexting, nonfinancial sextortion).

4.5 Help-Seeking for IBSA

Across IBSA types, posts mentioned prior remediation efforts and included some common patterns of help-seeking questions, such as informational questions (Section 4.5.3), as well as how to cope with emotions (Section 4.5.4) or navigate relationships (Section 4.5.5)—these and other less common help-seeking types are summarized in Table 4.5.

4.5.1 When posters sought help

Using a theoretical framework [296], we can describe patterns in *when* people sought help for IBSA. This framework defines four states of users experiencing safety events: *prevention, monitoring, crisis,* and *recovery* (see details in Appendix A.1.3). People may move through the states nonlinearly and experience multiple events at once. When coping with *multiple safety events* (57 of 257 posts;

Table 4.5: Types of help sought across the seven types of IBSA. Some posts discussed multiple forms of IBSA and were counted in multiple cells per row; some posts also included multiple types of help-seeking.

Help Sought	Description	FS	NFS	NCSEI	PS	CF	NCEI	RSA
Informational	Seeking general advice about options or potential actions.	37	25	29	9	22	20	16
Therapeutic	Seeking advice about emotions or other distressing elements.	17	11	16	19	10	7	25
Relational	Seeking advice about interpersonal dynamics or managing relationships with others.	3	17	13	25	14	18	5
Legal	Seeking legal avenues of recourse or asking legal questions.	5	4	8	3	7	9	7
Technical	Seeking advice about mitigating the abuse through technical means, potentially to prevent primary or secondary sharing of the image, or tracking the past sharing of the image.	3	1	2	1	1	4	1
Other	Seeking a type of help not listed above.	4	3	2	1	0	0	1

22%)[9], posters' panic and trauma were compounded.

Help was most often sought while the poster was in *crisis*, that is, when they were actively dealing with abuse (142; 55%). These posters expressed panic and needed immediate reassurance, as well as clear, simple guidance on the most critical steps to stop or limit further damage. Many posters also sought help while *recovering* from the abuse and feeling traumatized by it (112; 44%). Fewer posters described being in *monitoring*, for example, watching for new abuse or for images to appear in new places (42; 16%). Rarely, posters sought help *preventing* victimization (5; 2%).

4.5.2 Prior attempts at remediation

Posters mentioned prior attempts at remediation (in 176 of 261 posts; 67%), which were only partially successful or failed, prompting them to seek help on Reddit (Table A.1 in Appendix A.2).

(Dis)engaging perpetrators. Of posts that noted prior attempts at remediation, nearly half mentioned engaging with perpetrators through negotiation or mediation (76 of 176 posts; 43%). This strategy often occurred when IBSA was enacted by an intimate partner, including NCEI and

[9]Number of posts here is 257 because four did not specify the user state.

pressurized sexting. Another nearly half disclosed that the victim-survivor disengaged from the perpetrator, cutting off communication (82; 47%); some attempted engaging and then disengaging. Disengaging was common when the perpetrator was a catfisher or scammer, including financial sextortion and nonconsensual synthetic explicit imagery.

Technical strategies. Technical strategies, which included securing accounts or devices, platform reports, and deleting content, were sometimes employed (49 of 176 posts; 28%). They were infrequently reported across IBSA types, except for financial sextortion, for which victim-survivors more consistently reported deleting accounts or reporting. Often, sexually explicit images resided on a perpetrator's personal device or chat history, making it technically challenging for victim-survivors to access and for platforms to respond. As many help-seekers posed open-ended questions (e.g., *"what should I do?"* as described in Section 4.5.3), some may not have been aware of what technical strategies were possible.

Social & institutional support. Victim-survivors rarely mentioned seeking or obtaining social support—a common practice used to cope with other forms of abuse [592, 502]—or broader institutional support. The small number of posts that did, mentioned filing police reports (20 of 176; 11%), reaching out to family or peers (19; 11%), or reaching out to the victim-survivors' workplace (6; 3%). A common reason for not seeking social or institutional support was embarrassment: *"I only told a few friends about what happened, but not with any specifics because it's so humiliating. I really have no one to discuss this with."*

4.5.3 Information: To understand & stop abuse

Nearly half of posters sought *informational* help (in 127 of 261 posts; 49%), with questions focused on making sense of the IBSA and determining how to make it stop.

What should I do? Many posts included open-ended requests (*"someone help me, i have no idea*

what to do"), such as from victim-survivors who appeared panicked as they tried to stop or recover from IBSA.

What's happening to me? Another type of request—especially from those experiencing recorded sexual assault—asked whether their situation constituted assault, abuse, or coercion. These requests sometimes included misunderstandings, such as how intoxication affects consent: *"Was I sexually assaulted? I was okay going to his place even though I was totally wasted. When I was there, he made me do things and I consent after he convinces me. Then he takes out his phone to take a video, which I never consented to. Did I make the wrong choice or was it just wrong?"*

What can I expect? Posters also asked what to expect in the near and long term. For example, for financial sextortion, victim-survivors asked if perpetrators would follow through on threats to distribute imagery, if they were "safe" after waiting a certain period of time, or how they would know if the abuse was "over." For nonfinancial sextortion and NCEI especially, victim-survivors asked about what would happen if the explicit imagery were shared. They sought help anticipating and mitigating future harm.

Why did they do this? Questions about perpetrator motivations were common (e.g., why they created synthetic explicit imagery, or pressured people into sexting). Multiple cyberflashing victim-survivors—especially women—asked why it was *so* common in online dating: *"Why is it that within a few days of talking, men always send pen*s pics?"*

4.5.4 Therapeutic: Coping with emotions

About a third of posters sought emotional or *therapeutic* help (in 86 of 261 posts; 33%).

Trauma. Victim-survivors asked for help coping with intense feelings, particularly in cases of recorded sexual assault and NCEI. They were overwhelmed by emotions like shame (*"I'm melting*

from the shame and guilt") and disgust (*"I'm disgusted and feel so violated"*) upon discovering the details of the IBSA, such as who images had been shared with, or who possessed (and thus had control over) the images.

Empathy, self-doubt, & isolation. Posters sought empathy (*"Can anyone provide reassurance or share their experience, anything helps"*) and help with processing their feelings (*"I needed to vent to someone"*). Some felt conflicted and experienced self-doubt in understanding their own experiences: *"I don't know how to feel, I don't trust anything."* For many victim-survivors, experiencing IBSA was an isolating experience. They described being ostracized from social connections: *"I only wrote this post because I wanted to talk to someone. I have no friends right now"* or not receiving support from institutions like their employer's human resources department, their school, or law enforcement.

Fear of confronting the perpetrator. Posters recognized that confronting the perpetrator could result in even more harm and asked for help in managing fears while planning next steps: *"I'm scared he'll be mad and share my pics if I confront him."*

4.5.5 Relational: Navigating relationships

About a third of posters sought *relational* help (in 84 of 261 posts; 32%) for navigating situations where the victim-survivor was socially connected to the perpetrator.

Romantic relationships. Some victim-survivors experiencing NCEI as part of romantic relationships struggled with whether or not they should stay with their perpetrator-partner. In these posts, "love" was cited as a reason to stay, and victim-survivors questioned their own behavior: *"We really love each other. However I also feel bad because he's always spending time with me. Is he stressed out because we spend so much time together? Would it have been better if I broke up with him?"* Posters also cited the central role their intimate partner played in their lives (*"he's my*

everything" and *"everything i've done was with his support"*).

Victim-survivors experiencing pressurized sexting and cyberflashing—typically when the perpetrator was an intimate partner—also asked about staying in the relationship (e.g., was the perpetrator's behavior a *"red flag"* or *"deal breaker"*?). They also commonly asked for help with how to effectively confront the perpetrator, such as to ask them to stop, resist pressure, or negotiate for images to be deleted.

Nonromantic relationships. When the perpetrator was socially connected to the victim-survivor, but not an intimate partner, such as a work colleague or friend, victim-survivors asked how to manage the relationship moving forward: *"My [22M] BFF [22F] flashed me and now won't talk to me. What should I do?"*

4.6 Help-Giving for IBSA

We now turn to comments on posts about IBSA, that is help-giving. Help-giving appeared to be community-oriented, as 55 help-givers (out of 2,298 total) supported multiple posters in our dataset, and all but seven of these commented within the same subreddit. Help-givers and help-seekers generally did not overlap (or they used throwaway accounts); only one account in our dataset created a post *and* commented on someone else's post.

4.6.1 Types of support and advice

Across our 261 posts, help was almost always given: only 15 posts received no comments, and the median post received five comments authored by four distinct users. However, few help-givers asked for more details to inform or tailor advice (14 of 160 threads). We discuss the most salient forms of help-giving below (Table 4.6 provides a summary; some posts and associated threads

Table 4.6: Types of help given across the seven types of IBSA. Some threads responded to posts discussing multiple forms of IBSA and were counted in multiple cells per row; some threads also included multiple types of help-giving.

Help Given	Description	FS	NFS	NCSEI	PS	CF	NCEI	RSA
Informational	General details that explain a type of IBSA, a perpetrator's motives, or directing to a resource for more information.	15	11	17	12	17	14	7
Technical	Advice for security and privacy, blocking, reporting, deletion, and whether to engage with a perpetrator's account.	9	11	12	12	11	11	6
Relational	Advice for navigating relationships, whether to inform family and friends, or whether to reason with perpetrators.	4	7	10	16	11	14	5
Therapeutic	Advice for taking time for yourself, not to be afraid, that it is not your fault, and that you are not alone.	6	8	11	14	9	13	9
Institutional support	Advice around law enforcement, legal options, therapists, support centers, or human resources.	2	2	4	3	9	9	7
Counterproductive	Advice that cast blame on the victim-survivor, or otherwise ignored the needs of a victim-survivor.	5	2	4	4	1	2	5
Other	Help-giving in other ways not described above.	0	3	0	1	0	4	2

contained multiple IBSA types, so the row totals are greater than the denominators reported below).

Informational. The primary form of help-giving was sharing information (72 of 160 threads; 45%). For financial sextortion, informational advice highlighted the scripted nature of the attack: *"This scam is super common. You can't do anything except block them (and stop sending d*ck pics to randos)."* For cyberflashing and pressurized sexting, threads explained the changing norms around sexting and the choices available to victim-survivors (e.g., blocking, reporting). For NCEI and recorded sexual assault, threads instead focused on reinforcing that the experience was abuse: *"Revenge porn is pretty much always considered illegal"* and *"This is definitely sexual assault."* Help-givers' tone when sharing information, however, was not always comforting: *"Even if you told me to imagine the most fucked up thing I could, I still wouldn't think of that"*. Furthermore, help-givers rarely pointed the poster to other communities or resources (5 of 160 threads; 3%). We discuss these limitations more in Section 4.7.1.

Technical. Technical advice was common across most types of IBSA (52 of 160; 32%). Some threads focused on not engaging with the perpetrator (21) or blocking and reporting a perpetrator's account (26), e.g., in financial sextortion, pressurized sexting, or cyberflashing, where a perpetrator was a stranger or acquaintance. A less prevalent alternative was telling the victim-survivor to make their own account private to minimize contact (7). Such advice was mostly absent from IBSA involving an intimate partner, like nonfinancial sextortion, NCEI, and recorded sexual assault. Technical advice in these scenarios focused on how to delete photos if the victim-survivor had access to a perpetrator's accounts or devices (5), how to check for backups (2), or how to find and record evidence of abuse (9): *"Screenshot the convo in which your gf got the video before you delete it."* Advice to block or otherwise avoid perpetrators aligns with highly recommended advice by experts to stay safer online from harmful content [597].

Relational. Threads frequently discussed relationship advice (52 of 160; 32%), but not always in well-reasoned ways. Threads commonly directed victim-survivors to end their relationship (29) across all types of IBSA other than financial sextortion, no matter the circumstances shared by the victim-survivor, including their living or financial situation, their desire to make a relationship work (*"I love him so much, and I know he loves me too"*), or whether or not leaving was a safe option. For example, one victim-survivor clearly stated their desire to stay with their partner, who was not the perpetrator of the IBSA, but was still told to leave: *"That's fucking me up to the point I would dump her."* Furthermore, help-givers at times assumed worse behavior from the perpetrator than was specified in the post: *"Any guy who pressures you for nudes will share them with his friends."* Some threads were more nuanced, however, laying out how the victim-survivor had agency in setting boundaries in cases of cyberflashing or pressurized sexting (20). Others advocated engaging with a perpetrator (10), particularly for NCEI and pressurized sexting, to negotiate deleting images or ceasing requests: *"Send him a message asking him to delete what he*

filmed, that he didn't ask for permission and broke your trust." Finally, threads discussed creating a support network to prepare for the risk of leaked imagery or to intercede with the perpetrator (12).

Therapeutic. Therapeutic help-giving was also common (52 of 160; 32%), such as reassuring victim-survivors that a situation was not their fault (15) and telling them to not be alarmed (14) or look after themselves (12). Threads also reiterated that a perpetrator was harmful (8) or commiserated with similar experiences (8). In the case of financial sextortion, threads focused on reassuring the victim-survivor that not engaging was the right tactic and to not be afraid: *"If you haven't sent any money, they have no reason to go after you."* Meanwhile, threads responding to NCEI, recorded sexual assault, and nonconsensual synthetic explicit imagery emphasized general support of the victim-survivor as they navigated abuse: *"I hope that you heal and find sweet and honest love again."*

Institutional support. While less frequent, help-giving also encouraged seeking institutional support (31 of 160; 19%). This advice focused on contacting law enforcement (15) or accessing legal advice (13), therapists (8), advocacy groups (3), human resources (3), or immediate medical support (1). Such advice was largely given for cases of NCEI (9), cyberflashing (9), or recorded sexual assault (7).

Counterproductive. Not all help-giving was supportive of victim-survivors (19 of 160; 11%). Some community members admonished the victim-survivor for their behavior, asking what they expected by sharing explicit images with others: *"Honestly if you're sending dick pics to online randoms, does it matter that they're getting leaked?"*

Others minimized or de-legitimized victim-survivors, asking *"What's the worst that can happen?"* or claimed most recorded sexual assault videos found online are fake. While the majority of upvoted threads were helpful (144 of 160; 90%), these examples show how help seeking on social media

can expose victim-survivors to additional shame or risk.

4.6.2 Victim-survivor reactions to help given

In general, the advice given on Reddit seemed to resonate with the poster. Of the 246 posts with at least one comment, at least 64 had one or more comments from the poster, that is, where the poster engaged in some way with help-givers.[10] The median engagement by posters was 50% of threads.

Posters expressed a range of reactions to help given, including providing or requesting more information (20 of 160; 13%), expressing appreciation for the help they received (16; 10%), and outlining a concrete plan of action based on the help they received (9; 6%). These reactions were spread roughly evenly across IBSA types. In many of these comments, the poster offered only a short thanks, even if there was no clear resolution. Some threads seemed to have helped them make sense of their experience: *"Thanks for confirming what I thought."* One back-and-forth sextortion thread helped the victim-survivor better understand the nature of the scam and feel reassured as a result: *"Thanks for the help and info, I learned something new."* In very few cases (4; 3%), posters expressed feeling hopeless even after receiving support: *"It's just so rough that there's nothing we can do."*

4.7 Discussion

We now reflect upon both the challenges in this space (Section 4.7.1) and possible roles of technology in solutions (Section 4.7.2). While our results provide insights into help-seeking and help-giving behaviors on Reddit, we argue that it is too early to speak definitively on the full

[10]This number is a lower bound: our quantitative analysis was based on our delayed crawl of the live site and thus could not identify engagement by a poster who had later deleted their account, which was a majority of posters.

set of challenges or solutions. Thus, we encourage readers to view our discussion here as a results-informed exploration of the possible parameters of the challenge and solution spaces.

4.7.1 Existing challenges

Limited nuance or formalized advice. Victim-survivors were offered emotional support in the form of concern and empathy (Section 4.6.1), making it clear informal help-giving on Reddit provides a lifeline to those who might otherwise be unwilling or unable to disclose and seek help. This suggests findings from previous related work on online help-seeking for offline sexual abuse [19, 20, 228, 395, 426] are replicated for technology-facilitated sexual violence. Namely, the anonymity and visibility management often afforded by online spaces facilitates support seeking, both explicit and implicit. As in previous work [396, 595], we also observed the constraints associated with online support, such as help-givers who defaulted to certainty around the situation or the "right" outcome without asking for more details or centering the expressed needs or desires of the victim-survivor, thus overlooking important nuances. For example, the point at which a victim-survivor leaves an abusive partner poses the greatest risk of harm to the victim-survivor [570], so gender-based violence advocates typically scaffold the creation of a safety plan [90, 153]. However, help-givers commonly suggested leaving a relationship without mentioning safety plans or other risk mitigation strategies. Relative to other types of help given, referring to institutional support was the least common (Table 4.6). Formal organizations[11] may be better suited to providing long-term and holistic support, but it is clear that connecting victim-survivors to support outside Reddit remains a challenge.

On the other hand, our work contains hints of why victim-survivors may have sought help on Reddit rather than via institutions. Between the time at which our Reddit snapshot was collected

[11]Examples for tech-facilitated abuse more broadly include https://www.ceta.tech.cornell.edu/ and https://techclinic.cs.wisc.edu/

and our crawl of the live site in December 2023, 99 posts and 146 poster accounts had been deleted. Further, 41 posts were from "throwaway" accounts, i.e., accounts not used again on Reddit, suggesting that posters valued Reddit's perceived anonymity or ephemerality. We also observed in Section 4.5.1 that posters most often sought help when in active crisis, and may have viewed Reddit as a faster way to get help than finding and reaching out to an advocacy organization. Of course, the fact that posters sought help on Reddit does not preclude them seeking help from other sources as well.

Underscoring challenges in technology-facilitated abuse. Our analysis of IBSA help-seeking on Reddit exemplifies two key challenges from broader technology-facilitated abuse literature. First, prior work emphasizes the weaponization of new technologies to expand the scope of potential targets [413]. For example, in most financial sextortion and cyberflashing posts in our dataset, the perpetrator was someone previously unknown to the victim-survivor before the IBSA (Sections 4.4.1 and 4.4.5); perpetrators leveraged new platforms and social discovery algorithms to find new victim-survivors. Another poster asked for preventative strategies against generative AI specifically, because they were concerned: *"with all the AI hype, many people around me are making nudes of others, especially of women."* More generally, the imagery was synthetic in all NCSEI cases and victim-survivors expressed distress over the ability of perpetrators to portray them in ways they had not consented to. Perpetrators will continue to leverage new technology, necessitating ongoing research to prevent its misuse.

Second, our findings reiterate the occurrence of polyvictimization, as studied in technology-facilitated violence literature [262, 362, 392]. Many victim-survivors described experiencing other types of abuse that co-occurred with the IBSA. Echoing research on intimate partner abuse [202, 573], supporting victim-survivors requires holistic and trauma-informed [106] approaches that do not regard technology as a panacea.

Aligning interventions with sites of harm. While some technological interventions exist for different types of IBSA, our analysis raises challenges for their reach. For example, StopNCII [554] helps victim-survivors get images removed from social media platforms, but in only nine (of 45) NCEI cases did posters mention that images had been posted on social media. More often, posters were concerned about perpetrators nonconsensually retaining images (18 of 45) or distributing them via messaging apps (14 of 45) (Section 4.4.6). Generally, IBSA incidents in our dataset were most frequently carried out via dedicated communication apps or direct messages on social media (143 of 247, see Table A.2 in Appendix A.2). Fewer cases involved images being posted on one-to-many platforms: social media (72), porn websites (6), or unspecified websites (3). Nevertheless, reporting IBSA to platforms remains one of the few technical mechanisms to remove explicit images stored or shared on a platform, or to potentially take action against a perpetrator's account and prevent others from being targeted. Thus, while reporting can be an effective intervention in these cases, our findings suggest an additional need to create contextually specific technological interventions.

4.7.2 Role of technology in solutions

Expanding and integrating the support ecosystem. Our analysis shows IBSA help-seeking on Reddit occurs in hundreds of thousands of posts, some with tens to hundreds of comment threads. This creates a burden for communities to triage and potentially leads to inconsistent advice depending on who responds in the moment. Recent advances in LLM agents could amplify help-giving: triaging incoming queries, requesting additional information as needed, providing guidance where best practices exist (e.g., for financial sextortion), and connecting victim-survivors with relevant advocacy organizations who provide hands-on support. Results from our study could guide the design of such an agent or other interventions, accounting for help-seekers'

common emotional states, questions, and needs (Section 4.5). Such an agent may not replace the interpersonal therapeutic support that help-seekers received through other people on Reddit, but could be more consistently available while complementing other available help-giving pathways. Early research on a non-generative AI agent is underway,[12] but this remains a ripe area for exploration.

On-device detection and warnings. Given the complexity of remediating IBSA *after* an explicit image is shared, technology could play a *preventative* role. Given the privacy sensitivities—particularly around messaging where much of IBSA occurs (Table A.2 in Appendix)—this technology would be best *on-device* to minimize explicit images being sent to platforms. To this end, Apple recently announced opt-in, on-device detection of sensitive (e.g., "nude") content [28]. While these technologies are in an early stage, they might take the form of nudges against sending explicit images (e.g., pressurized sexting, financial sextortion); or blurring explicit images upon receipt (e.g., cyberflashing). The effectiveness of these nudges—and expanding the detection capabilities to dangerous interactions (e.g., early detection of sextortion patterns)—remain to be explored. Smartphones could also extend current alertness-detection methods (e.g., as used in face biometric systems to assess whether a person is awake) to prevent or make more challenging the taking of nude or explicit imagery of people who are asleep or otherwise unaware.

Expanded controls around explicit content. Technology might also play a role in mediating how, for how long, and with whom explicit images are shared to reduce the risk of IBSA like sextortion and NCEI. Qin et al. discussed options such as disappearing messages, screenshot notifications, and watermarking to prove ownership or track the origin of a leaked image [462]. In practice, the effectiveness of these technologies hinges on the origin of an image (e.g., initially consensual vs. covertly recorded), and the willingness of perpetrators to adhere to tech-mediated

[12] See Umibot, https://umi.rmit.edu.au

norms. However, any increased friction for perpetrators can still reduce harm [476].

Audio-visual alerts around recording. Victim-survivors of NCEI shared that perpetrators often covertly captured explicit imagery. One intervention explored by device manufacturers in Japan and South Korea has been to emit a "shutter" sound whenever a cellphone records imagery [540]. This strategy is similar to the alert emitted by AirTags, intended to prevent stalking [418]. Such a feature could be paired with on-device detection of explicit content. Whether such a feature is acceptable to users, and to what degree it discourages NCEI, requires further investigation.

Preventing generative content. Specific to nonconsensual synthetic explicit imagery, generative AI technologies require safeguards to prevent their use in IBSA. Solutions likely require a combination of preventing harmful model outputs and also detecting synthetic explicit imagery.

4.8 Conclusion

We examined Reddit conversations about IBSA experiences, exploring the (1) types of IBSA for which people sought help, (2) the help they asked for, and (3) the help they received. After identifying over 100,000 posts through combined LLM and manual review, we qualitatively analyzed a stratified sample of 261 posts about seven types of IBSA: financial sextortion, nonfinancial sextortion, nonconsensual *synthetic* explicit imagery, pressurized sexting, cyberflashing, nonconsensual explicit imagery, and recorded sexual assault. We synthesized similarities in types of help sought and given, finding that across the seven IBSA types we studied, victim-survivors most often asked for and were offered information, empathy and therapeutic support, and advice about managing existing relationships. Technical, legal, and other institutional support were comparatively less common, indicating opportunities for more comprehensive support. Our work informs existing challenges towards mitigating, preventing, and supporting recovery from IBSA, and we outline

the role technology could have towards potential solutions.

Acknowledgements

We thank our anonymous reviewers for their valuable feedback and Nicola Henry for reviewing our search keywords.

Part II

Mapping Societal Factors

Chapter 5

SoK (or SoLK?): On the Quantitative Study of Sociodemographic Factors and Computer Security Behaviors

Researchers are increasingly exploring how gender, culture, and other sociodemographic factors correlate with user computer security and privacy behaviors. To more holistically understand relationships between these factors and behaviors, we make two contributions. First, we broadly survey existing scholarship on sociodemographics and secure behavior (151 papers) before conducting a focused literature review of 47 papers to synthesize what is currently known and identify open questions for future research. Second, by incorporating contemporary social and critical theories, we establish guidelines for future studies of sociodemographic factors and security behaviors that address how to overcome common pitfalls. We present a case study to demonstrate our guidelines in action, at-scale, that conduct a measurement study of the relationships between sociodemographics and de-identified, aggregated log data of security and privacy behaviors among 16,829 users on Facebook across 16 countries. Through these contributions, we position our work

as a systemization of a *lack* of knowledge (SoLK). Overall, we find contradictory results and vast unknowns about how identity shapes security behavior. Through our guidelines and discussion, we chart new directions to more deeply examine how and why sociodemographic factors affect security behaviors.

This chapter originally appeared as the paper "SoK (or SoLK?): On the Quantitative Study of Sociodemographic Factors and Computer Security Behaviors" at the USENIX Security Symposium in 2024 [596]. 'We' in this chapter refers to me and the co-authors: Jaron Mink, Yael Eiger, Tadayoshi Kohno, Elissa M. Redmiles, and Franziska Roesner.

5.1 Introduction

Sociodemographic factors — people's social, cultural, or demographic attributes (e.g., gender, race, socioeconomic status, age, or internet skill) — shape their lived experiences, i.e., what happens in their lives, how they are impacted by what happens, and how they make decisions. Prior works find that sociodemographics *do* impact computer security behaviors, e.g., that women may choose weaker passwords than men [375] or that older users choose stronger passwords [61]. These findings suggest that gender and age influence password selection and thereby computer security, potentially motivating interventions that target the underlying causal mechanisms.

The field of computer security has considered the role of the human for decades, e.g., Saltzer and Schroeder's recognition of the importance of psychological acceptability of security solutions in the 1970s [499], and Whitten and Tygar's foundational 1999 paper catalyzed the formation of the field of *usable security* [603]. Focus on the role of sociodemographic factors in computer security behaviors is, however, comparatively new [478]. Given the potential impact of these factors, it is vital to examine the current state of knowledge with respect to sociodemographics

and computer security behaviors.[1] With this understanding, it becomes possible to focus future efforts on addressing knowledge gaps and, ultimately, to help improve computer security for everyone.

Our first two research goals are:

- **Goal 1.** Collect and synthesize current knowledge about the quantitative relationship between sociodemographic factors and computer security behaviors.
- **Goal 2.** Enumerate existing knowledge gaps about sociodemographics and computer security behaviors.

Through a focused literature review of 47 papers in selected technical security conferences and a high-level survey of 151 papers in the wider literature, we synthesize trends, e.g., that people of different genders may focus on different security behaviors, as well as identify open opportunities for future research. We focus on quantitative studies, a primary method researchers use to measure security behaviors in relation to specific sociodemographic factors, like gender. Knowledge gaps exist when pertinent sociodemographic factors are omitted in analyses. For example, in our set of 47 papers between 1999 and 2023, we find that 38 consider (binary) gender whereas only 3 consider income, 5 consider race, and 9 consider Internet skill; none consider non-binary gender. We also observe different levels of depth with respect to how sociodemographic factors are analyzed and how differences by factors (if any) are interpreted.

After reviewing the current state of knowledge sociodemographics and behaviors, we identify our third goal:

- **Goal 3.** Formulate guidelines for future research on sociodemographic factors and security behaviors.

To demonstrate the use of our guidelines in practice and at-scale, we apply the guidelines

[1]For brevity, we use 'security' to consistently refer to security and privacy.

to conduct and report the results of a case study. Our measurement study uses de-identified, aggregated log data from Facebook to analyze the relationship between security behaviors on the platform and selected sociodemographic factors. We confirm several trends observed through our literature review, while adding nuance to others. Finally, we critically consider the knowledge gaps illuminated by our investigation — particularly, the lack of understanding about *why* sociodemographic factors and security behaviors might be correlated — and chart directions for future research.

5.2 Background and Motivation

Demographics uses statistics to study trends in human populations [456, 451]. Sociodemographics encompass demographic factors as well as social factors defined by formal institutions, e.g., governments [517], or informal institutions, e.g., sociocultural norms [497]. Conventional studies of sociodemographic factors are positivist, i.e., asserts that knowledge can be empirically measured and there *exists* a correct measurement that scientists can strive for [429]. In presuming objectivity, conventional demography overlooks the historical and political processes that shaped the categories themselves [277, 497, 532].

Categorization abstracts away richness to allow scientists to focus on selected characteristics. As an inherently reductive activity, categorization renders some research more tractable but may not accurately represent lived realities [64, 273]. By assigning people to static, finite groups, those who shift between groups or exist outside those groups, for example, will be systematically misinterpreted [50, 310, 512]. Further, categorization schemes are typically designed by historically and socially privileged groups in ways that can embed power imbalances [416, 538, 498, 451].

***Critical demography,* as an alternative to conventional demography, incorporates the**

reflexive study of how categories are socially constructed. As such, it "necessitates an open discussion and examination of *power* in society. Specifically, critical demography elucidates how power both affects and is impacted by demographic processes and events" [277]. Thus, critical demography offers a theory-driven paradigm to study how people behave, informed by social and political history [277], towards epistemological diversity and addressing inequity [429]. Prior work has applied critical demography approaches to deepen knowledge and practice, e.g., in computing education [421]. We apply a critical demography approach to synthesize prior work on sociodemographic differences in security behaviors, but also to map what is not yet known.

5.3 Literature Review Methods

What is currently known about how sociodemographics affect behavior, and what gaps remain? To scope to studies of users' actual security or privacy behaviors, we excluded studies of intended behavior, concerns, knowledge, or attitudes. As we were also interested in quantitative studies, we only included works that compared behavior between sociodemographic groups, i.e., we excluded work that investigated only one group within a sociodemographic factor.

5.3.1 Identifying Relevant Work

To identify potentially relevant studies, we defined unique search queries for selected conferences (see Section 5.3.2) and used the advanced search features of the ACM DL, IEEE Xplore, and the USENIX databases to search full-length research articles (see Table 5.1). Since these databases do not contain NDSS papers, we also obtained an NDSS paper archive scraped by other researchers. We wrote a `pypdf` [190] script to extract and search text directly from the PDFs using the search strings shown in Table 5.1. Two researchers independently reviewed paper titles and abstracts of search results to apply the scoping criteria described above and iteratively resolved disagreements

Table 5.1: Summary of literature review search methods. We used Google Scholar to write custom **search strings** to identify relevant studies in selected computer science **venues** and **databases** as well as in non-CS venues. For each search, we show the **number of results** and **number of included** studies satisfying our scoping criteria. †We implemented manual keyword searches of PDFs scraped from ndss-symposium.org. *We manually reviewed only the first 3 search result pages.

Venue	Database	Search String	# Results	# Included
FOCUS DATASET: Selected Computer Science Venues				
ACM CHI	ACM DL	In abstract: [security OR privacy] AND [behavior OR habits OR practices] and in body text: [gender OR sex OR age OR technical expertise OR education OR race OR culture OR internet skill]	213	14
IEEE S&P	IEEE Xplore	*same as CHI*	90	2
USENIX Security	USENIX.org	[security OR privacy] AND [behaviors OR habits OR practices]	41	2
SOUPS	ACM DL	*same as CHI*	77	17
ACM CCS	ACM DL	*same as CHI*	508	7
ACM CSCW	ACM DL	*same as CHI*	62	4
NDSS	ndss-symposium.org†	*same as CHI*	62	1
FULL DATASET: Beyond Selected Computer Science Venues				
Various	Google Scholar	[cross cultural OR large scale OR demographic] AND [behaviors] AND [security OR privacy]	270K+*	97
Various	Google Scholar	[password OR authentication OR update software OR secure drop OR phishing emails OR encryption OR WiFi OR anti-virus OR HTTP SSL warnings OR tracker blockers OR information disclosure OR self disclosure OR IoT OR VPN] AND [behaviors].	5.8M+*	11

to select the final dataset.

Relevant studies are also published in venues beyond computer science conferences; we used Google Scholar to find popular studies from any venue, including journals of business, information science, social science, or grey literature. We then defined two sets of keyword searches (see Table 5.1), which yielded over 6 million results, so one researcher reviewed the first 3 pages of

search results. Finally, we followed citations from papers in our dataset that referred to relevant work, adding 20 studies not identified through search strings. We set no explicit time boundaries for our dataset.

5.3.2 Full and Focus Datasets

Our final "full" dataset consisted of 151 works. Most papers were published in academic venues such as conferences or journals, but we also included 4 theses, 3 Pew Research studies, and 1 arXiv paper. The full dataset reflects a growing interest in security behaviors with respect to sociodemographic factors across venues and academic disciplines. Much of the dataset (76 papers) was in information science or social science domains and spanned a wide range of venues, from computing (e.g., *Computers in Human Behavior*) to communications and media (e.g., *New Media & Society*) to marketing and business (e.g., *Journal of Interactive Marketing* and *Journal of Management Information Systems*) to social sciences (e.g., SSRN). Another 20 papers were in computer science publications. The distribution of venues is long-tailed since we had 70 papers each from unique domains. We defined a "focus" set of 47 papers by identifying seven conferences most likely to include papers of interest: four computer security conferences (IEEE S&P, USENIX Security, CCS, NDSS), two HCI conferences with a tradition of including security and privacy (ACM CHI, ACM CSCW), and one conference at the intersection of HCI and security (SOUPS).

5.3.3 Qualitative Analysis

We qualitatively analyzed papers in our "full" set by coding behaviors studied (dependent variables) and sociodemographic factors considered (independent variables). For further analysis on our "focus" set, we also coded whether a significant relationship was (or was not) found[2] as well as

[2]Because studies sometimes studied multiple sociodemographic factors for a given behavior, we coded each relationship separately.

where the study was conducted, the research methods used, and any research sample limitations.

We created our codebooks via a series of iterative coding sessions between two coders. First, a primary coder prepared an initial codebook by inductively coding all papers. A second coder then independently coded a subset of the papers using the same codebook. The two coders then met to resolve inconsistencies and, if necessary, clarify and adjust the codebook. If adjustments were made, the primary coder then recoded the rest of the dataset. This process was repeated until the codebook no longer changed. We also verified our resulting codebook with a prior work's codebook on security behaviors [485]. Table 5.2 presents the behaviors codebook and paper counts; codebooks for other topics are presented in Appendix E.1.1.

5.3.4 Positionality

The authors' particular social, cultural, political, and historical context influence the way we discuss sociodemographic factors in this work. As researchers who have predominately lived and worked in the U.S., have English as a first language, and have had opportunities to pursue or achieve academic degrees in computer science, our perspective is limited by the privileges these experiences afford, relative to different experiences. Our motivation for this work is also shaped by experiences of marginalization by gender, culture, race, and age. As researchers with substantial experience studying human factors in security, we aim to improve the security of all people, not only those who have been historically prioritized in security research (i.e., users who are predominantly men, white, wealthy, highly educated, and live in the U.S.). We seek to raise the voices of those at the margins, in alignment with standpoint theory's premise that non-dominant social groups contribute critical knowledge towards scholarship and action towards justice [122].

5.3.5 Limitations

We sought as exhaustive a list of papers as possible to study sociodemographic factors and security behaviors, but we likely missed some relevant papers. During paper collection, we ultimately included less than 25% of our search results because many papers use sociodemographic keywords without satisfying our criteria, i.e., they do not quantitatively compare groups within a factor. Relevant work is also published outside the seven venues of our focus papers, but we believe our methods captured a set of papers large enough for meaningful analysis and discussion.

Our goal was to focus on sociodemographic factors related to security behaviors. As such, we scope to papers that measure security behaviors directly, e.g., through observational data, log data, and self-reports about actual behaviors. Future work may seek to focus on the substantial literature on additional topics, such as attitudes, opinions, or perceptions.

Since our goal was to conduct a formative literature review of security behaviors with respect to sociodemographic factors, we did not attempt to evaluate the "validity" of any paper. The replication crisis in psychology [531] reminds us that robust quantitative findings must be repeatedly tested and confirmed, which we leave to future work.

5.4 Literature Review Results

We first survey the locations, methods, and behaviors investigated in the full dataset of 151 papers (Section 5.4.1) and then explore methodological considerations in our 47 focus papers (Section 5.4.2). Next, we synthesize results about eight sociodemographic factors—gender, age, education, technical expertise, Internet skill, geography, race, and income—from the focus papers (Sections 5.4.3-5.4.8). Table 5.3 provides an overview of sociodemographic factors studied with respect to security behaviors and whether differences among groups were found.

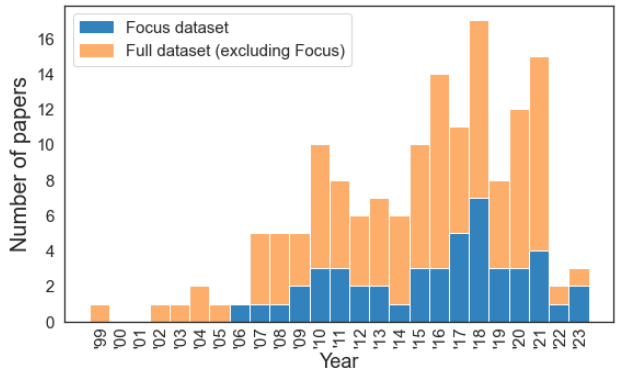

Figure 5.1: Number of papers in focus and full datasets investigating sociodemographics and security behaviors over time.

Throughout this section, we enumerate trends as **T#** and opportunities for future work as **O#** to facilitate later comparison with our measurement study results. We present a summary of all trends and opportunities in Table 5.4.

5.4.1 Survey of Full Dataset

Since 1999, the publication year of the first study in our dataset, papers on sociodemographics and security have steadily increased (see Figure 5.1).

Most studies are conducted in the U.S. or Western Europe O4. A plurality of studies in the full dataset (N=151) were conducted in the U.S. or Western Europe (60), followed by Asia (20), Africa and the Middle East (9), Australia (6), and South America (3).[3] Studies in the focus dataset (N=47) show the same trend: U.S. or Western Europe (33), Asia (5), Africa and the Middle East (3), Australia (3), South America (2).

Studies reflect a significant interest in network and web security behaviors as well as phishing and spam, but comparatively less in other behavior types O14. A complete breakdown of behaviors studied by papers in our full and focus datasets is presented in Table 5.2.

[3]Counts do not sum to 151 due to studies conducted in multiple locations.

Table 5.2: The number of security behaviors studied by papers in our **full** (N=151) and **focus** (N=47) dataset. Counts do not sum to 151 or 47 due to papers that study multiple behaviors.

Security Behavior	Full	Focus
Network and Web, e.g., VPNs, use private browsing, use anti-tracker tools	44	9
Phishing and Spam, e.g., phishing susceptibility	38	9
Social Sharing, e.g., disclosing info. on social media, changing privacy settings	35	11
Authentication and Accounts, e.g., password creation or reuse, 2FA	23	12
Device, e.g., anti-virus software, mobile lock screens	21	8
Composite: combined security behaviors from above that cannot be disaggregated	14	5

The largest portion of our full dataset investigated network and web-related behaviors (44 papers), such as using VPNs or avoiding public WiFi, or using private browsing and other anti-tracking tools or practices. The second largest portion investigated phishing or spam susceptibility, a much narrower set of behaviors than network and web yet studied by nearly as many studies (38 papers). Less commonly studied were social sharing behaviors, e.g., disclosing information or changing privacy settings on social media (35), authentication and account behaviors (23), and device-related behaviors (21). Finally, 14 papers studied composite security behaviors, or combinations of the above types of behaviors, with regression analyses that we could not disaggregate.

Security behaviors are primarily investigated through self-report methods O15. For our full dataset, the vast majority, i.e., 109 studies, were based on self-reports of security behavior. Less common were the 32 studies that used experimental measurement methods, e.g., in-lab or online experiments, or the 15 that used observational methods, e.g., log data or installed software on user devices. For the focus dataset, 32 were self-reports, 12 used experimental measurement methods, and 8 used observational methods. The significant preference for self-report methods likely reflects the relative convenience of collecting data from participants simply by asking them, but self-reporting may not be wholly accurate given participant biases, e.g., social desirability biases [477]. Future work should confirm these results with more ecologically valid methods.

5.4.2 Methods Considerations in Focus Dataset

For the focus dataset, we were interested in how prior work ensured that the research was robustly designed and conducted for the sociodemographics and security behaviors of interest.

A sizeable proportion of investigations did not mention sociodemographic factors until the results section of the papers O16. We analyzed focus dataset papers to determine if they motivated their study of sociodemographics in the research questions or introductory section (i.e., considered) or whether sociodemographic factors were mentioned only in the results section (i.e., post hoc). Because clearly defining independent variables, e.g., gender, and their levels or conditions, e.g., woman/man/nonbinary, is essential to control for confounding factors between conditions [460], not considering the role of sociodemographics until the results may imply an incomplete methodology. Further, consistently reporting the variables and relationships of interest before beginning statistical testing is important to avoid cherry-picking non-null results [377]. We found 74 instances where factors were considered in advance and 45 instances where sociodemographic factors were studied post hoc.[4] Studying sociodemographic relationships to security behavior is not warranted merely because relevant data was collected; therefore, confirming the results of papers that conducted a post hoc study of sociodemographics is an opportunity for future research to ensure that the findings were not spurious.

The majority of papers had limited samples, i.e., samples not representative of broader populations or balanced among groups O17. To assess the generalizability of studies in our focus dataset, we coded whether papers were representative (i.e., sample attributes matched broader population attributes), or balanced (i.e., equally across factor groups) for analysis. We found only 7 instances of factors being balanced and 18 representative; the vast majority (96) were

[4]Papers would have multiple instances if they studied multiple factors; thus, counts do not sum to 47.

limited,[5] i.e., not controlled in any way (snowball or convenience samples), or had no description of recruiting considerations. Future work should expand on addressing limited samples.

5.4.3 Gender

Gender is used in varying contexts and includes a person's gender, but also how gender is constructed in a societal context, i.e., referring to socially established gender roles [311]. It is often conflated with sex, i.e., bodily attributes, though these are distinct concepts [311]. Gender has received the most attention in security research (38 of 47 focus papers) relative to other sociodemographic factors.

Existing research primarily uses self-report methods, which could be biased by gendered differences in self-reporting O2. Prior work finds gender stereotypes that men are overconfident when it comes to security and privacy [598]. As described above, sociodemographic differences in behavior tend to be investigated through self-report methods, a trend that holds for gender specifically: of 28 papers that found gender differences, 21 used self-report methods. Thus, gender differences could be because men are more likely to self-report behaviors than women, regardless of true adoption rates.

Prior work suggests men may focus more on technical security behaviors, while women may focus more on privacy behaviors T1. One set of papers found that men were more likely to take certain protective actions related to network and web security, e.g., use tracker blockers [371], take protective actions against trackers [129], and use private browsing [236], although no differences were found for heeding SSL warnings [545]. Two papers studied composite security behaviors, finding that men were more likely to adopt predefined sets of security and privacy protective practices [630, 594] (although no differences were found in "triggers" prompting

[5] Counts do not sum to 47 because of papers studying multiple factors.

security behaviors [150]). Finally, one paper investigated how the sources for security advice differ between genders [481], finding that men were more likely than women to source advice from service providers. Taken together, these findings suggest gender differences in "technical" security behaviors, though it is unclear whether these differences result from self-reporting biases, prior computing experience, attitudes towards computers [607], or something else.

Another set of papers found that women were more likely to engage in security and privacy behaviors on social media and personal devices. Women were more likely to have private profiles on Facebook or Myspace [218, 556], post non-publicly on Facebook [193] and Snapchat [237], and avoid actions that expose online profiles they viewed [278]. Teen girls were found to be more likely than teen boys to adopt risk-coping behaviors (e.g., deleting posts, untagging photos, faking personal information) as well as seek privacy advice [295]. Prior work also found differences in disclosure content: men were more likely to disclose COVID-19 distress in May 2020 than women [624], but generally women were more likely to share memes portraying subjects positively [247]. Though prior work found inconclusive evidence about gender differences in device behaviors — adopting lock screens [241, 583], Android updating [370] — in others, women were found to be more likely to use webcam covers [354] and women 18-23 were more likely than men or people of other age groups to deny Android permission dialogs [60]. Taken together, these findings align with prior work (outside our literature review) [503, 356, 305, 567, 475] indicating that women focus on information protection and engage in privacy-preserving self-censorship.

Results were mixed on gender differences with respect to authentication and susceptibility to phishing and spam. Research on authentication behaviors is mixed: two papers found that men's passwords were stronger against offline attacks [61, 375], but men aged 46-49 were more likely to share passwords than others [309]; another paper did not find sharing differences by gender [435]. Further, researchers found that women were more likely to reuse passwords with

slight modifications [524], less likely to remember graphical passwords [109], and less likely to enable 2FA in response to experimental prompts [224], but other researchers found no difference in password reuse [443] or in whether they change password managers [404]. Future work should investigate whether gender differences in authentication behaviors are due to methodological differences, context, or other reasons **O8**.

Two papers found that women were more likely to click on phishing and spam [480, 527], although three other papers did not find significant gender differences in this regard [164, 323, 528], and one found that women were less likely to visit malicious URLs than men [523]. Papers not finding gender differences were published in 2006, 2007, and 2009, while papers finding differences were published in 2010 and 2018. One explanation could be that phishing and spam increasingly targeted women in 2010, and people of different genders now receive different types of phishing and spam [480] **O9**.

Existing research primarily investigates binary (assumed cisgender) individuals, excluding non-binary and transgender people O1. Most papers mention only women and men, and few papers conduct statistical testing with non-binary individuals, often opting to filter them out during data processing. Non-binary people constitute a far smaller proportion of study participants, posing a challenge for parametric statistical testing that could be resolved with use of nonparametric tests or different study designs. Further, almost no papers discuss transgender individuals, while other work conflates gender and sex by referring to participants as female and male when discussing gender, against best practices [511]. Research should distinguish between cisgender and transgender people only when relevant, but given that transgender people experience significant harm [584] and erasure [557], omitting this aspect of gender may reflect cisnormativity.[6]

[6] *Cisnormativity* is the assumption that everyone is or should be cisgender.

5.4.4 Age

Age granularity in the security literature varies from a single year to multiple decades and can be modeled as a numeric or categorical variable. Age is the second most studied sociodemographic factor (in 30 of 47 focus papers).

Prior work suggests that age may have been correlated with differences in password behaviors in the past, but is no longer T2. Two papers published after 2017 found no significant differences by age in switching password managers [404] or password reuse [442] (also found by a 2010 paper [524]). Supporting a theory of change in the past decade, a 2012 paper found that older users chose stronger passwords [61] but a study from the following year did not find such correlations [375]; similarly a 2011 paper found older users were more likely to share passwords [309], but evidence from a study seven years later did not support this finding [435],

Older users may behave more securely, while younger users focus on privacy T3. When studying a combined set of internet *security* behaviors, prior work found that older adults behaved more securely [594, 630], while younger users were more likely to adopt a combined set of *privacy* practices [630], e.g., use private browsing [236], use tracker blockers [371], and have Android lock screens [241] (though older users in Singapore may differ as they were more likely to adopt privacy practices [111]). The distinction between security and privacy behaviors may be partially explained by the finding that people 60+ were more likely to learn from automatic requirements or service providers than younger people [481]: formal sources may emphasize security as prevention of universal harm but privacy as a personal choice. Thus, older users were found to be more likely to enable 2FA in response to prompts [224] and deny Android permissions dialogs [60] as well as be more likely to be prompted by social triggers to behave securely [150]. However, differences by age were not found in responses to SSL warnings [545] or Android auto-updating [371].

With respect to online sharing, older users were more likely to post publicly than younger

users [193] though less likely to specifically disclose distress during the COVID-19 pandemic [624] or share a security news event [152].

Similar to mixed results for phishing susceptibility by gender, prior work presents inconclusive findings about the relationship between age and phishing susceptibility O10. Three papers found that younger participants were more susceptible to phishing [527, 323, 249], while two found no correlations by age [164, 528].

Table 5.3: Relationships between sociodemographic factors and security behaviors for papers in our focus dataset. For each sociodemographic factor (rows) and category of security behaviors (columns), we show X / Y, where X is the number of papers that found differences by factor for behavior, and Y is the total number of papers studying that factor and behavior. Summing counts do not sum to totals due to papers that study multiple factors or behaviors. Auth. = Authentication, Tech. Exp. = Technical Expertise, Composite = multiple behaviors studied together.

	Accounts & Auth.	Device	Network & Web	Phishing & Spam	Social Media & Sharing	Composite	Total
Gender	6 / 9	4 / 6	3 / 4	3 / 6	10 / 10	3 / 4	28 / 38
Age	3 / 9	2 / 3	2 / 3	3 / 5	5 / 6	4 / 4	19 / 30
Education	- / 4	- / 2	1 / 2	- / 2	- / 1	3 / 3	4 / 14
Tech. Exp.	3 / 6	2 / 3	4 / 4	4 / 5	0 / 0	1 / 1	7 / 12
Geography	1 / 1	1 / 1	2 / 2	2 / 2	2 / 2	2 / 2	10 / 10
Internet Skill	2 / 2	- / -	1 / 1	2 / 4	1 / 1	1 / 1	7 / 9
Race	- / -	- / -	- / -	- / -	2 / 3	2 / 2	4 / 5
Income	- / 1	- / -	- / -	- / -	- / -	2 / 2	2 / 3
Total	12	8	9	10	11	5	47

5.4.5 Education

Formal education imparts knowledge and skills to students and increasingly includes information about computing. Educational systems and institutions vary greatly, including nationally and internationally, but can be broadly grouped into primary, lower and upper secondary, and tertiary (also called higher ed) [579].

Education does not seem to be correlated with secure behavior T4. Of the 14 papers that investigated relationships between education and security behaviors, four found significant correlations: more educated users were more likely to delete cookies and history [107] as well

Table 5.4: A summary of trends and opportunities for future research from our literature review.

T#	Trends in Findings
T1	Women seem to focus more on information protection, while men seem to focus more on technical security.
T2	Older users may have had different password behaviors in the past, but no longer.
T3	Older users seem to exhibit more security-related behaviors while younger users focus more on privacy.
T4	Education does not seem to be correlated with secure behavior.
T5	Users with more tech expertise/use seem more likely to adopt technical security tools and take protective actions.
T6	Geography seems to be strongly correlated with differences in security behaviors.
O#	**Opportunities for Future Research**
Who is being studied: Lack of Focus around Specific Groups	
O1	Lack of research on non-binary and transgender people's security behaviors.
O2	Lack of research on gender differences in self-reported behaviors.
O3	Lack of research on education at levels besides secondary or post-secondary.
O4	Majority of papers conducted in U.S. and Western contexts; relative lack of research in other locations.
O5	Lack of research on geographical differences beyond granularity of countries.
O6	Lack of research on race and security behaviors.
O7	Lack of research on income and security behaviors.
What was found: Contradictory or Unclear Results	
O8	Mixed results on authentication behavior ~ gender.
O9	Mixed results on phishing and spam susceptibility ~ gender.
O10	Mixed results on phishing susceptibility ~ age.
O11	Mixed results on phishing susceptibility ~ internet skill.
O12	Mixed results on password behaviors ~ technical expertise.
O13	Unclear patterns of geographical influence on security behaviors.
How: Methodological Issues	
O14	Significant interest in network/web behaviors and phishing/spam, but less on other behaviors.
O15	Security behaviors are primarily investigated through self-report methods.
O16	Many papers did not declare an interest in sociodemographic factors in the motivation of the work.
O17	Most papers had limited sample generalizability.

as adopt a composite of 30 security, privacy, and ID theft practices [630]. However, studying a composite of four behaviors to combat viruses and hackers, Wash et al. find that compared to those with a high school diploma, those who did not complete high school were more likely to adopt security behaviors [594]. Similarly, compared to those who held a BA, those who did not hold a BA were more likely to report learning security advice from automatic software updates [481].

On the other hand, 10 papers do not find significant correlations between education and account sharing [435], password strength [608], password reuse [442], switching of password managers [404], Android auto-updating [370], webcam cover use [354], public sharing behaviors on Snapchat [237], SSL warning behaviors [545], or phishing susceptibility [164, 323].

Differences in correlations between education and security behavior are not well understood. There may exist several reasons for these disparate results. First, while prior work notes that those with lower educations are more concerned about being the victim of a computer scam, losing financial information, and being the target of harassment [357], it is unclear how the varying computer knowledge held by those with different educational backgrounds affects the ability to employ secure behaviors. Education does not necessarily include computing or security education; indeed, prior work found that while people with less education rely on less authoritative sources of security advice, they report fewer negative incidents, perhaps suggesting that formal advice sources — including formal educational environments — fail to provide effective security education [482]. Further, a bachelor's degree education varies significantly by institution, such that high-level education categories reduce critical nuances. Education may not result in a linear increase in security behavior but may vary by context. Future work should investigate the relationships between education and security behaviors to better understand the underlying causal mechanisms at play.

There is a lack of research on students at levels besides secondary or post-secondary **O3** **.** Of 14 papers studying education and security behaviors, four conducted studies in U.S. and Canadian universities (e.g., university students, staff, faculty) and another seven conducted studies with U.S. recruitment/crowdworker populations, which are more likely to have attended or completed college than the average [271]. Only the remaining three papers were not conducted in the U.S. or Canada, revealing a striking over-representation of Western university-affiliated users in education-related results. Future work should consider a wider range of educational levels, in different or outside of educational systems.

5.4.6 Technical Expertise, Use, and Skill

Aside from general education, users have varying levels of experience with technology (i.e., technical expertise) or the internet specifically (i.e., internet skill).

Users with more technical expertise may use more technical security tools and take more protective actions T5. People with greater technical expertise were found to be more likely to use private browsing [236], identify security threats [425], and cite school (as opposed to required sources or device prompts) as a source of security advice [481]. Those with computer and mobile skills were more likely to take defensive security measures [107]. Greater technical expertise was also associated with higher adoption of multiple security practices [289, 84], although no correlation was found between technical expertise and webcam cover use [354].

Relatedly, internet use may suggest higher adoption of security practices, e.g., users who logged in from multiple locations chose stronger passwords [61], and users more active on Facebook were more likely to enable 2FA in response to prompts [224]. Prior work also demonstrates that people with more internet skill cite different sources of advice [481], which may contribute to these behavioral differences.

We observe an inconclusive relationship between technical expertise and password-related behaviors O12. Unlike other security behaviors, technical expertise did not have a clear relationship with password choices. Two papers studied people affiliated with universities, one finding that participants in the computer science department chose stronger passwords than those in business [375], but the other found no conclusive evidence that technical expertise (including departmental affiliation) was correlated with stronger passwords [608]. Further, two papers found no correlation between technical expertise and password reuse [524] or password manager switching [404].

There is an inconclusive relationship between internet skill and phishing susceptibility

O11. Two papers found that greater internet skill or knowledge about phishing correlated with less phishing or spam susceptibility [480, 527], while three others did not find correlations between internet skills or phishing susceptibility [164, 528, 249]. A potential explanation comes from outside the literature review: prior work suggests activity level on a platform (which is itself weakly correlated to internet skill) may have more explanatory power than the coarser measure of internet skill [480].

5.4.7 Geography and Race

Geography is a proxy and umbrella term for a range of sociodemographic factors, including nationality, language, population density, political history, culture, internet penetration, freedom of speech, and more. Geography also shifts over time since politics and culture reshape the societies living between socially constructed boundaries.

Race refers to groups of people who share cultural, social, and physical similarities. It has been shaped through historical narratives of identity to be a tool of power, particularly for discrimination and the justification of colonialism [416, 538, 498]. Though racialized science continues to advance myths of biological differences between races, race is a powerful determinant of the privileges that an individual has access to, e.g., education, wealth, health.

All papers in our focus dataset studying geographic factors with respect to security behaviors were significantly correlated with behaviors T6, but effects lacked clear cross-cultural patterns O13. While the ten papers investigating correlations between geography and security behaviors find differences in many types of behaviors, these results are often sparsely populated, and it is not clear why these patterns appear or how they do, or do not, generalize to other geographical regions. German and French participants were found to be twice as likely to take protective actions against tracking than those in the UK [129]. Compared to U.S. and U.K.

users, German internet users tended to adopt more advanced, active privacy methods, such as proxies, Tor, and providing false information [128]. U.S. users were more likely to take security-protective actions because of proactive triggers, whereas people in India were more likely to act in response to social triggers [150]. Password strength varied by primary language spoken: passwords chosen by Indonesian-speaking users were found to be the weakest; German- and Korean-speaking users tended to choose relatively strong passwords [61]. Android lock screen usage also varied by country, e.g., 76.4% in the U.K. compared to 50.4% in Italy [241].

Phishing and spam susceptibility also differed by geographic location. Users who live in countries that have more spam are less likely to click it [480]. South Koreans were more likely to fall for phishing attacks in Korean than English, while Japanese participants were more likely to fall for phishing in English than Japanese [249]. On social media, rural U.S. users were more likely to set profiles to private than urban U.S. users [218], and Saudi women were more likely to block people on WhatsApp than Indian women [161]. Compared to U.S. users, U.K. users were less likely to dismiss cookie banners but more likely to not read consent text [63].

Few papers discuss geographical factors beyond the granularity of a country O5. Geographical factors describe a wide range of sociodemographic variance beyond national identity; however, the majority of papers focus only on these differences. Of ten papers, eight segregate geographical differences by nationalities [63, 129, 150, 161, 480, 241, 249], while only two discuss variations by language spoken [61, 249], and only one considers urbanization differences within the same country [218]. Future work can continue to illuminate how security behavior changes based on sociodemographics other than national identity, such as within a country, in cultures that extend beyond nations, or WEIRD vs. non-WEIRD societies [344].

Race is an infrequently studied sociodemographic factor in research on security behaviors O6. Race is a function of culture and was only studied in five papers. Trends are difficult to

ascertain because these papers investigated distinct behaviors and used different racial categorizations (we report using those papers' terminology). With respect to security, prior work found that white people were more likely to take certain protective security actions than Asian Americans and Pacific Islanders as well as Black or African Americans, though American Indians and Alaska Natives were more likely than white people to use security settings [594]. White people were more likely to solicit security advice from family and friends than Hispanic people [481]. With respect to privacy, while one paper found that racial minorities were more likely to publicly post on Snapchat [237], another found that compared to African Americans, white people were more likely to disclose distress on social media [624]. One paper did not find racial differences in Facebook profile privacy settings [435].

5.4.8 Income

Income determines not only the financial resources that one has to spend, but it may also indirectly influence the time or energy that one can put towards security behaviors.

More research is needed on relationships between income and security behaviors O7. Only three papers studied relationships between income and security behaviors: one found no differences in account sharing [435], while another found that people with lower incomes were more likely to adopt a combined set of security and privacy behaviors [630]. People at different income levels learn from different sources; those with higher incomes were more likely to learn from school, work, or device prompts [481].

5.5 Guidelines for Future Sociodemographic Research on Security Behaviors

Our literature review documents a significant and growing interest in studying how sociodemographic factors relate to security behavior. Based on our review, our own domain expertise, and sustained discussions amongst the research team, we developed guidelines to support strong and valuable research contributions. We iteratively refined these guidelines throughout our research process, including during our measurement study (see Section 5.6). We offer these guidelines to assist researchers in both their research and reviewing process. However, we caution: the guidelines are not a checklist to guarantee quality work, there may be cases when they do not apply, and norms and best practices continually evolve.

Factor Selection. The selection of which sociodemographic factors to analyze should be done deliberately and stated as a research interest in the motivation (e.g., in the introductory section) for the work. Many papers in our literature review did not explicitly declare studying sociodemographic differences but presented correlations with sociodemographic factors in the results (see Section 5.4.2), which may indicate spurious correlations **O16**. Additionally, multiple studies are necessary to establish robust evidence of factor correlations, as demonstrated by the replication crisis in psychology research [374, 531].

G1: Identify at the beginning of the study the specific sociodemographic factors, if any, you intend to study. If you investigate sociodemographic differences, commit to reporting the results even if they do not show differences, i.e., null results. Consider study pre-registration [119].

Group Selection. Within all sociodemographic factors, there are groups that are privileged or

marginalized. We found many opportunities for research about different groups, e.g., groups marginalized by gender (see Section 5.4.3) **O1** or race (see Section 5.4.7) **O6**. Researchers choose to study a subset of groups for practical or other reasons. If so, describe how the scope was chosen and how the sample studied relates to the broader population.

G2: Consider and justify which groups are included in or excluded from your study.

Method Selection. Epistemic diversity allows researchers to explore a wider range of research questions. Consider research methods that make different types of contributions [613], including but not limited to: quantitative, qualitative, or mixed methods [330]; descriptive, experimental, or speculative; cross-sectional or longitudinal [215, 131]. If relevant, consider causal inference methods [441, 142].

Most papers in our literature review used statistical hypothesis testing, which is primarily valuable to identify factors for correlations but not causation. Few papers in our literature review modeled sociodemographic factors as control factors (see Cho et al. as an exception [111]). Further, many papers we reviewed used self-report methods **O15**, which are convenient for formative work but not suitable for establishing robust results.

G3: Consider using diverse research methods, acknowledging the benefits and limitations of each.

Result Interpretation. When interpreting results, remember that complex factors could lead to any observed differences. Avoid "essentializing" (reducing individuals to assumed group characteristics) and over-generalizing findings. In interpreting results, state not only what can be implied from the results, but also what cannot: for example, "We found a significant correlation between this factor and this behavior, which might be due to methodological choices or factors outside the scope of this study." This is particularly important for studies conducted on limited

samples O17.

G4: When sociodemographic differences are observed, exercise caution in describing the results. Consider posing several causal interpretations for observed differences.

Author Positionality. Weighing the advantages and disadvantages of disclosure [343], if appropriate and safe to do so, include positionality statements in your work. In some cases, the risks to researchers may not merit disclosure. Further, we caution against positionality statements that merely list identities without reflexivity as to how these identities influenced the research process. When included thoughtfully, such statements provide context for readers about researcher motivations and the potential influence of researcher backgrounds. For example, a majority of existing research in our literature review is U.S.-centric and studies people affiliated with universities O3 and O4, which is likely the result of the (undiscussed) positionality of researchers as primarily professors and graduate students at Western universities.

G5: Be aware of your own positionality and identity as a researcher and critically reflect on how it might affect your research questions, hypotheses, and interpretation of your findings [277].

5.6 Case Study: Measuring Sociodemographics and Security Behaviors on Facebook

We now instantiate our guidelines in our own case study to concretely demonstrate their application for future researchers. We iteratively refined the guidelines in the process, resulting in the version in Section 5.5.

Unlike most prior work that uses self-reports, we leverage de-identified, aggregated log data to

shed light on how users' real security behavior correlates with sociodemographic factors. Security is often a secondary goal, so users may incorrectly recall actions and self-report based on social desirability [408, 318] or researcher demand [428]. Thus, real-world security behavior offers high ecological validity and an important complement to self-report studies.

5.6.1 Measurement Methods

This study was conducted by combining de-identified, aggregated log data about security behavior with the results of a 16,829 respondent survey run on Facebook in 16 countries during December 2019. Respondents were recruited through both web and mobile interfaces via a message at the top of their social media feeds. The survey was translated into the respondent's local language by professional translators with native language proficiency.

Our *dependent variables (DVs)* were four security behaviors: **Security settings visited** (± 45 days of survey date), **Security settings acted on** (± 45 days of survey date, only among respondents who visited), **2FA enabled** (ever), and **Stronger password** (i.e., those not yet identified as potentially more vulnerable to attack[7]).

Based on the available log data about Facebook users, we chose six sociodemographic factors to study **G1**. These factors were: **Age**, **Gender** (binary[8]), **Educational attainment**, **Geographic location** (16 countries in four regions), **Internet skill**, and **Technical knowledge**. Appendix E.2.1 details these factors and how they were determined. We also included four available *independent variables (IVs)* regarding Facebook use based on the de-identified platform data: **Tenure** (how long the respondent had an account), **L30** (how many of the last 30 days the respondent had logged in

[7]See https://www.facebook.com/help/124904560921566 for details on Facebook's password guidelines and https://www.facebook.com/notes/760840091433907/ on identifying potentially vulnerable passwords.

[8]Due to cross-cultural differences in prevalence of non-binary gender reporting, we study only those who reported a binary gender to allow for interpretable comparisons across countries. As underscored in **O1**, we encourage future work on those of non-binary genders.

to their account), **Time spent** (how much time the respondent spent on the platform over the last 30 days), and **Friend count** (number of social connections on the platform).

Analysis. We analyzed our data with logistic regression models because of the scale of our data **G3**. We weighted our sample to represent the population of the broader social media platform on age, gender, tenure, and L30 in order to maximize the generalizability of our results. To examine the relationship between security behavior and our independent variables, we constructed weighted logistic regression models, with security behavior as the boolean DV and the other variables listed above as the IVs (see Appendix E.2.2). We also controlled for two interactions that had correlations with $\rho > 0.3$: *l*30 $*$*time spent* — there is a correlation between the number of days and amount of time spent on the platform — and *location*$*$*tenure* — there is a correlation between geographic location and platform tenure since the platform was introduced to different markets at different times. Regression models were fit using 5-fold cross validation. The variance in AIC between the five folds was always less than 3%.

Limitations. Our measurement study considers users of only one social media platform, although this platform is one of the largest and most diverse online platforms. Though we studied users in 16 countries, this represents a minority of countries globally **G2**. Further, racial categories differ greatly by sociocultural context, which is why our measurement study across 16 countries did not study race **G2**.

Ethics. We analyze de-identified, aggregated log data of users on Facebook who voluntarily completed survey data. There was no manipulation of any user's experience, and no personal identifying information was used. Our research procedures were vetted and approved through an internal review process.

Positionality. We echo our positionality statement from Section 5.3.4 in conducting this measurement work **G5**. Additionally, we note that one author engaged in a paid collaboration with

Meta, which allowed them to access and analyze the de-identified, aggregated log data.

5.6.2 Measurement Results

In interpreting our results, we emphasize that all findings describe only associations between sociodemographics and behaviors, and we do not make causal claims **G4**.

Gender: On Facebook, women were more likely than men to take action regarding security settings, but no gender differences were found with respect to password strength or use of 2FA. We do not find significant differences in likelihood to *visit* security settings, but women were more than 1.4 times as likely to *action* security settings than men ($OR = 1.44$, $p < .01$). These results may support **T1** if actioning security settings is interpreted as an information protection behavior. Given that other work on the same platform we study finds that people tend not to make a clear distinction between security and privacy [479], women actioning security settings would align with other information protection behaviors. We found no significant differences in likelihood to have a stronger password or use 2FA by gender. While this null result could mean there is no relationship between gender and these behaviors, it could also mean that Facebook users are unique in not having gender differences, but differences could be found in studies of users of other services.

Age: While older Facebook users were less likely to visit their security settings, action their security settings, and use 2FA, those age 50+ were more likely to use stronger passwords. Compared to those aged 25-34, older adults were significantly less likely to *view* their security settings, with the odds of those between 35-49 being 0.74 times as likely to visit their security settings ($OR_{35-49} = 0.74$, $p_{35-49} < .05$) and those 50+ being 0.63 times as likely ($OR_{50+} = 0.63$, $p_{50+} < .05$). We also found significant differences in their use security settings, with the odds of older adults *actioning* their security settings being lower ($OR_{35-49} = 0.55$, $p_{35-49} < .001$;

$OR_{50+} = 0.38, p_{50+} < .001$) and the odds of them using 2FA being lower, as well ($OR_{35-49} = 0.79, p_{35-49} < .05$; $OR_{50+} = 0.63, p_{50+} < .05$). However, the odds of adults 50+ having a stronger password was higher ($OR_{50+} = 2.08, p_{50+} < .05$). These findings appear to support **T2**, i.e., that older adults are more likely than younger to adopt security behaviors like passwords.

Education: On Facebook, education levels were correlated with the likelihood of using 2FA. Compared to users with no post-secondary education, users with some college ($OR = 7.14, p < .01$) or a bachelor's degree or more ($OR = 5.40, p < .01$) were more likely to use 2FA. However, education *was not* correlated with visiting or actioning security settings or having a stronger password, in alignment with **T4**.

It is possible that users with higher educational levels had to previously comply with their institution's 2FA IT policy and thus were more likely to reengage with 2FA on Facebook. Those with higher educations may also be more comfortable with computer systems and security tools like 2FA. Future work can continue to investigate post-secondary institution's influence on 2FA adoption by comparing with users not affiliated with post-secondary institutions, towards **O3**.

Technical expertise: On Facebook, technical expertise was correlated with stronger passwords and 2FA use. Technical knowledge of passwords was correlated with having a stronger password ($OR = 1.88, p < .05$) and using 2FA ($OR = 1.33, p < .05$), as was knowledge of the reaction feature on Facebook ($OR = 1.75, p < .05$; $OR = 1.37, p < .01$for stronger password and 2FA, respectively). Knowledge of QR codes was also correlated with greater use of 2FA ($OR = 1.49, p < .001$), while knowledge of downloads was not correlated with any security behavior. Since downloads are the oldest technology feature we asked about, the trends we find in our measurement study seem in alignment with our literature review, i.e., that technical expertise correlates with increased likelihood to take secure actions **T5**.

Internet skill: On Facebook, internet skill was correlated with all behaviors except

having a stronger password. Internet skill was correlated with visiting ($OR = 1.41, p < .01$) and actioning ($OR = 1.44, p < .05$) security settings as well as using 2FA ($OR = 1.84, p < .001$).

Platform-specific use: Tenure on a platform was correlated with all security behaviors, while use in the past 30 days was not correlated with any. Platform tenure in years was correlated with all four security behaviors, specifically to be less likely to visit ($OR = 0.95, p < .01$) or action ($OR = 0.95, p < .05$) security settings, less likely to have a stronger password ($OR = 0.91, p < .05$), but more likely to use 2FA ($OR = 1.84, p < .001$). This may be due to those with longer standing accounts having already adjusted their settings and due to changes over time in password advice (those creating accounts earlier may have received less password education at the time of account creation). Friends and time spent were also positively correlated with use of 2FA ($OR = 1.02, p < .01$; $OR = 1.13, p < .05$), though use in the last 30 days was not correlated with any security behaviors.

Geography: On Facebook, users in Africa, the Middle East, and Asian geographic markets differed significantly from the Western market in terms of security behavior. The odds of users in Asia visiting security settings were higher than users in the West ($OR = 1.94, p < .05$), lower compared to the same group to have a stronger password ($OR = 0.16, p < .001$), and no different for actioning security settings and using 2FA. Users in Africa and the Middle East were less likely to have a stronger password ($OR = 0.24, p < .05$), but other behaviors were not significantly different. Users in Latin America were not significantly different from Western users on any of the four security behaviors we studied. Geographic differences in our case study broadly align with **T6**, i.e., that geographic differences are significant but with unclear patterns **O13**.

5.7 Discussion

Having presented a systematization of knowledge of sociodemographics and security behaviors (Section 5.4) and guidelines for researchers (Section 5.5) and applied them in our own measurement study (Section 5.6), we now critically consider our *lack* of knowledge in this space.

5.7.1 The Missing "Why?"

This work reveals many correlations between sociodemographic factors and security behaviors, but little insight into *why* these correlations exist. The trends we synthesize and the opportunities we highlight begin to pose hypotheses for underlying causal relationships, but much work remains. Without understanding why, interpreting results becomes arduous and different studies can yield seemingly contradictory results. For example, when correlational studies find that one sociodemographic group adopts a security behavior less than another, is this the result of sample differences, different threat models, user interface designs that assumed one group as "default" users [130], or some other reason? Sociodemographic factors also do not exist in isolation but are correlated and influence each other; though this work did not investigate (reflecting papers in our literature review) intersectionality [137], only through understanding *why* each identity influences behavior can intersectional analyses be conducted.

Drawing implications for interventions to change behavior when the *why* is still missing is a tenuous proposition. We can neither know what interventions might encourage adoption of security behaviors nor whether such interventions are necessary, desired, or even helpful. Worse, if we assume incorrectly, subsequent actions or discussions may have a negative impact, e.g., perpetuating gendered stereotypes about computer security and privacy behaviors [598]. Recent related work on underlying causes of differences in security threats, rather than on behaviors, takes initial steps toward a causality-focused framework, e.g., identifying higher-order

factors like prominence and marginalization that put particular groups at higher risk of security threats [592, 506].

5.7.2 Towards Answering "Why?"

What should be next for this field of research on sociodemographics and computer security and privacy? To close the knowledge gap, future work should explore not only *what* differences exist among sociodemographic groups in security and privacy, but *why* these differences exist.

Epistemic Diversity of Methods. Seeking to understand the causal relationships underlying sociodemographics and security behaviors cannot be achieved solely through quantitative methods. That does not mean establishing correlations has no value; the field as a whole must grapple with *why*, and individual papers provide incremental steps towards an answer.

In addition to the inferential methods used in the quantitative papers we analyzed, qualitative methods, e.g., in-depth interviews, observational studies, and ethnographies, can be used to explore the missing *why*. Such methods are increasingly used in security and privacy research to study the needs and practices of specific marginalized and vulnerable user groups but should also be used to draw out the underlying sociodemographic factors and their relationships to behavior. Especially by critically comparing privileged and marginalized groups, qualitative methods can assess existing hypotheses about causal relationships or pose new relationships and mechanisms of effect.

There are also other quantitative methods to consider beyond correlation and regression analyses. For example, structured equation modeling (SEM) involves constructing a model with causal relationships and statistically evaluating relationships as well as effect magnitudes [577]. Other analyses include causal inference methods [440], Bayesian methods [319], or quantitative meta-analyses. Each method has strengths and limitations that future work can explore.

Towards Social Theories. Overall, we recommend that security and privacy researchers learn from other fields that rely on social theories. Social theories, i.e., scientifically plausible principles that seek to explain certain phenomena by posing causal hypotheticals, pose richer explanations for how people behave, which avoids essentializing a group of people. For example, when women take fewer security measures than men, some might interpret this to mean that men are fundamentally better suited to security tasks. Instead, women's choices may reflect systemic educational inequities, where women were discouraged from learning about technical topics, or other reasons. Research must be careful to avoid attributing differences to innate group characteristics, e.g., racial essentializing [498]. Relevant social theories support robust interpretation when differences are found and can also indicate how a lack of difference can be meaningful.

Social theories from other fields can also help illuminate gaps in security behavior research, i.e., understudied factors that also merit study. Papers we analyzed focused most often on factors such as gender and age, but factors that have been more deeply studied in other fields and could inform security research include (dis)ability, marital status, religion, migration status, socioeconomic status, and race.

Finally, social theories facilitate critically challenging assumptions inherent in some perspectives on sociodemographics and security. This includes (1) questioning whether certain security behaviors are desirable for certain groups and in certain contexts since "spending more time on security is not an inherent good" [266]. Further, (2) the security behaviors studied may not address (or be trusted to address) the needs of all communities, especially those most marginalized [379, 605], and (3) sociodemographic categorizations themselves, and the types of security behaviors studied, are not the only ways to organize the space and may not be the most salient to users. As research continues to explore sociodemographic differences in security, incorporating theoretically informed inquiries presents the greatest opportunity to build on current methods and knowledge.

5.8 Related Work

Qualitative Studies of Marginalized Populations in Security and Privacy. A sizeable and growing body of literature investigates the experiences, behaviors, and needs of populations underrepresented in security and privacy research. These works often overlap with sociodemographic factors, e.g., targets of intimate partner violence [575] who are disproportionately women, refugees [533], LGBTQ+ individuals [212], and Muslim-American women [503]. Other studies investigate vulnerable populations due to their work, e.g., journalists [386], content creators [566], and sex workers [379]. These studies have been overwhelmingly qualitative, i.e., providing rich insights rather than quantitatively generalizable results.

Meta-Analyses of User Studies. The need for cross-study synthesis grows as the number of user studies of security behavior increases, e.g., about which methods are common [167], or expert vs. non-expert users [306]. Prior meta-analyses also investigated marginalized [506] and at-risk users [592], specifically developing unifying frameworks. We focus on sociodemographics as the unifying frame because they are a powerful latent cause of differences; ultimately, marginalization relies on our contemporary, socially constructed sociodemographic categories. Aside from a recent preprint investigating geographic diversity in security and privacy research [248], we are unaware of other meta-reviews taking a sociodemographic lens, though sociodemographic meta-reviews in HCI are more common, e.g., culture [304] as well as gender, race, and class [513].

5.9 Conclusion

We broadly survey scholarship (151 papers) that quantitatively studies sociodemographic factors and computer security behaviors, and we synthesize methods and results in a focused review of 47 papers. Taking a critical demography approach, we enumerate five trends in existing research and

fifteen opportunities for future research (Table 5.4). We establish five guidelines for conducting quality sociodemographic research investigating security behaviors (Section 5.5) and apply those guidelines in a case study of the real security behaviors of 16,829 Facebook users. Taken together, this work documents the current state of knowledge on how people's identities relate to the security and privacy actions they take and charts new directions towards greater security, privacy, and equity.

Acknowledgements

We thank our reviewers and especially our shepherd for their helpful feedback. We are also grateful to Alannah Oleson, Matthias Fassl, and the UW Security and Privacy Research Lab (including Rachel Hong, Alexandra E. Michael, Christina Yeung) for insightful conversations on framing, methods, and impact. We thank Maximilian Golla and Aleksei Stafeev for sharing an NDSS paper archive. This work was supported in part by the U.S. National Science Foundation under Awards 2205171 and 2206950 and the Graduate Research Fellowship Program (DGE-1746047). The fifth author did a portion of this work while working as a contractor for Meta.

Chapter 6

Skilled or Gullible? Gender Stereotypes Related to Computer Security and Privacy

Gender stereotypes remain common in U.S. society and harm people of all genders. Focusing on binary genders (women and men) as a first investigation, we empirically study gender stereotypes related to computer security and privacy. We used Prolific to conduct two surveys with U.S. participants that aimed to: (1) surface potential gender stereotypes related to security and privacy ($N = 202$), and (2) assess belief in gender stereotypes about security and privacy engagement, personal characteristics, and behaviors ($N = 190$). We find that stereotype beliefs are significantly correlated with participants' gender as well as level of sexism, and we delve into the justifications our participants offered for their beliefs. Beyond scientifically studying the existence and prevalence of such stereotypes, we describe potential implications, including biasing crowdworker-faciliated user research. Further, our work lays a foundation for deeper investigations of the impacts of stereotypes in computer security and privacy, as well as stereotypes across the whole gender and identity spectrum.

This chapter originally appeared as the paper "'Skilled or Gullible? Gender Stereotypes Related

to Computer Security and Privacy" at the IEEE Symposium on Security and Privacy in 2023 [598]. 'We' in this chapter refers to me and the co-authors: Pardis Emami-Naeini, Franziska Roesner, and Tadayoshi Kohno.

6.1 Introduction

Stereotypes are reductive beliefs about social groups, e.g., people of a certain gender or age. Gender stereotypes have been widely studied in numerous areas of society (e.g., medicine [521], law [589], education [314], politics [183], STEM [551]) and have documented impacts on a multitude of attitudes and behaviors. For example, researchers in other domains have found that gender stereotypes can significantly alter behavior by boosting or hindering self-efficacy, i.e., an individual's belief in their ability to achieve their goals [79, 324, 355]. In STEM, stereotypes also have adverse consequences, e.g., on girls' interest in computing [369].

Given the widely-documented existence of gender stereotypes and associated harms in other domains, we hypothesize that gender stereotypes exist for computer security and privacy, contributing to gender inequities. However, these issues have not been rigorously studied, leaving open questions about how gender stereotypes manifest in our field. This work provides a critical theoretical foundation for understanding gendered differences in attitudes and behavior, and thus exemplifies how gender analysis can foster scientific discovery [561] in security and privacy.

We investigate what specific security- and privacy-related gender stereotypes exist and how widely they are held. Our research questions are:

1. What gender stereotypes (about women or men) do members of the general U.S. public hold that concern everyday computer security and privacy issues?
2. What explanations or rationales do people give to justify gender stereotypes?

Though we do not aim to compile a comprehensive list of stereotypes with respect to computer

security and privacy, our investigation lays the necessary groundwork to study the harms of specific stereotypes. Further, we investigate the rationales for stereotypes in order to inform efforts to combat stereotypes and mitigate their impacts.

The computer security and privacy research field must ultimately consider gender beyond the binary to contend with gender's full multiplicity [311]. We begin by investigating binary genders in order to build on existing research instruments on sexism, which primarily consider binary genders, as well as our own experiences and identities. Further, we note that considering gender as a binary is itself a widely held stereotype [417, 214].

Contribution one: specific instances. Through a pre-study of 202 U.S. Prolific participants, we surface specific instances of potential gender stereotypes with respect to computer security and privacy. These reside in three categories — general engagement, personal characteristics, and specific behaviors — and lay a foundation for our next phase.

Contribution two: quantitative evidence. We provide quantitative evidence that people hold gender stereotypes about computer security and privacy. Among other results from our second, 190-participant Prolific study, we find that:

- Men were expected to be more engaged with security and privacy topics, including being more skilled at protecting their security and privacy. Women were expected to be gullible and emotional about these topics.
- Participants believed men were more likely than women to behave in security- or privacy-enhancing ways, e.g., to verify HTTPS, install software updates immediately, and enable two-factor authentication.
- Most negative stereotypes we observed were negative towards women, but we also found negative stereotypes towards men: e.g., participants expected men to be more overconfident and less likely to ask for help.

Furthermore, we found beliefs correlated with other factors:

- Many stereotypes were held by both women and men — including negative stereotypes about women.
- Sexism (measured with the validated ASI scale [221]) was strongly correlated with belief in gender stereotypes with respect to computer security and privacy.

Contribution three: characterizing rationales. In order to combat stereotypes, we must first understand why people hold such stereotypes. To sample our findings:

- Many rationales were adapted from gender stereotypes outside of computing: "Men are more likely to be more logical when it comes to computer security and privacy because men are natural born problem solvers and always try to explore the best possible means to fix a problem" (P189).
- Other stereotyping rationales included flawed reasoning about biology: "Women are less biologically driven to use technology and thus may not be as aware of the risks of sharing too much information online" (P31).

In addition to characterizing the rationales to combat stereotypes, surfacing these rationales deepens our understanding of how people evaluate and manage their own security and privacy, as well as how people view others.

To conclude, we reflect upon our findings and make recommendations for system designers and researchers by compiling ten guidelines for the future. We make suggestions for short-term work to be conducted to study the implications of gender stereotypes, as well as for long-term efforts to combat gender stereotypes in research and design processes. In particular, we highlight how gender stereotypes could bias the results of user studies conducted with crowdworkers. Though potentially linked, we advocate distinguishing gender stereotypes from empirical measurements of gender, attitudes, and behaviors, because stereotypes cause harm regardless of the status quo.

Ultimately, we hope this work validates the experiences of people who have been at the receiving end of gender stereotypes in security or privacy, and serves as a call to action to combat these stereotypes.

6.2 Related Work

6.2.1 Gender in security and privacy research

Prior research in computer security and privacy has found that gender can be a contributing factor in security behaviors, e.g., password choice [376], usage of private browsing [236], usage of two-factor authentication (2FA) [455], interpreting security warnings [22], susceptibility to phishing [238, 527], as well as in security intention [230], attitudes [182], and risk perception [208]. Privacy research has also found that gender may influence self-disclosure on social media [158, 333, 471], information disclosure generally [500], protection strategies [436, 427], or privacy concerns [525]. Most of these prior works do not primarily focus on gender; instead, they include it among other demographic factors. Our focus in this work is not on the direct study of gender differences in security and privacy behaviors, but the biased assumptions and stereotypes that people hold about them — which may play a role (alongside other factors) in disproportionate adoption of security and privacy behaviors by gender (e.g., due to stereotype threat, a psychological threat of confirming negative stereotypes [549], and the barriers they form [40, 522]).

Other security and privacy work focuses on gender through the lens of specific marginalized populations, e.g., the cultural context of women in South Asia [503, 501] or ways women are vulnerable, e.g., as survivors of intimate partner violence [253], users of menstruation [389, 287] or women-specific apps [620], or victims of gender-based harassment [565, 53]. Our work studies gender through a different specific lens, i.e., U.S.-based internet users.

6.2.2 Gender stereotypes

Gender is a social construct that exists distinct from, but may be related to, biological differences between women and men [154, 85]. In many societies, gendered expectations exist about the ways that women and men should be [154, 42] and manifest as cultural stereotypes. An abundance of research continues to theorize about the creation and reinforcement of gender stereotypes (e.g., [313, 300, 195]). The Stereotype Content Model posits stereotypes are composed of two dimensions: competence and warmth [195]. Decades of research study stereotypes that men are more competent but women are interpersonally warmer (e.g., [195, 140]).

Stereotypes create two classes of implications: distorted perceptions by stereotype holders, i.e., for "perceivers", and the experience of targets, i.e., for "experiencers" [141]. With respect to perceivers, gender stereotypes may negatively influence perceptions of others (e.g., [48, 394]) or change what people value in others [317]. Gender stereotypes also become more apparent when people are asked to assess others as opposed to themselves [299]. With respect to experiencers, gender stereotypes may contribute to various individual outcomes, e.g., career paths [345], as well as generally decrease performance via stereotype threat [549].

Our research is grounded in feminist theory and practice [42, 154, 33] and takes a feminist perspective on gender stereotypes by viewing them in the broader U.S. social, political, and cultural context. Feminist theory holds that identity is intersectional [137, 121] (connected to multiple identities) and closely and inextricably linked to structural oppression, and its goal is to end these forms of oppression [42]. Our work is motivated by the desire to contribute to the awareness and combating of gender stereotypes.

6.2.3 Gender stereotypes in STEM

Gender inclusivity in Science, Technology, Engineering, and Math (STEM) is drawing significant attention, as evidenced by recent handbooks, guides, and reports on gender inclusion and other identities (e.g., [338, 96, 1]). Education and economics research confirm the existence and extent of gender biases, including implicit biases associating men with STEM fields [187, 188, 186], stereotypes [450], stereotype threat [38, 549], and other barriers to participation (e.g., [563, 108, 405]). Other work also characterizes how stereotypes affect self-efficacy perceptions of women in STEM [83, 606, 355], including sense of belonging [367] and the interest [369] of girls in computing. Though this forms a considerable literature, the existence of gender biases is not always accepted, and its denial in STEM persists despite evidence [402]. Moreover, people who do not believe this bias exists may be more likely to perpetuate it [39].

Stereotypes lead to significant negative consequences in STEM for women and gender minorities [252], e.g., lower pay and less mentoring [401], increased stress [254] and other physical health problems [288], harassment [331, 565], and depressed performance [549, 40, 522].

Within the computing field, human-computer interaction researchers have found that gender stereotypes change perceptions of image search results [307, 430] or trust in robot voices [544]; machine learning researchers have found that gender stereotypes are also detectable in natural language with machine learning classifiers [87, 139].

A recent NSF report shows that in the U.S., while some fields in STEM are close to, or have even achieved, gender parity in education and employment, e.g., math and biology respectively, computer science remains one of the farthest from parity, with less than 20% of CS bachelor's degrees in the U.S. going to women [1]; this percentage decreased from 27% in 1998 [1]. Emerging research suggests that stereotypes about robotics may be stronger than stereotypes about STEM generally [368], calling for further identification and investigation of gender stereotypes in other

specific areas of computing, such as security and privacy.

6.3 Motivation

Having taken stock of work on gender in security and privacy research, as well as gender stereotypes in other fields, we now motivate the scope and goals of this chapter.

Explore an explanation for gendered differences in security and privacy behavior. As described in Section 6.2.1, a cluster of usable security and privacy research has identified gendered differences in behavior, but does not explain what accounts for such differences. For example, Sheng et al. found that women may be worse at identifying phishing [527] and Mazurek et al. found that women may choose weaker passwords [376]. One category of explanations could originate from *biological essentialism*, or intrinsic differences predetermined by one's gender (e.g., [232, 163]), and another from *social constructionism*, or cultural differences arising from societal expectations or other non-biological factors (e.g., [85, 75]). In other words, if women are worse at security and privacy behaviors, is it because of their biology or their society? Debate between proponents of each continues in academia (e.g., [160, 155, 174, 559]) and in society [86]; here, we study gender stereotypes, which has been posited to be an explanation for gendered differences in the style of social constructionism [272]. Our study asks participants about stereotypes related to previously found gendered differences, e.g., who is more likely to fall for scams or reuse passwords, thereby contributing to this literature by investigating gender stereotypes as a potential contributor to gendered differences.

Identify specific stereotypes whose impact should be evaluated. Initially, the research goal of our team was to measure the impact of gender stereotypes in security and privacy, and we conjectured research questions such as:

1. Do gender stereotypes in computer security and privacy negatively impact users themselves?
2. Do people who hold gender stereotypes in computer security and privacy cause negative impact to others?
3. To what degree do user interfaces reinforce gender stereotypes in computer security and privacy?

However, as we designed preliminary experiments, we encountered the following fundamental challenge: while we hypothesized the existence of gender stereotypes with respect to computer security and privacy, we did not know *what* gender-related beliefs were commonly held and should be included in our experiments. This observation led us to the need for a foundational, broad, and general study of gender stereotypes with respect to computer security and privacy. Our work empowers future researchers hoping to study impact with *specific, concrete, precise* gender stereotypes in security and privacy.

Inform future security and privacy research and practice. Understanding whether security and privacy-related gender stereotypes exist (and which, specifically) has the potential to help researchers and practitioners. Armed with knowledge about specific gender stereotypes, researchers can account for stereotypes in their methodologies, and designers can avoid unintentionally reinforcing them.

6.4 Pre-Study Method and Results

We conducted a pre-study in late 2020 to identify potential stereotypes to evaluate in our main study. Our institution's IRB determined this survey to be exempt; we followed the same ethical considerations and positionality statement as described in more detail in Sections 6.5.5 and 6.5.6.

6.4.1 Pre-study method

We sought to explore gender stereotypes with respect to security and privacy but found no prior work to examine. Thus, we recruited 202 U.S. participants from Prolific and asked: "What stereotypes can you think of about men, women, and people of different genders, when it comes to computer security or privacy? Please list as many as possible, including ones you don't believe in, but think others might."

One member of the research team followed a thematic coding process [67] to surface potential stereotypes. One coder led this process because our goal, i.e., to identify potential stereotypes for investigation in the main study, was subjective and generative [29, 400, 364]. To balance researcher subjectivity with thoroughness and integrity [602], the main coder reviewed the pre-study results with other research team members throughout the process. The other members corroborated that selected items would be meaningful and interesting to evaluate in the main study.

Our final codebook included 17 codes (i.e., potential stereotypes) across two high-level themes: stereotypes about why men would be better, or about why women would be better. For our main study, we selected only potential stereotypes that were mentioned by at least 5 participants.

6.4.2 Pre-study results

Participants reported potential stereotypes that men were more likely to be **logical**, but **overconfident**[1] and **lazy**, while women were more likely to be **perceptive**, but **emotional** and **gullible**. These were the six *personal characteristics* stereotypes in the main study. Participants also reported potential stereotypes that men **knew more**, were **more interested in**, and were **more skilled at protecting** their own security and privacy: these were the three *general engagement* stereotypes. Based on our participants' qualitative responses, we also interpreted that these stereotypes were

[1]Participants used "overconfident" and not "confident," which may contribute to a gendered interpretation; for fidelity, we use "overconfident."

either positive or negative, as indicated in Table 6.1.

6.5 Main Study Method

In early 2021, we conducted another online survey to evaluate gender stereotypes related to computer security and privacy surfaced from the pre-study. We submitted our study protocol to our institution's IRB, which determined that our study was exempt (Section 6.5.5).

6.5.1 Affinity diagramming

In one section of our main study, we sought to investigate stereotypes about specific security and privacy tasks. Few behaviors were surfaced organically by participants in our pre-study, likely because enumerating specific security and privacy tasks is much more salient to researchers and practitioners than to the general population we sampled in the pre-study. Thus, we reviewed the security and privacy advice literature and performed affinity diagramming to identify further tasks to include. Affinity diagramming is a method suitable for consolidating a large number of ideas through an iterative grouping process [275].

We gathered potential tasks from three recent papers from usable security and privacy: Ion et al.'s review of security and privacy advice (14 items, see Figure 1 in [289]), Redmiles et al.'s work on the same topic (35 items, see Figure 1 in [485]), and Egelman et al.'s standardized scale to measure end-user security behavior (16 items, see Table 4 in [180]). We additionally added two behaviors that were mentioned by some participants in the pre-study: falling for shopping scams and falling for dating scams.

We collected all security and privacy behaviors from the aforementioned sources and grouped similar ones, e.g., tasks related to internet safety, authentication, privacy, or finances. We iteratively pared down the list and removed those that were not applicable to all internet users (e.g., use

parental controls, set up IoT devices) or that were too vague (e.g., act anonymously online, remove unnecessary programs). To sample a range of behaviors, we selected five behaviors that are beneficial for one's security and privacy (marked as positive in Table 6.1), and five detrimental (negative).

6.5.2 Survey structure

Participants first completed a consent form and read the following instructions: "While we understand there are many genders, for the purposes of this study, we will ask about specifically men and women." We further clarified that we were interested in participants' honest thoughts and opinions, that there were no right or wrong answers, and that their responses would have no impact on compensation. We emphasized this information at the beginning of the survey to minimize the potential for social desirability biases [477] to influence participants' responses. The full survey instrument is shown in Appendix D.1.

Three stereotype categories. The first three sections of the survey asked about three different categories of potential stereotypes regarding security and privacy perceptions and behavior, totalling 19 potential stereotypes (full list in Table 6.1). Participants saw these sections in a randomized order. We asked: "Based on your personal beliefs and experiences, who is more likely to be more [potential stereotype] when it comes to computer security and privacy?" Answer choices were "Definitely men," "Probably men," "Men and women equally," "Probably women," "Definitely women," "Another gender, please specify," and "Don't know or not sure."

Follow-up questions about stereotype sources and rationales. In the fourth section of the survey, we asked participants who had *not* responded that women or men were equally likely (i.e., who expressed a gendered stereotype) for the prior questions to elaborate on why they believed the gender stereotype with respect to computer security and privacy existed. Prior work highlights a

Table 6.1: We studied 19 potential stereotypes related to security and privacy, in three categories: general engagement, personal characteristics, and specific behaviors. Items in first two categories were generated from our pre-study (see 6.4.2), and items in last category from affinity diagramming security behaviors listed in prior work (see 6.5.1).

Potential stereotype	**Pos.**	**Neg.**
General engagement [from pre-study]		
Interested in learning about protecting	×	
Know how to protect	×	
Skilled at protecting	×	
Personal characteristic [from pre-study]		
Be logical	×	
Be lazy		×
Be overconfident		×
Be perceptive	×	
Be emotional		×
Be gullible		×
Specific behavior [from [289, 485, 180]]		
Verify HTTPS	×	
Install software updates immediately	×	
Use antivirus software	×	
Enable 2FA	×	
Ask for help if have questions	×	
Fall for shopping scam		×
Fall for dating scam		×
Leave device unlocked		×
Reuse password		×
Share sensitive info on social media		×

divide between biological and non-biological reasons for gendered expectations, so we focused on this distinction [51]. Replicating a prior (more general) Pew research study [437], we offered the following answer choices: "Biological reasons," "Non-biological reasons," "Other reasons, please specify," and "Don't know or not sure." We also asked participants to explain their choice with a free-text response.

In the fifth section of the survey, we asked some general questions about sources where participants may have heard gendered stereotypes regarding security and privacy.

Ambivalent Sexism Inventory and demographics. The survey concluded with the Ambivalent Sexism Inventory (ASI) [221, 222], a standardized measure of individual sexism scored from a low of 1 to a high of 6 (reproduced in full in Appendix D.2), as well as demographic questions.

Survey pre-testing. To pre-test the survey, we conducted 5 expert reviews with researchers familiar with security and privacy user studies, as recommended by best practice [477]. This process allowed us to catch best-practice errors and validate that our survey was serving our research questions. We further conducted 10 pilot tests with Prolific participants (data excluded from our results) to identify any remaining misunderstandings or technical issues, and we updated question wordings or survey code accordingly.

6.5.3 Participants

Table 6.2: Breakdown of participant demographics by gender, age, education, race/ethnicity, and technical background.

Gender		Age		Education		Race/Ethnicity		Tech background	
Woman	38.9%	18-24	19.9%	High school	13.0%	White	59.5%	No	58.4%
Man	56.3%	25-34	34.7%	Associate's or some college	7.1%	Asian	11.1%	Yes	19.5%
Non-binary	2.1%	35-44	24.2%	Trade/technical/vocational	1.9%	Black or African American	5.8%	Prefer not to say	3.2%
Multiple genders	1.0%	45-54	9.3%	Bachelor's	39.6%	Hispanic, Latino, or Spanish Origin	4.7%		
Genderfluid	<1%	55-64	6.8%	Master's	16.9%	Mixed	1.6%		
Prefer not to say	1.0%			Professional degree or doctorate	3.9%	Middle Eastern or North African			
						Other	<1%		

We recruited participants from Prolific, a crowdworking platform shown to be better than other crowdsourcing platforms, such as Amazon Mechanical Turk, in terms of comprehension, attention, and honesty of its participants [459, 444]. We recruited participants who lived in the United States and were fluent in English. For the main study, we collected responses from a total of 190 U.S. participants. We verified participants were paying attention to our survey by checking the coherency of their responses to open-ended questions. Participants took 17 minutes, on average, to complete the survey. We compensated them $2.50, which was calculated based on the average length of our pilot tests (10 minutes) at an hourly rate of $15/hour. 74 were women, 107 were men, 4 were non-binary, 1 was a woman and non-binary, 1 was a woman and man, 1 was genderfluid, and 2 preferred not to say. 11.6% of participants reported having an education or working in security and privacy in particular. Table 6.2 shows additional demographic information.

6.5.4 Data analysis

We used a mixed quantitative and qualitative approach. For statistical analysis with participants' gender, we excluded responses not from women or men because we did not have enough responses in these categories to have adequate statistical power to make accurate claims. However, we report qualitative data from all participants.

Quantitative. For all quantitative analyses, we binned responses into those towards women ("Definitely women" / "Probably women") or towards men ("Definitely men" / "Probably men") to increase our statistical power, but we report these gradations in figures for context. We did not perform statistical testing with the other responses ("Another gender, please specify" and "Don't know or not sure").

To understand whether significantly more participants believed stereotypes about women or men, we conducted two-sided exact binomial tests to determine whether the proportions of

responses towards either differed significantly (dropping the "Men and women equally" option). We performed Holm's correction to reduce Type I error.

We were interested in how participants' gender and sexism scores impacted their stereotype beliefs, but we did not include gender and sexism scores in the same model because they were significantly correlated ($p < .05$). To identify how participants' self-identified gender affected their security and privacy stereotype beliefs, we conducted two two-sided exact binomial tests on responses towards women and men for each stereotype, one for the subset of women participants, and one for the subset of men participants. We performed Holm's correction within each family. To identify how participants' sexism impacted their stereotype beliefs, we constructed 19 multinomial logistic regressions models, one for each stereotype. The dependent variables (DV) were responses to the stereotype question, retaining the "Men and women equally" option to account for participants with low sexism scores. The independent variable was the numeric overall ASI score.

Finally, we investigated whether participants believed stereotypes for biological or non-biological rationales with two-sided paired t-tests. We also conducted two mixed logistic regressions to investigate whether participants' sexism score correlated with their selected rationales. The independent variable for both regressions was sexism score; the dependent variable for one was selecting biological rationales (dummy-coded to 0 or 1), and for the other, selecting non-biological rationales (also dummy-coded). Regressions were separate because rationales were not independent, and we performed Holm's correction within each family.

Qualitative. For participants' free-text rationales for the stereotypes, we used qualitative thematic analysis to describe and interpret (but not necessarily verify or evaluate) [445] themes in how they justified their beliefs. Our goal was to facilitate deeper explanations of why participants believed in gender stereotypes regarding security and privacy, beyond the choice of "biological"

and "non-biological."

We followed a thematic coding process [67]. One researcher read and re-read all data, noting initial thoughts about the rationales participants gave. The researcher then generated a set of themes, applied them to the full dataset, and iteratively defined and refined each theme (full codebook in Appendix D.3). One member of the team performed the analysis, consistent with viewpoints from qualitative research theory and practice about the potential for multiple coders to reduce interpretive nuance [29, 400] or the semantic power of the codebook [364]. To balance researcher subjectivity with thoroughness and integrity [602], another team member reviewed the codebook, independently coded 25 randomly selected responses, and discussed and resolved differences with the main coder.

6.5.5 Ethical considerations

Our institution's IRB reviewed our study and determined it to be exempt. However, IRB review is not sufficient to guarantee ethical research. We identified the following ethics-related questions: would our research instrument cause our participants to believe (1) harmful stereotypes that they did not believe prior to participating in our study, or (2) that harmful stereotypes apply to others or themselves? We carefully constructed our survey to avoid suggesting any gender differences were true; rather, our survey was designed to be neutral and elicit the participant's unprimed responses (full survey instrument in Appendix D.1).

6.5.6 Positionality statement

Aligned with feminist methodology, we recognize that our position as researchers and our identities influence our research [33, 274].

All researchers have observed instances of gender stereotyping with respect to computer

security and privacy, either directed at ourselves or via our roles as instructors of computer security courses. We have the most personal experience with gender stereotypes as it relates to people who are women or men and thus focused our study on these genders. Two researchers were born outside of the U.S.; all of us have lived in the U.S. for at least the last six years. Our work focuses on stereotypes in the U.S. cultural context for this reason. We further discuss how our positionality limited our research perspective in Section 6.5.7.

6.5.7 Limitations

We must consider standard survey-based limitations, including survey fatigue and social desirability bias. We attempted to mitigate these concerns by pre-testing our survey to optimize its length and by explicitly stating that there were no right or wrong answers. However, our acknowledgement at the beginning of the study that there are many genders may have signaled our positionality and influenced some responses. Further, we studied only perceptions that participants were willing to report in our survey, suggesting that our results are a lower bound on gendered perceptions people consciously or subconsciously hold. In terms of fatigue, we received a large amount of free response text (24,180 words) from our 190 main study participants, suggesting that many engaged deeply with the survey.

Our results are also limited by the characteristics of our Prolific sample. Crowdworkers have more internet experience than the general U.S. population, but are still representative in terms of security and privacy experiences and knowledge [483]. Prolific has emerged as an alternative to other crowdworking platforms like Amazon Turk [433] for implementing features to improve participant recruitment specifically for scientific researchers. We studied only English-speaking U.S. participants; gender stereotypes with respect to computer security and privacy may look different in other cultures and contexts.

Finally, our work is limited by our own identities, perspectives, and experiences as researchers. We hypothesize that intersections with race or ethnicity (e.g., for Black women), gender identity (e.g., for transgender women), age, sexuality, dis/ability, and other identities would strongly modulate how gender stereotypes in security and privacy are experienced. Our research team is composed of people in their 40s or younger; two are white, one is Asian, and one is Asian, Native, and white. While we partially made the choice to focus our work on stereotypes of binary genders to build on existing sexism research instruments (the ASI only considers binary genders), our own identities also shaped the limited scope of this work. Future work should investigate other critical aspects of identity in an intersectional way [137]; extending a study to a full spectrum of genders or identities will require at minimum a thoughtful redesign or even a different method entirely.

6.6 Main Study Results

First, we analyze our participants' sexism scores (Section 6.6.1). We then report stereotypes that our participants believed about women, men, or that were not strongly associated with either (Section 6.6.2), whether these differed by sexism or participant gender (Section 6.6.3), and the sources of these stereotypes (Section 6.6.4). We conclude with participants' rationales for these stereotypes (Section 6.6.5).

6.6.1 Sexism scores

The Ambivalent Sexism Inventory (ASI) is scored from 1 (low) to 6 (high) [221]. Overall, our participants scored an average of 2.7 (SD 1.0; range 1-4.8). Further broken down into benevolent and hostile sexism, our participants scored an average of 2.8 (SD 1.1, range 1-4.7) and 2.6 (SD 1.33; range 1-5.5), respectively. Men's overall sexism scores were higher than for women (Table 6.3). Additionally, both men's benevolent sexism and hostile sexism were higher than women's. Gender

correlated significantly with participants' overall ASI ($Z = 3.01$, p-value $< .01$) and hostile sexism ($Z = 3.31$, p-value $< .001$), but not for benevolent sexism.

Table 6.3: Sexism scores, as mean (SD), of all participants, just women, and just men. We used the Ambivalent Sexism Inventory (ASI), measuring overall sexism, benevolent sexism, and hostile sexism from 1 (low) to 6 (high).

	All	Women	Men
Overall sexism	2.7 (1.0)	2.5 (1.0)	2.9 (1.0)
Benevolent sexism	2.8 (1.1)	2.7 (1.1)	2.9 (1.0)
Hostile sexism	2.6 (1.3)	2.2 (1.2)	2.9 (1.4)

6.6.2 What stereotypes exist about how women and men protect their security and privacy?

Stereotypes about women. Out of nineteen stereotypes we investigated, we found five regarding security and privacy characteristics or behaviors about women. Participants expressed that women would be more likely than men to:

- Share sensitive information on social media (-)
- Be emotional (-)
- Fall for shopping scams (-)
- Ask for help if they have questions (+)
- Be gullible (-)

For these stereotypes, 37%-68% of participants responded women would definitely or probably be more likely to be or do so, compared to men (Figure 6.1). One cluster of these stereotypes about women regard their personal characteristics, i.e., that they are more likely to be emotional or gullible. Another cluster regards specific behaviors, but none of the stereotypes about women included positive stereotypes from the category of general engagement with security and privacy.

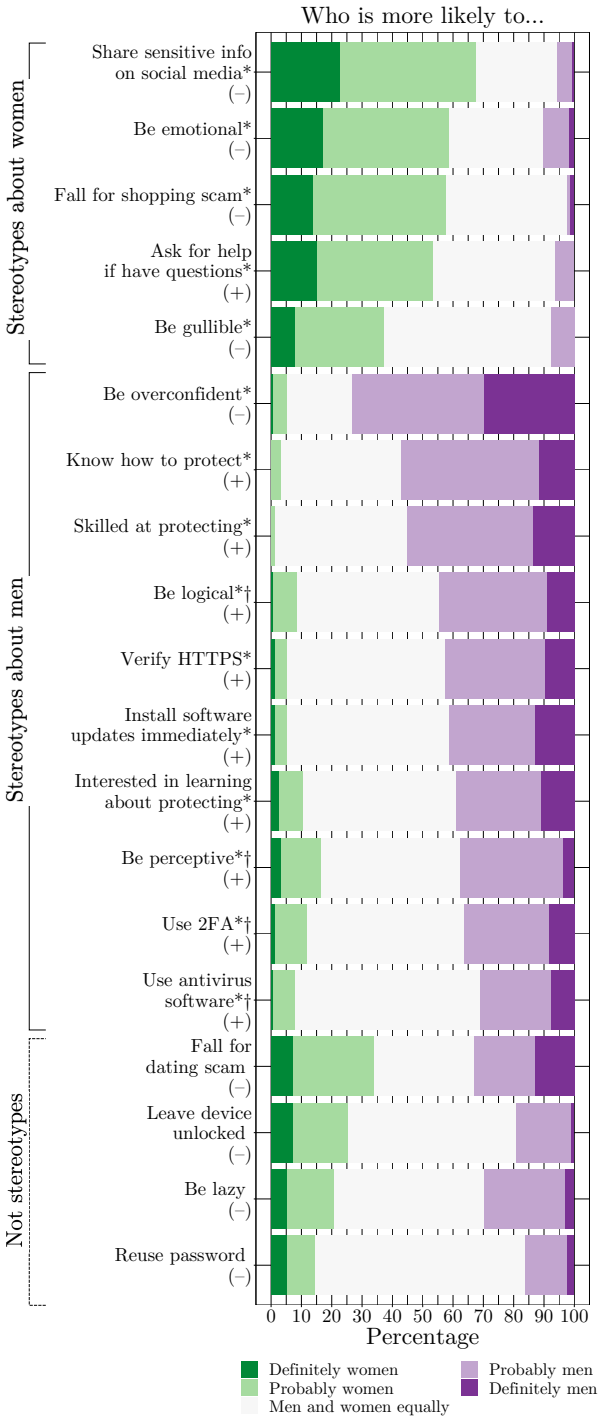

Figure 6.1: Stereotypes in security and privacy. * represents stereotypes we found, defined by a significant difference (p-value $< .001$) in proportion of participants who selected either "Definitely women" / "Probably women" compared to "Definitely men" / "Probably men". † represents stereotypes believed more by men; see further detail in Figure 6.2. (+) represents positive stereotypes, and (–) negative stereotypes.

From our original interpretations (of participants' responses in the pre-study, or of related work we referenced) about potential "positive" or "negative" stereotypes, four of these five were negative stereotypes about women, with only one positive stereotype: asking for help.

Stereotypes about men. For ten stereotypes, significantly more participants associated the characteristic or behavior with men over women. We found that at least 30% of participants associated ten security and privacy stereotypes with men (Figure 6.1), i.e., that men would be more likely to:

- Be overconfident (-)
- Know how to protect their security & privacy (+)
- Be skilled at protecting their security & privacy (+)
- Be logical (+)
- Verify HTTPS (+)
- Install software updates immediately (+)
- Be interested in learning about protecting security & privacy (+)
- Be perceptive (+)
- Enable 2FA (+)
- Use antivirus software (+)

These stereotypes about men include all three of our potential stereotypes about general engagement with security and privacy, as well as three personal characteristics (i.e., overconfident, logical, and perceptive). Another cluster of these stereotypes regarding men are about a range of protective security and privacy behaviors, including verifying HTTPS, installing software updates, and enabling 2FA. From our original hypothesis about the stereotypes being "positive" or "negative", all stereotypes about men were positive except overconfidence.

Our finding that men are more logical contradicts prior work (in STEM broadly, not security

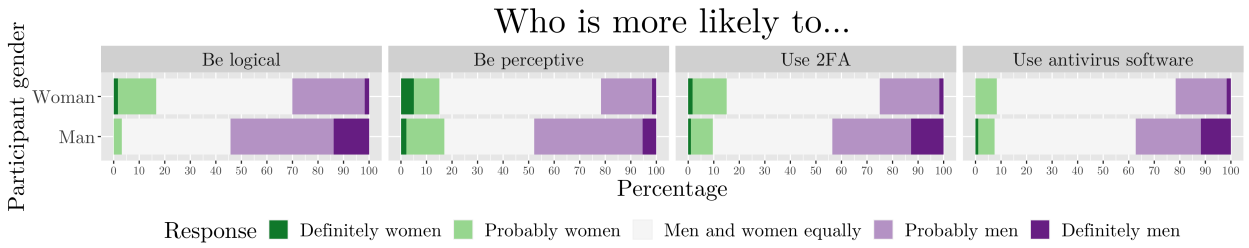

Figure 6.2: Four stereotypes held by men, but not by women, in our sample (all other stereotypes were held by both men and women). Differences in proportions of participants who selected either "Definitely men" / "Probably men" compared to "Definitely women" / "Probably women" were statistically significant (p-value $< .001$) for men but not for women.

and privacy specifically) that logical thinking was perceived to be a gender-neutral personality trait, i.e., not a gender stereotype [450].

Stereotypes not strongly held about women or men. There were no statistically significant differences in the proportion of participants who believed gender was associated with four characteristics or behaviors:

- Being lazy (-)
- Falling for dating scams (-)
- Leaving devices unlocked (-)
- Reusing passwords (-)

Given that differences in the proportion of responses were not significant, these characteristics or behaviors could be described as gendered prejudices that are held by a minority.

6.6.3 How do stereotyped beliefs vary by participants' gender and sexism level?

Stereotype beliefs by participant gender. Building on the identification of stereotypes in the prior section, we now turn to whether the beliefs in stereotypes were correlated with participants'

gender. We find four stereotypes about men that men believed but women did not:

- Be logical (men: p-value $< .001$, women: *n.s.*)
- Be perceptive (men: p-value $< .001$, women: *n.s.*)
- Use 2FA (men: p-value $< .001$, women: *n.s.*)
- Use antivirus software (men: p-value $< .001$, women: *n.s.*)

Figure 6.2 shows participant responses to these stereotypes, comparing women and men in our sample. We found no stereotypes that women believed but men did not, indicating that — of the stereotypes we studied — men held more gender stereotypes regarding security and privacy than women.

Men and women alike held the 11 remaining stereotypes in Section 6.6.2. This suggests that most stereotypes are widespread; however, select stereotypes are only held by men, which, further, are positive stereotypes about men.

Stereotype beliefs by participant sexism. In addition to participants' gender, we wanted to know whether higher levels of participant sexism, measured via Ambivalent Sexism Inventory (ASI) scores [221], correlated with having (or not having) belief in gender stereotypes.

Overall, we found that as sexism scores increased, so too did the belief in fifteen of the stereotypes we studied: seven about women, and eight about men (see Appendix, Table D.1). Participants who the test identified as more sexist were significantly more likely to believe that women would be emotional (estimate = 1.00, p-value $< .001$), gullible (estimate = 0.85, p-value $< .001$), lazy (estimate = 1.32, p-value $< .001$), fall for shopping scams (estimate = 0.89, p-value $< .001$), ask for help (estimate = 0.80, p-value $< .01$), reuse passwords (estimate = 0.86, p-value $< .05$), and leave devices unlocked (estimate = 0.95, p-value $< .001$). Note that beliefs about women being lazy, reusing passwords, and leaving devices unlocked were not found to be stereotypes overall but were views more likely to be held by participants who scored higher on

the sexism scale.

For stereotypes about men, participants who the test identified as more sexist were more likely to believe that men would be more likely to know how to protect (estimate = 0.71, p-value $< .01$) and be skilled at protecting their security and privacy (estimate = 0.70, p-value $< .01$), be perceptive (estimate = 0.76, p-value $< .01$), be logical (estimate = 1.26, p-value $< .001$), verify HTTPS (estimate = 0.70, p-value $< .01$), install software updates immediately (estimate = 0.78, p-value $< .01$), use 2FA (estimate = 0.62, p-value $< .05$), and use antivirus software (estimate = 1.10, p-value $< .001$).

6.6.4 Personal exposures to stereotype beliefs

We asked participants to select all sources (not mutually exclusive) where they had heard about people of one gender being better than others at performing security and privacy behaviors. 82 participants reported not hearing about gender differences from any source. Social media was the most commonly cited source (31), followed by friends (25), TV/movies (23), family (20), work or job (20), or the news (16). For the 10 "other" responses, participants mentioned hearing about stereotypes from teachers, co-workers, the military, their own experiences, nowhere in particular, and "ambient cultural osmosis" (P116). One participant wrote, "of course i have heard, what a silly question to ask." [sic]

6.6.5 What are participants' rationales for their stereotype beliefs?

We now turn to our second major research question: What are participants' rationales for gendered stereotypes? Towards the end of combating gender stereotypes, we sought to understand what types of evidence are used to rationalize the stereotype beliefs. Table 6.4 summarizes these results.

Table 6.4: Participants' rationales for gender stereotypes. For stereotypes about women or men, closed-response rationales are shown for participants who believed that stereotype and whether there was a significant difference between participants selecting biological or non-biological reasons (% do not sum to 100% because choices were not mutually exclusive and "other" is not shown here). For example, 26% of participants who believed women would be more likely to share sensitive information on social media believed so for biological reasons; this choice was significantly less than the 72% that selected non-biological ($t(45) = -4.12$, $p < .01$). Selected quotes from participants are shown for open-response rationales. For not gender stereotypes, rationale data is reported from all participants. *Note:* * $p < .05$, ** $p < .01$, *** $p < .001$.

	Stereotype	Closed-response rationales					Participant quote
		Reasons	%	*t*	*df*	*p-value*	
Stereotypes about women	Share sensitive info on social media	Biological	26	-4.12	45	**	"Women are less biologically driven to use technology, and thus may not be as aware of the risks of sharing too much information online."
		Non-biological	72				"[Women] like more of the attention"
	Be emotional	Biological	75	4.46	62	***	"Due to their genetics, women tend to be much more emotional, their brains are created in such a way that emotions are much more intense in them"
		Non-biological	29				"it's more socially acceptable for women to be more emotional... Therefore I'd expect women to be more emotional over computer security than men"
	Fall for shopping scam	Biological	19	-6.77	58	***	"Women gather and make clothes/food for there families, to clothe their children...over time these things leave biological signatures for survival."
		Non-biological	78				"There are more shopping-related scams targeting women."
	Ask for help if have questions	Biological	19	-3.84	26	**	"Women's brains generally do not think that way and they don't have a problem asking for help."
		Non-biological	74				"[women are] more familiar with and more culturally comfortable with asking for answers."
	Be gullible	Biological	40	-1.67	51	(*n.s.*)	"Women are biologically programmed in many ways over thousands of years to trust their intuition over logic... Once they feel something is right.. they take that path repeatedly and incautiously... and as such can easily be manipulated"
		Non-biological	62				"[Women] give someone the benefit of doubt mostly. We don't mean to be gullible just polite"
Stereotypes about men	Be overconfident	Biological	30	-4.42	53	***	"I think men think that they are just stronger and incapable of someone coning them even when it is in a technology situation."
		Non-biological	76				"Men are generally socialized to have more confidence than women, especially about technology. They are much more likely to be overconfident."
	Know how to protect	Biological	28	-3.76	70	**	"Men are just protectors in general. Its just in their blood."
		Non-biological	66				"Culturally men have been the ones more responsible for protection for any sort, so I would think it would extend to this as well"
	Skilled at protecting	Biological	26	-4.33	65	***	"[men are] more biologically driven to use computers"
		Non-biological	70				"women generally have less access to computer security and privacy... career choices... lesser participation in STEM categories and it's no fault on their part"
	Be logical	Biological	54	0.85	55	(*n.s.*)	"I feel men are more likely to be more logical when it comes to computer security and privacy because men are natural born problem solver and always try to explore the best possible means to fix a problem"
		Non-biological	43				"Men have been at the forefront of technological advancement"
	Verify HTTPS	Biological	25	-5.89	59	***	"Men probably use those sites that need to be protected than women, so they are more used to what it is"
		Non-biological	82				"There are more men that are into computers, thus they would be more likely to know to look for this."
	Install software updates immediately	Biological	25	-4.14	59	**	"Men due to their natural skepticism are more likely to recognize the danger of not keeping software up to date."
		Non-biological	68				"Men have time - they aren't as busy with children or taking care of the household. They like to take care of their 'toys' and tech."
	Interested in learning about protecting	Biological	21	-4.30	47	**	"Men have been wired... cognitively to be protectors, of themselves first and foremost, [if they] sense threat they deal with it way thoroughly than women."
		Non-biological	69				"Men tend to be more interested in things whereas women like to learn and study people"
	Be perceptive	Biological	22	-4.55	44	***	"that's just the way guys are, nerdy and techy, Women are careless on the computer, not as much knowledge about geeky stuff as men"
		Non-biological	76				"Because of societal factors, men are given more training and confidence in computer-related fields... hence why they are more perceptive."
	Use 2FA	Biological	32	-3.19	52	*	"[2FA] is too complex for women"
		Non-biological	70				"Men are more likely to be targeted by technology news and are more likely to have been informed of the benefits of two-factor authentication."
	Use antivirus software	Biological	27	-3.47	47	**	"naturally men are always security conscious and can go the extra length to secure their devices"
		Non-biological	71				"[men] seem more like the type to download more sketchy items from the internet."
Not gender stereotypes	Fall for dating scam	Biological	47	0.15	57	(*n.s.*)	"women more subject to being swayed by their emotions"; "[mens'] testosterone may temporarily inhibit sound decision when it comes to dating-related financial scams"
		Non-biological	45				"Women are seen as softer targets by scammers"
	Leave device unlocked	Biological	32	-3.01	59	*	"Women tend to be more trusting and less skeptical"; "Men tend to be more careless"
		Non-biological	65				"Women rarely have things to hide"; "Men more likely to take risks"
	Be lazy	Biological	26	-3.91	57	**	"Women more concerned about posting a photo and how many likes they get"
		Non-biological	69				"A lot of men think they should have things done for them due to personal and societal standards"
	Reuse password	Biological	40	-1.15	46	(*n.s.*)	"Women is more likely to use the same password for multiple accounts. Because it is easy to remember for women"
		Non-biological	55				"Men want things to be as simple as possible"

Closed-response rationales

As a first pass, we asked participants to select whether they believed gender differences were due to biological or non-biological factors, replicating prior work [437].

For stereotypes about women, significantly more participants believed that non-biological reasons explained why women would be more likely to share sensitive information on social media ($t(45) = -4.12$, p-value $< .001$), fall for shopping scams ($t(58) = -6.77$, p-value $< .001$), and ask for help ($t(26) = -3.84$, p-value $< .001$). On the other hand, stereotypes about women's personal characteristics – being more emotional and gullible – were attributed by more participants to biological reasons (75% and 62%, respectively), although this difference was only significant for being emotional ($t(62) = 4.46$, p-value $< .001$). Participants may have perceived *actions* to be more related to societal expectations (non-biological factors), e.g., that women use social media and shop more and thus would fall for more shopping scams, while *personal characteristics* were seen as biologically determined. Further, participants with higher sexism scores were more likely to consider women to be emotional (estimate $= 0.57$, p-value $< .01$) or gullible (estimate $= 0.72$, p-value $< .01$) as a result of biological reasons, while participants with lower sexism scores were more likely to attribute women being more emotional (estimate $= -0.47$, p-value $< .05$) or sharing sensitive information on social media (estimate $= -0.77$, p-value $< .001$) to non-biological reasons.

Regarding stereotypes about men, significantly more participants attributed nine of the ten to non-biological reasons: being overconfident ($t(53) = -4.42$, p-value $< .001$), knowing how to protect ($t(70) = -3.76$, p-value $< .001$), being skilled at protecting ($t(65) = -4.33$, p-value $< .001$), verifying HTTPS ($t(59) = -5.89$, p-value $< .001$), installing software updates immediately ($t(59) = -4.14$, p-value $< .001$), being interested in learning about protecting ($t(47) = -4.30$, p-value $< .001$), being perceptive ($t(44) = 4.55$, p-value $< .001$), using 2FA ($t(52) = -3.19$, p-value $< .05$),

and using antivirus software ($t(47) - 3.47$, p-value $< .05$).

Open-response rationales: Sources of and evidence for stereotype beliefs

Participants were also asked to explain their rationales for holding stereotypes; we collected a total of 1,159 free-text rationalizations from 150 participants and now present results of thematically analyzing these responses. Aligned with qualitative methods, our analysis is intended to be generative, surfacing themes about the sources of and evidence for participants' beliefs, rather than measuring pervasiveness. As such, we report whether themes were expressed by a few (less than 25%), some (25%-49%), or many (more than 50%) of the 150 participants that provided free-text rationales. We also apply our own interpretive lens to develop shared themes across participants' responses that build on, but ultimately rise above and enrich, the closed-response rationales.

Other stereotypes. Many rationales for stereotypes were based on other stereotypes. P117 explained that women were more likely to be gullible because:

> "Women have a tendency to be compassionate...and listen to others and that often gives scammers the opportunity to fool them."

Often, participants rationalized their beliefs for who would be more likely to reuse passwords based on which gender they perceived to be more lazy, e.g.,

> "women are naturally lazy in issues of internet matters and always tend to seek the easy way out" (P189).

"Science". A few participants' rationales referenced biologically essentialist effects of estrogen, testosterone, and hormones. These also spanned scientific disciplines including biology, psychology, and anthropology:

> "Women are biologically programmed in many ways over thousands of years to trust their intuition over logic" (P170).

> "Men might have a certain drive to explore, and so often venture into new territory like technology" (P142).

> "Men seem to be the protectors in anthropological terms" (P108).

Societal expectations. Some participants rationalized their beliefs by referencing social discourses about the ways that women or men should be. The most common was that men were expected to understand and enjoy technical topics and were provided support and encouragement to have interests in STEM — in P122's words:

> "the social coding of those hobbies as 'masculine' "

which led to participants deducing that men would be more likely to verify HTTPS, use 2FA, install software updates, and more. On the other hand, we observed stereotype rationales about societal expectations that women be family-oriented, e.g., that women would be more likely to fall for shopping scams because:

> "Women gather and make clothes/food for their families, to clothe their children, as these responsibilities often fall on women" (P142).

Personal observations and experiences. Many participants' rationales came from their observations that women or men in their lives tended to have certain traits, e.g.,

> "With the way social media is, women are known to 'overshare' information about their lives. I don't see too many men doing this" (P159).

or take certain actions, e.g.,

> "From all my friends the male ones concern more about their privacy and security, so they look it up about it more" [sic] (P43).

Assumptions of knowledge, level of experience, and interest. Many participants wrote that men likely had more knowledge, experience, or interest in technical topics, which then influenced

stereotypes they held about men. Participants assumed that men were more interested in software, gaming, and the internet, and thus would be more knowledgeable about computers, security, and privacy. Some commented on women's apparent lack of interest, e.g.,

> "Women consider technology a tool, something to use but not spend too much time on" (P129).

> "Women have more things on their mind than computers, ie: home life, kids, errands, friends. Most leave it up to their husbands to take care of the techy geeky stuff" (P132).

Threat models. The development (or lack thereof) of three aspects of threat models contributed to participants' stereotype rationales. First, some participants referred to innate valuations of security and privacy, e.g.,

> "men value security more [so they will use 2FA]" (P21).

> "women may care less about this topic than do men [so they will be more gullible]" (P144).

A range of assets contributed to valuing security, e.g., women's personal information that could be abused to harass, or men's financial information or browsing activities:

> "men probably have more to hide on their devices, honestly, ...to lock up porn history" (P142).

Second, a few participants believed one gender had a better understanding of threats, e.g.,

> "men tend to... understand how security plays a role and the consequences that come if you are not protected" (P72).

Other participants highlighted negative experiences that contribute to threat awareness, e.g.,

> "Women are more often targets of cyber stalking, doxxing campaigns, and scams than men, so they have a more obvious reason to avoid sharing sensitive information and probably learn more quickly how to do so effectively" (P186).

Third, a few participants observed threats external to individuals, e.g., scammers target women on shopping sites or men on dating sites.

"Just because". Finally, a few participants did not rationalize the stereotypes they held with a unique reason, e.g., P186's rationale for why men were more likely to leave devices unlocked:

> "That was just a gut feeling, I have no reasoning to back it up."

In another example, P130 uses explicit and non-inclusive language to explain why men reuse passwords more:

> "Their bodies are different, women have [slang term for body part], men have [slang term for body part]! isnt that enough." [sic]

We find the lack of rationales meaningful because they reflect internalized biases; people may not have thought consciously about gender stereotypes in security and privacy, and yet, they exist.

6.7 Discussion

6.7.1 Summary and key findings

Using two surveys, we studied the beliefs that Prolific crowdworkers in the U.S. hold with respect to gender and computer security and privacy. To answer our initial research questions:

What gender stereotypes (about women or men) do members of the general U.S. public hold that concern everyday computer security and privacy issues? Participants in our study believed that men are more likely or more able to protect their computer security and privacy than women, e.g., that men are more interested in and more skilled at computer security and privacy, or that women are more emotional and likely to fall for scams (see Section 6.6.2). Because these beliefs were held by statistically significant proportions of our sample, we identify these as gender stereotypes. Additionally, we find that gender stereotypes are held by both women and men (see Section 6.6.3).

More sexist participants, based on their responses to the ASI [221], are more likely to believe these stereotypes (see Section 6.6.3).

What explanations or rationales do people give to justify gender stereotypes? A sizeable proportion of participants rationalized gender stereotypes about security and privacy topics by either reiterating gender stereotypes from outside of computing or invoking essentialist claims. Many participants also reflected on societal gender expectations and personal experiences or assumptions as contributors to the existence of gender stereotypes. Overall, rationales for gender stereotypes spanned the spectrum from biological to non-biological and were deeply entrenched in participants' perceptions of others (see Section 6.6.5).

6.7.2 Guidelines for the future

Though the existence of gender stereotypes with respect to computer security and privacy is not surprising, given the documentation of stereotypes in other contexts, our work uniquely captures the existence of specific gender stereotypes in the field of security and privacy. We hope this work inspires other researchers to explicitly consider the impact of stereotypes on the design and evaluation of future computing systems, and to further investigate the relationships between gender or other identities and computer security and privacy. Building on the implications of our work, we suggest ten guidelines for future work.

Familiarize research and design teams with the principle that stereotypes and facts are related but separate concepts

Stereotypes, or reductive beliefs about a population, are distinct from facts, or empirical measurements of that population. For example, we found a stereotype held by our participants that men would be more likely to use 2FA. This stereotype is distinct from empirical measurements

suggesting more men may use 2FA than women [455]. This distinction distinguishes our work from prior work making empirical measurements because whether or not stereotypes align with empirical measurements of a population, stereotypes can cause harm. Even if multiple studies corroborate that men are indeed more likely to use 2FA than women, thereby ostensibly providing "evidence" for this gender stereotype, the stereotype may discourage scores of women from even attempting to set up 2FA for their accounts. We recommend that future researchers be mindful the distinction between stereotypes and empirical measurements, and study the relationship between the two.

Investigate the potential role of stereotypes when gender gaps are uncovered

Our work demonstrates that individuals' gender can have significant impact on their likelihood of believing stereotypes. Thus, we suggest that when gender gaps are uncovered in security and privacy (e.g., in adoption rates, in preferences, in attitudes), researchers explore whether gender stereotypes contributed to those gaps. Gender stereotypes may have also contributed to prior work that found gendered differences, such as in individuals' password choices [376] or susceptibility to phishing [527]. Keeping in mind guideline 6.7.2, gendered differences in individuals' behaviors could be a result of gendered stereotypes or other types of gender discrimination.

Familiarize research and design teams with the potential harms of stereotypes

We investigated the existence of specific gender stereotypes in our realm of computer security and privacy to create a foundation for studying the potential harms of these stereotypes. While we look forward to a multitude of future research examining what and how harms manifest in security and privacy, we recommend that research and design teams familiarize themselves with the harms of gender stereotypes in other domains, e.g., on self-efficacy [606, 355] or interest [522] in STEM, as well as feminist primers [42] that provide contextual theory.

Explore harms arising from people believing gender stereotypes about themselves

We encourage future research to explore the gender stereotypes' harms that arise for *experiencers*, or members of the stereotyped group. For example, we discovered stereotypes that women would be more likely to fall for shopping scams. Does this stereotype then contribute to women developing learned helplessness in avoiding such scams? Prior research on stereotype threat suggests that this could be the case; equally qualified women performed worse on a math test after being reminded of negative stereotypes about women and math [549]. We urge future work to explore the harms of the specific stereotypes identified in this work.

Explore harms arising from people believing gender stereotypes about others

We encourage future research to also explore gender stereotypes' harms that arise from *perceivers*, or people who hold stereotypes that distort their perceptions of others. For example, we found stereotypes that men would be more likely to use 2FA and other common security tools. Does this contribute added barriers for women who seek to use such tools, in opposition to the stereotype? Additionally, we urge future researchers to consider potential harms from gender stereotypes for people of non-binary genders. Gender is multiplicitous in its "many meanings and relations to individuals and communities" [311], and study of gender stereotypes and non-binary genders may necessitate research approaches and methods beyond those used in this work, which were intended as an initial investigation and does not adequately contend with gender's multiplicity.

Combat gender stereotypes that reduce adoption of positive security and privacy behaviors

Gender stereotypes are well-documented to present barriers to participation, e.g., in the field of STEM [108, 405]. The gender stereotypes identified in this work suggest that they may also have

negative effects on the adoption of positive security and privacy behaviors, e.g., using 2FA, being interesting in learning about security and privacy. Especially for (but not limited to) topics where stereotype belief or individual sexism correlate with disproportionate adoption by gender, we call for the combating of those stereotypes. These efforts may align with a growing playbook to increase representation in computing and STEM, such as through outreach campaigns, diversity in marketing, and much more.

Acknowledge that participants in security and privacy user studies may hold gender stereotypes with respect to security and privacy

Our study finds that U.S. participants on Prolific, a commonly used crowdsourcing platform, believe gender stereotypes with respect to security and privacy; it is imperative that future researchers and designers take this into consideration. Specifically, researchers or designers making gendered assumptions (e.g., using gendered personas, embedding assumptions about users' aptitude or knowledge), could trigger gender stereotypes about who is more likely to fall for scams, have security and privacy interest and knowledge, or adopt security tools. We recommend avoiding gendered assumptions that could bias resulting outcomes.

Develop tools for measuring individual belief in gender stereotypes in security and privacy

Tools to quantify belief in gender stereotypes in computer security and privacy could be a significant resource for researchers and practitioners working with users. Such tools could include validated scales, similar to SeBIS [179] or SA-6 [182], as well as experimental procedures, similar to the Implicit Association Test (IAT) [233]. Ideally, these validated scales or experimental procedures could be easily incorporated into a range of future research, including surveys, interviews, or other user study methods. The development of such tools is an essential long-term goal requiring

extensive effort from researchers in our field. In the meantime, our work indicates that ASI [221] or other sexism measures can proxy belief in gender stereotypes.

Request identity information as late as possible

To minimize the risk of stereotype threat, designers should ask for user identity information as late as possible in a security- or privacy-related UI flow. Otherwise, asking for a user's gender could contribute to the negatively stereotyped groups making poorer decisions. For example, if women were asked for their gender prior to an option to sign up for 2FA, given our finding that men are perceived as more likely to use 2FA, women might feel discouraged from doing so. Designers should also first ensure that collecting users' gender is actually necessary for their design's functionality.

Consider gender stereotypes throughout research and design processes

Though the mitigation of gender stereotypes for research and design resists simple solutions, we advocate for the consideration of gender stereotypes throughout security- and privacy-related processes and design flows. To promote such consideration, we advocate for researchers, designers, and practitioners to reflect on the following categories of questions.

- Laying context: How does gender appear in each part of the process? What kinds of impact will gender have in each of those places? If gender is not explicitly considered, what assumptions could be going unsaid?
- Setting goals: What is the ideal outcome, with respect to gender, for your process? How will this ideal outcome support people of all genders, not just one gender or people of binary genders?
- Heeding stereotypes: How might your process relate to gender stereotypes found in this work? Do they trigger or accidentally reinforce them? How can your process combat

stereotypes?

These questions are intended as a guide and not a comprehensive list of requirements. For further background on incorporating gender analysis into research, we refer readers to Tannenbaum et al. [561].

Stepping back. We hope that this work (a) serves to validate the experiences of people who have been at the receiving end of harmful stereotypes in computer security and privacy, and (b) serves as a call to action for researchers and technology creators in security and privacy to actively combat these stereotypes as we create and discuss products, research results, and future technologies.

6.8 Conclusion

We conducted two studies with U.S. participants on the Prolific platform to surface specific gender stereotypes regarding security and privacy characteristics and behavior. We focused on binary genders as a first investigation and empirically measured beliefs in stereotypes. We found that participants believed women were more likely to be emotional and gullible, and to take poor security and privacy actions, while men were more likely to be engaged with security and privacy topics and take protective actions. While a significant minority of participants attribute various stereotypes to biological reasons, overall, many participants believed in the validity of stereotypes for non-biological reasons. This work suggests a new direction for security and privacy research, which centers gender and other identities as critical factors in how people manage security and privacy on their computers.

Acknowledgements

We thank Chris Geeng, Kurt Hugenberg, Elissa Redmiles, and Eric Zeng for providing feedback on various drafts. We are also grateful to Joe Eckert, Philip Garrison, and Anna Lauren Hoffman for brainstorming and framing insights throughout the evolution of this paper. This work was supported in part by the National Science Foundation under grant CNS-1565252 and by a gift from Google.

Part III

Characterizing Emerging Online Abuse Threats

Chapter 7

Anti-Privacy and Anti-Security Advice on TikTok: Case Studies of Technology-Enabled Surveillance and Control in Intimate Partner and Parent-Child Relationships

Modern technologies including smartphones, AirTags, and tracking apps enable surveillance and control in interpersonal relationships. In this work, we study videos posted on TikTok that give advice for how to surveil or control others through technology, focusing on two interpersonal contexts: intimate partner relationships and parent-child relationships. We collected 98 videos across both contexts and investigate (a) what types of surveillance or control techniques the videos describe, (b) what assets are being targeted, (c) the reasons that TikTok creators give for using these techniques, and (d) defensive techniques discussed. Additionally, we make observations

about how social factors – including social acceptability, gender, and TikTok culture – are critical context for the existence of this anti-privacy and anti-security advice. We discuss the use of TikTok as a rich source of qualitative data for future studies and make recommendations for technology designers around interpersonal surveillance and control.

This chapter originally appeared as the paper "'Anti-Privacy and Anti-Security Advice on TikTok: Case Studies of Technology-Enabled Surveillance and Control in Intimate Partner and Parent-Child Relationships" at the Symposium on Usable Privacy and Security in 2022 [600]. 'We' in this chapter refers to me and the co-authors: Eric Zeng, Tadayoshi Kohno, and Franziska Roesner.

7.1 Introduction

"Is my partner cheating on me?" "What is my teenager doing right now?" "How do I access something my parents restricted?" Questions like these have long existed in interpersonal relationships, and to answer these questions, some people turn to methods of surveillance and control. In recent years, the availability and accessibility of new technologies have enabled lay users to implement increasingly invasive surveillance and control over others. For example, tracking apps like Life360 facilitate precise location tracking of other individuals, and Apple AirTags can be misused to enable the same. These tools enable violations of security and privacy boundaries through unauthorized or unintended use of technology, or by otherwise transgressing others' expectations.

In this work, we investigate a novel source of advice on how to surveil and control others' through technology: the social media platform TikTok. We find that on TikTok, users post detailed tutorials for surveilling their partners or children. Consider this suggestion to turn on the auto-answer call accessibility feature on a partner's phone to detect cheating:

> welcome to toxic tiktok 😭😭 i promise this isn't me anymore! but lemme help you out!! if he's not picking up, change this setting, it will automatically pick up all his calls! and if you hear stuff you didn't want to hear... i'm so sorry bb 😭 (TT45)

We call such videos "anti-privacy advice" or "anti-security advice": *anti-privacy* or *anti-security* because the techniques often involve violating privacy or breaking device and account security, and *advice* because the videos are presented as guidance intended to be widely seen (more examples in Figure 7.1). We sought to answer the following research questions:

1. What information or systems are being targeted in anti-privacy or anti-security advice on TikTok and by whom? How are these attacks carried out and for what reasons?
2. How do anti-privacy or anti-security advice videos fit into the ecosystem of videos on TikTok, and how do they relate to a broader societal context?

To scope our study to a meaningful yet manageable size, we use case study methods to identify two interpersonal relationships as the contexts for our investigation: intimate partner and parent-child. We collect a dataset of 98 English-language TikTok videos and use qualitative methods to answer our research questions. First, we use a deductive approach to thematic analysis to apply a threat modeling framework to understand the assets, stakeholders, techniques, and motivations. Second, we use an inductive approach to thematic analysis to generate themes about how these videos are situated in the broader TikTok and societal context.

We find that surveillance in the intimate partner context is usually surreptitious and for the purposes of detecting cheating. Techniques used include leveraging tracking apps, obtaining unauthorized access to messages, and manipulations via physical access. In the parent-child context, surveillance by the parent used family tracking apps and parental controls, is typically overt, and for ensuring child safety or restricting access to certain types of content. Meanwhile, teenagers in particular tended to resist these measures, and manipulated settings or broke authentication

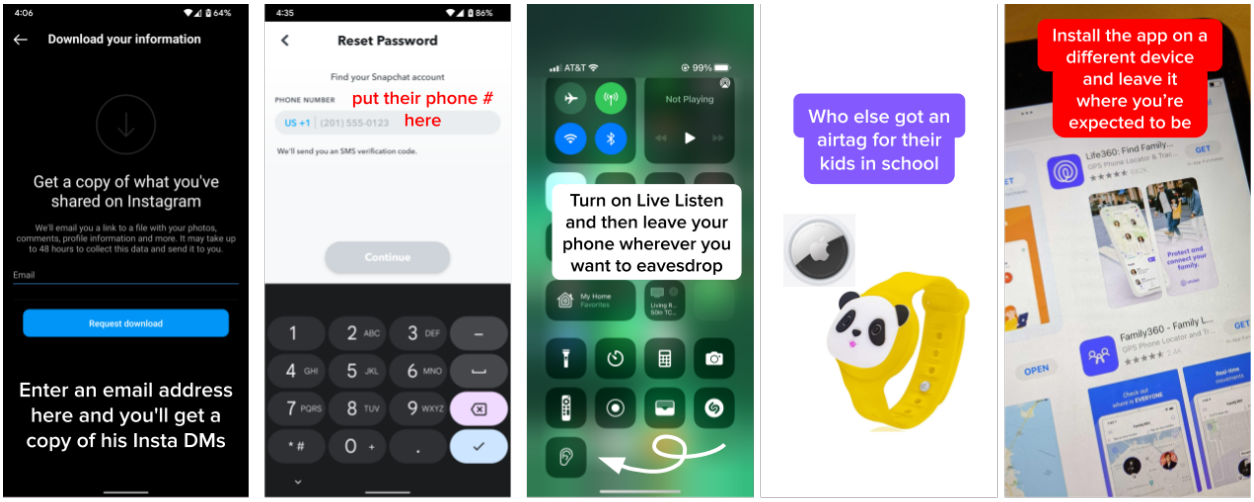

Figure 7.1: Examples of TikTok"Anti-Privacy and Anti-Security Advice," recreated to protect creators' anonymity. a) Surveil an intimate partner's ("his") Instagram messages by requesting a data download from the target's phone, and sending it to the attacker's email. b) Hijack an intimate partner's Snapchat account to view their messages by recovering targeted account from the attacker's phone, selecting "phone call" to verify the identity, and picking up the call on the target's phone without unlocking it. c) Using AirPods' Live Listen accessibility feature to surveil someone in another room. d) A parent using an AirTag necklace to track their child's location. e) A teenager evading the Life360 app by installing it on an iPad that remains at home.

measures to evade tracking.

We generate themes about three social factors contextualizing the anti-privacy and anti-security advice we found. First, we identify that social acceptability influences framing of such advice: videos in the intimate context joked about being "toxic" because surveillance of other adults is transgressive, while videos in the parenting context framed techniques as helpful "#momhacks" for child safety. Second, we examine the influence of gender, given that a majority of language in the videos was feminine-coded, and how gender expectations could have contributed to the motivations of detecting cheating and protecting children. Third, we associate the engaging, easy-to-follow, and sometimes controversial characteristics of the anti-privacy and anti-security advice videos with TikTok's competitive culture of creating viral content.

Our investigation sheds light on an ecosystem of people sharing anti-privacy and anti-security advice on TikTok. We close by discussing our findings' implications for the computer security and

privacy community and surfacing opportunities to address the risks introduced by anti-privacy and anti-security advice, while also recognizing that technical fixes will not fully address the associated social and societal challenges. We also reflect on the benefits and challenges of TikTok as a qualitative data source.

In summary, we make the following contributions:

1. We identify TikTok as a source for rich qualitative data about "anti-privacy" or "anti-security" advice, and conduct case study investigations of 98 videos about technology-enabled surveillance and control. We study two interpersonal contexts: intimate partner and parent-child relationships.
2. We identify assets, stakeholders, techniques, and motivations in anti-privacy and anti-security advice.
3. We generate themes about how these videos are situated in the broader TikTok and societal context.
4. We discuss our findings' implications, identifying opportunities in security and privacy research and practice.

7.2 Related Work

7.2.1 Interpersonal Security and Privacy

Most closely related to our work are other studies of security and privacy as indexed by specific interpersonal relationships.

Intimate Partners. A growing body of scholarship studies adversaries and their methods in intimate partner relationships. Freed et al. categorize attacks into four categories based on the resources abusers leverage and their intentions [205]. Other studies investigate spyware apps

for intimate partner surveillance (IPS) [101], as well as creepware for interpersonal attacks [494]. Tseng et al. [573] create a taxonomy of IPS tools discussed on IPS forums. In our work, we do not know if the TikTok creators giving anti-privacy or anti-security advice actually use such techniques to abuse, but we highlight the potential for such advice to do so. Our context of study is also different: TikTok is an open platform, compared to narrower populations in prior work, e.g., survivors contacting Family Justice Centers [205] or those on dedicated forums [573].

Other work examines how to effectively design interventions supporting survivors [574, 629, 253], particularly by working in consultation with survivors to map concerns [205]. Complementing these intentional efforts, the observational nature of our work allows us to see attacks organically discussed on TikTok, for informing countermeasures and support.

Many scholars studying the intimate context draw attention to its complexities. For example, intimate partner violence (IPV) targets must negotiate tensions such as seeking distance despite social, financial, or other connections to abusers [203]. Levy & Schneier highlight common privacy assumptions made by computer scientists that do not hold in intimate relationships [336]. We join these scholars by diving into the murkiness of interpersonal relationships through the content that perpetrators and targets themselves create and post on TikTok.

Contrasting prior IPV and IPS work, our dataset includes social media stalking techniques used *before* a relationship begins, perhaps more akin to the privacy of online dating [117] or online status indicators [118]. This may speak to the normalization of intimate surveillance [337] with new technologies.

Parent-Child. Many scholars have also investigated familial privacy boundaries. One body of work interrogates the information sharing that some parents engage in — sharenting — when children are younger and unable to consent [57, 77, 78], as well as the normalization of parental monitoring [332, 552]. Some scholars draw attention to the increased risk of "dataveillance" from

parents [351, 625]. Studies of parental control apps find that apps are purportedly for safety, but may favor parents' desires at the cost of childrens' [611], contributing to negative experiences [216], especially if designed incorrectly [590].

Between parents and their teenaged children, user studies of privacy boundaries find different technology understandings and preferences for monitoring or autonomy [133, 144], but also expectations that parents and children will collaborate to find the right balance [558]. The tension between parents' desire for information and control to ensure safety with teens' desires for autonomy and privacy has also been documented in the context of specific technologies, e.g., IoT entryways [582, 255], smart speakers [329]. The openness of TikTok creators allowed us to observe parents' opinions and suggestions for surveillance and control, as well as the teenagers' countermeasures.

7.2.2 Security Advice

Security and privacy researchers have studied what *pro*-security advice exists, its sources, and its quality [487, 485, 481, 484]. Other work also investigated advice for specific communities, e.g., queer individuals [212], or contexts, e.g., in workplaces [147, 148], after "triggers" [150], during civil rights protests [588, 66]. In this work, we instead study *anti*-security advice, or advice on how to compromise others' security and privacy through methods of surveillance and control.

Aside from Tseng et al.'s work on IPS forums [573], we are aware of little academic work studying how security and privacy adversaries learn. Some low-tech techniques in videos we study call to mind advice from other contexts, e.g., social engineering and low-tech hacking guides [348].

7.2.3 TikTok

As TikTok is only 5 years old, TikTok research is still in its early stages. Some study specific subcommunities, e.g., populations with disabilities [171], healthcare workers [546], or aspects of TikTok's culture, e.g., authorship practices [308], visibility [3]. Other work leverages TikTok as a respository for specific content, e.g., public heath messaging [342, 36, 58], social activism [125], science memes [623], political communication [520]. We add to this growing body of work by studying anti-privacy and anti-security advice: content that teaches how to surveil or control others through technology. De Leyn et al. study tween privacy perceptions, but in conjunction with parents [340], whereas this work studies when parents may pose the privacy risk.

7.3 Background

TikTok is a social media platform on which users post short-form videos (also called "TikToks"). In early 2020, TikTok became the most downloaded app in the world, and reached 1 billion monthly users in late 2021 [352], demonstrating enormous growth relative to older social media platforms. As of early 2022, 35% of TikTok's users are between 19 and 29 years old and an additional 28% are under 18; only 18% are between 30 and 39, and 19% are over 39 [290].

Usage. TikTok's primary interface is the For You Page (FYP), an infinite scroll feed of autoplaying videos. The FYP serves videos using a recommender system, which personalizes recommended videos based on engagement metrics such as dwell time, likes, and comments. Content can also be viewed in the Following tab (to see content from previously followed creators) or the Discover tab (to search for videos or see trending topics). TikTok displays videos full screen (on mobile), and it is only possible to watch TikToks one at a time, swiping up to display the next video.

In addition to the video (often showing the creator in portrait mode), TikToks frequently

include overlaid text (which may be read aloud by a built-in voiceover feature), TikTok's own set of sounds (including licensed music), and various visual effects. Users can interact with content by liking, commenting, or sharing videos; following TikTok creators; or remixing other TikToks.

TikTok subcommunities. Subcommunities on TikTok are loose associations of creators and followers interested in a specific topic, often organizing around certain hashtags, e.g., #egirl (rebellious women gamers turned fashion aesthetic), sometimes with a play on the platform name, e.g., #momtok (moms on TikTok), #fittok (fitness TikTok). Relationships are one such subcommunity, with users posting anything from inspirational relationship content, to giving advice, to calling out toxic behaviors. The top relationship-related hashtag is #relationship with 90.1 billion views. Another subcommunity discusses various aspects of parenting, including sharing advice or personal experiences. The top parenting-related hashtag is #parenting with 13.0 billion views.

7.4 Methods

We investigate anti-privacy and anti-security advice on TikTok through case studies of two interpersonal contexts. We selected these contexts informed by case study methods and collected a total dataset of 98 TikTok videos (see 7.4.1). For data analysis, we performed procedures from the qualitative methods family of thematic analysis (see 7.4.2). Although our research did not directly recruit participants, and as such, our institution's IRB determined our work not to be human subjects research, we still recognize that we are studying real people: we carefully made ethical considerations to protect the subjects of our research (see 7.4.4). We conclude by contextualizing the goals of this work with its limitations (see 7.4.5).

7.4.1 Case Selection and Data Collection

We summarize our overall approach to data collection, which occurred between November 2021 and February 2022.

We used progressive focusing [550], an approach from case study methodology, to iteratively narrow our research questions as well as select which cases we used. In his influential 1995 book, *The Art of Case Study Research*, Stake describes progressive focusing to place a high emphasis on interpretation that allows for flexibility during the research process because "the aim is to thoroughly understand [the case]. If early [research] questions are not working, if new issues become apparent, the design is changed." [550]

In this work, our case was centered on English-language TikTok videos that described technology-enabled techniques for harming others' digital security or privacy, i.e., anti-privacy or anti-security advice. Our criteria for inclusion of a TikTok video as anti-privacy or anti-security advice were: (a) does the video describe a technique that requires technology,[1] (b) does the technique involve violating privacy or security measures or boundaries, and (c) does the technique implement (or evade) surveillance or control?[2]

Initially, we tried searching for security and privacy related terms using the built-in TikTok search interface to surface relevant videos: e.g., "hacking," "security," "violate privacy," "surveillance." These terms are meaningful to the computer science community, but we discovered they were not to TikTok creators nor viewers. Instead, we realized that we would need to first identify contexts in which anti-privacy or anti-security advice could be common, and then find videos in those contexts that included technology-enabled techniques.

We conducted a literature search to identify contexts in which anti-privacy or anti-security

[1] Thus, we excluded videos without a technology element.

[2] Thus, we included videos where the technique was been demonstrated in the video with consent, but could also be used without consent.

advice could be common. We considered the following contexts (that we did not include): smart homes, proctorware, hidden cameras in vacation rentals. We searched for videos in these contexts, finding the most qualitatively rich videos in intimate partner and parent-child relationships, which we finalized as our cases.

We collected more data by adding context-specific search terms to our original set: in the intimate partner context, e.g., "toxic," "relationships," "cheating," and in the parent-child context, e.g., "parental controls," "life360," "kid tracking." Data collection was an iterative process between two members of the research team, who recorded relevant search terms and frequently met to discuss data collection efforts.

The majority of data collection concluded when we felt that we had exhausted the relevant search terms and could not find more videos, and that we had a rich enough dataset for analysis. Drawing from case study methods, we continued triangulating — "working to substantiate an interpretation or to clarify its different meanings" [550] — throughout our analysis and writing. By iteratively searching for relevant videos to confirm or deny our findings and interpretations, we continued to make refinements and added 21 videos in this manner. Our final dataset consisted of 98 anti-privacy or anti-security advice videos: 66 videos in the intimate partner context, 27 videos in the parent-child context, and 5 relevant to both. Altogether, our dataset accounts for 60 minutes and 14 seconds of audio-visual content, with a total of over 16 million likes (mean = 171K, median = 4.5K, max = 3.2M). For reporting, we abbreviate the xth TikTok in our dataset to TTx. We note that our dataset is a case study, and prioritizes qualitative depth over quantitatively measurable claims.

7.4.2 Data Analysis

We conduct thematic analyses of our data, a broad family of methods that is flexible with respect to conceptualization of the data and its meanings, inductive or deductive orientations, and the procedures that can be used [70, 71].

Deductive Thematic Analysis. The first part of our analysis focused on our first research question about (a) what information or systems are being targeted, (b) by whom, (c) using which techniques, and (d) for what reasons. We used a codebook approach [70, 71] to deductively (theory-driven) apply a security threat modeling framework to our data. Because of the significant theoretical value of this framework to security and privacy researchers and practitioners, the codebook approach permitted us to develop these questions early in the research process. First, two coders familiarized themselves with the videos by watching them multiple times, taking notes separately (this initially began concurrently with data collection). They then met multiple times to develop four codebooks: stakeholders, assets, motivations, and techniques. Using these codebooks, one coder coded intimate partner videos, the other coded parent-child videos. Lastly, both coders reviewed each others' work, discussing and resolving concerns.

Inductive Thematic Analysis. For the second part of our analysis, we used a less structured approach to inductively (data-driven) generate themes about the social factors that contextualize the anti-privacy and anti-security advice we collected on TikTok. We did this by continuously meeting with all members of the team to discuss higher-level observations we made about the data, and drafted memos about these broader ideas. Through this iterative process [397], we developed three themes about the social context of such advice (Section 7.7).[3] To ensure thoroughness, we also triangulated [550] these themes by going back to do more data collection, or add new elements

[3]Due to the deductive thematic analysis approach we used for applying the threat modelling framework to our data, as well as the observational nature of TikTok videos, we did not conduct a fully reflexive thematic analysis [68].

of analysis, as necessary. For example, to triangulate our findings about the gender in Section 7.7.2, we went back to the data with a gendered lens.

7.4.3 Positionality Statement

In the process of our inductive thematic analysis in particular, as well as our overall research approach and perspective, we acknowledge our active role as researchers in the process of knowledge production [68] and regard our "subjectivity as analytic resource"[71]. Our research analyses and interpretations are the result of our particular social, cultural, historical, disciplinary, political, and ideological positionings [71]. Here, we describe our identities and how they relate to the interpersonal contexts (i.e., intimate partner and parent-child) and research data (i.e., TikToks) we study. Our research team is composed of two cisgender women and two cisgender men. Two researchers are in their 20s, one is in their 30s, and one is in their 40s. All researchers have experience with intimate partner relationships and two are parents. One researcher has 24 months of experience with TikTok, another has 6, and another has 3 at the time of these analyses. [4]

7.4.4 Ethical Considerations

We consulted with our institution's IRB, which determined that our study did not require review as human subjects research because the videos that we analyzed were publicly available at the time that we collected them. However, we recognize that IRB review is not sufficient to guarantee ethical research. In particular, there are ethical considerations with studying public data that was created and shared for purposes other than research [81], even if many of the videos we study have reached large audiences in the context of TikTok (and beyond — we observed some news articles about creators in our dataset). To mitigate potential harms that may come from

[4]The other co-author first heard about TikTok through his collaborators and only accesses it through links provided by the other three.

exposure of the content we study to unexpected audiences, we paraphrase creator quotes and recreated screenshots of the videos in this chapter, to preserve semantic meaning while obscuring the original source. We also aim to present our data in broadly descriptive or interpretive, rather than individually judgmental, ways — we recognize that there is additional context behind the motivations and situations of creators and viewers of the content we study that we may not fully understand. Ultimately, our goal is not to study the specific people who post or engage with this content, but rather to use this data as a window into popular use of interpersonal control and surveillance techniques more generally.

Our research also surfaces complicated social ethics considerations. The surveillance and control techniques we study have a tangled relationship with the interpersonal situations they are embedded in, including non-consensual surveillance, cheating, child safety, and fostering trusting familial relationships. Our work cannot resolve these ethical questions, but as security and privacy researchers, our goals are to enable an informed conversation about security and privacy risks, and hope that our findings contribute to a better understanding of the use of surveillance and control techniques.

7.4.5 Limitations

Our investigation necessarily considers only a slice of data from TikTok, focusing on specific subcommunities, at a specific point in time, and limited by the videos we were able to surface via our data collection methodology and TikTok's search capabilities. There are likely relevant videos on TikTok that are not included in our dataset, so there may be motivations or techniques that we missed. Moreover, there may be other related subcommunities that our searches did not surface, e.g., communities who respond to the videos we analyze or create similar videos in other contexts. Accordingly, our analysis focuses on surfacing the breadth and depth of interpersonal surveillance

and control motivations and techniques that the videos we study cover, not on understanding TikTok as a whole or on comparisons with different subcommunities.

Additionally, content on TikTok is, as on any social media platform, created and edited in order to present people and the topics they are discussing in a certain way. Our study uses TikTok data as a window into people's motivations, techniques, and responses to interpersonal surveillance and control, but (of course) does not give us information about the creators' actions or opinions beyond what is projected in the videos.

Finally, we come to TikTok and to our research questions as observers, not as TikTok content creators ourselves. There are likely unique aspects of content creation that we do not understand. However, as mentioned, several of us have significant experience immersed in TikTok as passive users.

7.5 Findings from the Intimate Partner Context

We collected a total of 66 TikTok videos in the intimate context. Of these, 64 were about implementing methods of surveillance and control, while 2 were about defenses. These videos were created by 25 unique TikTok creators: 18 came from Creator A, the most prolific creator; 9 came from Creator B, the second most prolific; 8 came from Creator C; and 1 video each came from seventeen creators.

7.5.1 Stakeholders, Assets, and Motivations

We present a summary of the stakeholders, assets (and associated technologies), and motivations in Table 7.1.

Explicit and Implicit Concerns about Cheating. In the videos we collected in the intimate

Table 7.1: A summary of the stakeholders, assets (and their associated technologies), and motivations we observed in our dataset. This table is intended to give a sense of the broader context and attack space; we note that our methods were qualitative and thus these results are not able to make exhaustive claims about what attacks are possible, nor quantitative claims about frequency.

	Intimate Partner Context	**Parent-Child Context**
Stakeholders	**Instigators** surveil **targets**' data or digital footprint, or otherwise exert control on targets' digital activities	**Parents** are the caretakers of **children**; childrens' ages ranged from early school age to teenagers
Assets	Location; social media accounts; social media data (who targets followed, messaged, or content targets posted); web browsing history; photos; live audio; dating app usage	Location and location privacy; access to specific types of content; access to communications; privacy about digital activities
Technologies Targeted or Used	Apple software and devices (iOS, iPhones, AirTags, AirPods, Apple Watches); Android (Google Maps); social media platforms (Instagram, Facebook, Twitter, Snapchat, Tinder); email; phone calls; family monitoring apps	Apple devices (AirTags); Life360; Bark; FamiSafe; parental control features; VPNs
Motivations	**Instigator**: Detect cheating; general surveillance; control contact with targets **Target**: Evade surveillance; maintain autonomy	**Parent**: Child safety in the physical world and online **Child**: Autonomy; privacy

partner context, **instigators** are interested in obtaining information about **targets**, primarily to detect cheating. Cheating concerns were sometimes made explicit by using the words "cheating" or "suspicious" (or variants thereof). We observed that many videos began with this motivation, e.g., "Do you wanna find out if your partner cheats?" (TT36), potentially to capture a viewer's attention. Sometimes this motivation arose later, e.g., the instigator in TT18 says, "keep watching if you wanna find all Twitter conversations between your partner and someone you're suspicious of." The creators also made their motivation as instigators explicit by naming an audience member's relationship to a target, e.g., "How to figure out if your partner is cheating on you" (TT10).

In other videos, concerns about cheating were implicit: for example, by implying a target's identity by their gender: "Trying to get into his Snapchat?" (TT38). Some videos included techniques that were substantively similar to those in videos explicitly motivated to detect cheating, or sought to find evidence of cheating behaviors (e.g., communicating with someone else, being at

certain locations) or contained context clues about catching a target, e.g., "Heh you can't hide from me dummy 🤪" (TT34).

Targeted Assets. Instigators sought to compromise a variety of targets' assets: aligned with the motivation of detecting cheating, instigators creatively postulated all the digital traces that could be treated as proof, including sexually explicit photos or emails from hookup websites. Location in particular was treated as more conclusive proof if instigators used technology to verify that targets had been at suspicious locations. Social media assets, such as who targets followed or messaged, were used sometimes as less conclusive evidence, e.g., "as a preliminary step to confirm or deny my suspicions, before I get into a full investigation" (TT40).

Other Motivations. A minority of videos were not motivated to detect cheating, and were instead about general behaviors of surveillance and control in intimate relationships. These behaviors may cross targets' personal boundaries, breaking their existing security measures or invading their privacy, either because a target would reasonably assume certain information to private, or in some cases, because a target had explicitly set that boundary. Some instigators sought to surveil targets at all hours of the day, even absent suspicions of cheating, or generally spy on as many of their target's digital activities as possible. Targets' motivations were to maintain autonomy, especially in the face of potential surveillance.

7.5.2 Intimate Surveillance and Control

Next, we break down the specific surveillance goals and techniques of instigators. We observed at least 24 distinct techniques for surveillance and control, underscoring the variety and creativity of instigators in this context. Though we do not pose this is an exhaustive list of all techniques discussed on TikTok, we detail these techniques to surface the breadth of how instigators surveil and control their targets. The full set of goals and their associated techniques are in Appendix B.1.

Goal: Surveil Digital Communications

Instigators were interested in learning who targets were communicating with, and what those communications contained, (presumably) to determine whether they were texting with a affair partner. Several methods were suggested for obtaining information about the targets' SMS or social media messages.

Technique: Exploit Data Downloads. One method for obtaining a target's messages and communications was through the data download feature of social media platforms: GDPR's Right of Access requires data subjects to be able to download archives of their data. Instigators noted that on platforms like Instagram, Snapchat, and Facebook, these data downloads can be used to obtain a copy of their messages, allowing them to search for evidence of cheating (Figure 7.1a). Three separate creators made tutorials for locating the data download in the settings interfaces of the above platforms. This attack relies on having physical access to the device or account access.

Technique: Gaining Direct Account Access. Another method for obtaining a target's messages was to obtain direct access to the target's social media account to view the target's messages in the app. One video describes hijacking the target's Snapchat account through the account recovery process, which only requires physical access to their phone (Figure 7.1b). The instigator attempts to recover the account password on their phone. Snapchat sends an authentication code via phone call, which the instigator can pick up without unlocking the phone. After confirming, the instigator can reset the target's password, accessing the target's Snapchat messages. Another approach suggested is to add the instigator's phone number to the target's iCloud account, which may enable the instigator to get a copy of their messages.[5]

Technique: Emoji Side Channel. Two TikToks suggest the target's frequently used emojis in their keyboard as a side channel for detecting cheating. If sexually suggestive emojis (e.g., 🍑,

[5]This technique does not work without also enabling message forwarding, which requires additional authentication.

🍆, 💦) were present and the target did not use them while communicating with the instigator, it suggests the target is sexting with someone else. This technique only requires non-privileged physical access to the phone: TT40 suggests opening an iPhone's "Today's View," accessible from the lock screen and containing a keyboard in the search bar.

Goal: Stalk on Social Media

Another goal for instigators was to stalk a target's activities on social media, either for generally monitoring their online presence, or for specifically finding evidence of cheating.

Technique: Read Twitter Conversations. One video suggests using Twitter's advanced search to find conversations between two specific people, to look for evidence of cheating.

Technique: Anonymous Viewing of Instagram Profiles. Instigators may be interested in viewing their targets' Instagram profiles; however, activity like following or viewing stories is visible to the target. To view stories anonymously, one video suggested creating a fake Instagram account to watch stories, while another suggested using a third-party site that claims to allow anonymous viewing. A different third-party site was suggested for enlarging a target's profile picture, which are usually only shown in a small size through the app.

Technique: Side Channels in Social Media Platforms. Other videos highlight side channels that leak information about the target's activity. For example, an instigator could determine the order in which a target follows other accounts, by viewing their "following" list on the web version of Instagram, which shows follows in chronological order.[6] Another video suggests that instigators can infer whether a target is sending Snapchat messages (e.g., sexts) to a large number of people or to an individual, by tracking the target's Snapchat score over time, and observing how much it increases.

[6]This is no longer works as of the writing of the original paper.

Technique: Track Online Status Indicators. Instigators may want to know when a target is online on a messaging app to infer other aspects of their behavior (e.g., are they actually asleep, or did they lie about it?). One instigator names a third-party app that specifically sends notifications each time a WhatsApp contact signs on or off.

Technique: Contact Someone Who Blocked You. One video demonstrates texting someone who blocked you by sending from an associated iCloud email address.[7]

Goal: Surveil Dating App Usage

Instigators presented techniques to infer whether targets were using dating apps despite being in a relationship with them.

Technique: Find Target's Profile on Dating App. One approach is to find the target's profile on the dating app. One video suggests creating a fake account on the dating app, and swiping through profiles manually. They also suggest setting the search radius to the minimum while physically near the target narrow down the available profiles as much as possible. Another suggests a paid third-party service called "CheaterBuster" that will look for the target automatically.

Technique: Infer Dating App Usage. Other videos suggest more indirect approaches. One video suggests attempting to create a dating app account with the target's email address to see if the email address is already in use, indicating they are signed up for that service. Another suggests looking through the App Store for dating apps — the list of downloaded apps shows not only which apps were installed, but when they were first purchased or installed. This would indicate if they recently installed a new dating app.

[7] According to many comments, this technique does not seem to work.

Goal: Surveil Other Digital Activities

Instigators also aimed to surveil targets' other digital activity, including monitoring their browsing history for watching porn, and searching their phones for sexually explicit content.

Technique: Searching for Explicit Content. Some videos instructed viewers to look for explicit photos in the photo gallery, as well as explicit content in the target's email and web browsing history. One video warned viewers of an app that could hide explicit photos while appearing to be a calculator, and noted that observing a target's reaction to being asked about whether they had this app might be informative enough.

Technique: Photo Metadata. One video suggested an app that automatically parsed EXIF data to show when a photo was originally taken, which allows inferring whether a sexually explicit photo had been, according to the instigator, "reused": "let's say you get a pic of their nuh-uh today, but if the pic was taken five months ago, who else might've gotten that pic, hm?" (TT16).

Goal: Manipulate Social Media

Instigators creatively manipulated the functionality of social media and messaging apps to obtain outcomes they desired.

Technique: Restrict and Unrestrict. Two videos advocate reading an Instagram direct message by blocking the sender, which then sends the message to a request inbox that does not send read receipts. Similarly, another advocates manipulating a target's Instagram story feed by hiding a story from the target, and then unhiding, which makes the story appear first.

Technique: Fake Tags. One video describes creating a fake "tag" with the poll feature in an Instagram story that appears to be tagging another user, but instead tallies how many people clicked on the fake tag.

Technique: Message Deletions. One video describes how to delete WhatsApp messages more than an hour old: changing the system time to within an hour of the message timestamp.

Goal: Surveilling Physical Activities

Instigators were also interested in surveilling targets' physical-world activities, such as their physical location, or hearing their conversations, which could provide evidence of cheating.

Technique: Tracking Location with Apple Products. A very common technique described by instigators is to use AirTags, AirPods, or Apple Watches to track a target's location. This is done by secretly hiding one of these in the target's belongings or car (one video demonstrates hiding it in the side pocket specifically). TT25 acknowledges that this would be "super toxic," but one could "forget, on accident of course, an Apple device in their car and then track their every move." In another notably overt example, an instigator makes an AirTag necklace with a customized design, names the AirTag "Cutie pie 🥧", and gives it as a present to her boyfriend. We also observe one instigator discussing an unsuccessful attempt, as Apple's mitigation alerted their target that they were being tracked, and later found the AirTag discarded in a bush.

Technique: Abusing Accessibility Features to Spy on Audio. Instigators developed techniques for surreptitiously listening to their targets' conversations. Some videos advocated for using Live Listen, an accessibility feature which enables an iPhone or iPad to act as a microphone to send sound to AirPods (intended for use with hearing aids, or in a noisy location). An instigator could leave their phone with the target, leave the room, and listen via AirPods (Figure 7.1c). Others suggested taking the targets' phone, enabling Auto-Answer for phone calls (intended for Touch accessibility), and calling them whenever they wanted to listen to what they were doing.

Technique: Use Tracking or Monitoring Apps. Three videos advocate installing location monitoring apps (e.g., Life360) or using OS-level tracking features (e.g. Find my Friends) on

partners' phones. These videos report the location of a target in real time. Another strategy suggested by instigators was to use the iOS Significant Locations feature or Google Location History to identify locations that the target visited in the past, which could reveal if the target had been dishonest about where they had been.

7.5.3 Countering Intimate Surveillance

We now review targets' strategies. In the 2 videos we collected, targets' goals were to counter surveillance. These defenses do not counter any of the instigator techniques we found, which could be a result of our methods (Section 7.4.5), and does not necessarily mean such content is not on TikTok.

Technique: Detect call surveillance. Two TikToks described checking phone carrier settings to check for call forwarding or redirection. However, the videos did not suggest purposeful next steps if found: "if any are enabled... scream" (TT54).

7.6 Findings from the Parenting Context

We collected a total of 27 videos in the parent-child context; 16 from parents, and 11 from children. These videos were posted by 25 unique TikTok creators, distinguishing this context from the intimate partner context where three creators accounted for over half of videos.

To facilitate comparison with the intimate context, we standardized our terminology to use "surveillance" and "control" for methods used by parents to track, monitor, or restrict their children's activities. In the parent-child context, these methods are more ethically ambiguous than the intimate partner context, and may not always be adversarial. The appropriateness of certain methods may depend on the age of a child or the overall nature of the parent-child relationship.

Though some creators shared techniques with positive intentions, viewers may not necessarily share those intentions. Further, such videos may contribute to the normalization of parental surveillance [552].

7.6.1 Stakeholders, Assets, Motivations

In the parenting context, we observed videos from **parents** and **children**, primarily teenagers (old enough to have a smartphone and a TikTok account). Tensions centered around parents having the right level of information and control to ensure childrens' safety, while children wished to have enough autonomy to ensure their own privacy. A summary of the stakeholders, assets, and motivations is again in Table 7.1.

Parent Perspective. When children were younger, parents were concerned about physical safety and leveraged technologies to track their location, especially when not in their supervision, e.g., riding the bus to school. Some captions alluded to more general concern: "Extreme measures are essential these days. Track kids with #airtag bracelets" (TT57). As children got older, concerns centered more on access to certain content, so some parents relied on family tracking apps, parental control features, or other technologies made for these concerns. Parents were concerned about children accidentally downloading malware or making purchases, messaging strangers, using rude or profane language, encountering explicit material, and having excessive screen time.

Child Perspective. Children's videos were motivated to evade tracking or restrictions by a desire for greater autonomy, particularly in the face of restrictions (e.g., on internet and app usage) and tracking software (e.g., for location) on their phones. Children were also motivated to hide their apps and texts from low-tech monitoring, like manual inspection by parents.

7.6.2 Parental Surveillance and Control

We now describe the specific goals parents had regarding child safety, and the techniques and tools used to reach those goals. Again here, we do not pose this is an exhaustive list of all possible techniques, but rather detail them to surface their breadth. Generally, parents used commercially available tracking and parental control tools, or parental control features built into mobile operating systems. Compared to the intimate partner context, parents typically used these features as intended, rather than abusing features. The full set of goals and their associated techniques are in Appendix B.1.

Physical Surveillance

Parents were interested in knowing the exact physical location of their children, for emergencies or general peace of mind.

Technique: Location Tracking with AirTags. Many of the videos from parents advocated using AirTags in order to keep track of their children's location, touting how cheap, accessible, and effective they were: "#Apple #AirTag this is so smart, only $30, so worth it ❤️🙌🏼" (TT60). Essentially all of these were made by moms for younger children (younger than preteen) and a few described this technique as a "mom hack." As noted above, the motivations were to keep children safe. The parents mainly showed their personal experiences of making an AirTag bracelet, keychain, or necklace and putting it on their child (Figure 7.1d), while a few also showed putting (or hiding) and AirTag in their child's bag or shoes. One in particular noted that a keychain attached to their child's belt loop, instead of backpack, was the best option "because backpacks are always left behind when something happens" (TT65). We suspect that parents chose to use AirTags with younger children because they do not yet have smartphones with which tracking apps can be used.

Technique: Location Tracking with Apps. For older children, parents described using specialized mobile apps, especially Life360, to monitor their activities. Life360 is advertised as a family location sharing app, which also provides emergency assistance alerting and digital safety tools to monitor identity theft or credit scores. One parent described using Life360 to monitor their kids while they went to school and extracurriculars (TT63).

Goal: Online Safety and Monitoring

Parents are also concerned about kids' online safety, and employed a variety of apps and tools to restrict access to the internet and apps, and to monitor communications.

Technique: Monitoring and Parental Control Apps. Some parents described using third party apps to impose parental controls and monitoring to their kids' smart phones. Apps mentioned include FamiSafe and Bark, which are advertised as online safety apps that monitor social media content for appropriateness as well as time limits on certain apps. Bark alerts them if profanity was detected: "privacy with a safety net" (TT97). Another set of parents created a sponsored video where they describe using FamiSafe's app download allow list to restrict their kids to trusted apps (fearing that their child might install malware on their phone).

Technique: Fully Locking Down Phone. One parent advocated for a fully locked down phone from Gabb Wireless, which had built-in parental control tools for screen time restrictions and content filters (including no access to any social media platforms), while still allowing for some phone functionalities like calling and texting.

Technique: Monitor Messages with System Features. Parents could also use built-in operating system features to perform monitoring of their children. One video explained how to monitor a child's text messages: parents can add their phone number to the child's iCloud account, and then update the settings to forward all messages to the parents' device(s).

7.6.3 Children's Defenses

Teenagers' primary goal in our dataset was to evade surveillance or restrictions placed on their phones by the parents; such as location tracking apps or parental controls. These techniques were generally reactive, not proactive, to parents' usage of certain commercial products or device features.

Technique: Disrupting Location Tracking Apps. Children described a number of ways to evade location tracking apps like Life360, e.g., disabling cellular data and motion and tracking permissions for Life360, while leaving location and WiFi permissions on. This prevents the app from reporting back real time location updates, but does not notify parents that the location permission was disabled. Another technique was to install the Life360 app on another device that could be left at home (Figure 7.1e). Another video claims that putting the iPhone in Do Not Disturb mode would disable tracking, though commenters disputed this method.

Technique: Bypassing Parental Controls. Teens also found techniques to bypassing parental controls, which may restrict screen time, app downloads, or access to certain websites, depending on the software and how the parents configure it.

Two children described guessing the parental control passcode by examining the fingerprints left by their parents. One suggested wiping a screen perfectly clean, and another by getting a screen very dirty, and then asking parents to unlock or temporarily allow access to apps. Then, by looking at the location of the fingerprints, they systematically guessed the possible combinations. For parental controls that use a VPN to intercept web and message history, like Bark, one video suggested removing the VPN in the system settings. Lastly, to bypass App Store restrictions on which apps can be downloaded, one user suggested signing out of their iCloud account, logging into a new iCloud account to download the app, and then signing back into their usual account.

Technique: Hiding Digital Activity with OS Features. Two children advocated for a technique

specifically for when parents ask to see their phone. To hide certain apps, the children described an iOS feature that hides certain homepage screens, so that the parent would not see certain apps.

7.7 Social Context of Anti-Privacy and Anti-Security Advice

We now present themes from all 98 videos across both settings, stepping back to consider broader social contexts.

7.7.1 Social Acceptability

Though on a technical level, videos in our dataset all contain advice on breaking or potentially misusing computer security and privacy features, we saw notable differences in how socially acceptable the creators perceived their advice to be, and whether the techniques were meant to be covert.

Intimate Partner Hacking: Socially Unacceptable, Covert. In the intimate partner context, creators often demonstrated performative self-awareness about how their videos were taboo, transgressive, or could be illegal or considered violations of privacy. Captions for these videos often included hashtags or phrases like "#toxic", "#stalker", "#crazygirlfriend" (referring to self), or "#hacks". Some creators put disclaimers at the beginning of videos or in their account profiles, declaring that their videos were not to be taken seriously:

> Disclaimer: Techniques shown here should not be replicated. If you are actually crazy, you should probably get medical help. These videos are only for entertainment and informational purposes. Use this as you will. (TT19)

Techniques used by instigators in the intimate context often had covert objectives, such as viewing content anonymously, secretly getting unauthorized access to a device or account, or

abusing existing features like platform user blocking.

Parental Surveillance and Restrictions: Socially Acceptable, Overt. In contrast, videos about anti-privacy or anti-security advice in parent-child relationships were not framed as deviating from social norms. For parents' videos, because the motivations of child safety are widely accepted, creators tended to frame their videos as helpful tips: "I really strongly recommend using AirTags if you have a kid going to school on public transit" (TT64). The techniques and tools used by parents, such as Apple AirTags, parental controls on smartphones, and apps designed for family tracking or child safety, like Life360, are commercially available, and used for their intended purpose, rather than covertly used or misused. Rather than secret surveillance methods, parents openly put AirTags on their childrens' wrists or clothing or enabled parental controls on their childrens' phones.

Teens Evading Surveillance and Control: Socially Acceptable, Covert. In teenagers' videos on evading restrictions and tracking, although their techniques were often intended to be covert and undetectable by parents, none of the creators framed their videos as socially unacceptable. For example, multiple videos gave advice for disabling location monitoring in the Life360 app so they could leave the house without alerting their parents. The techniques were intended to be discreet, but the creators did not portray doing so as ethically wrong.

Why These Differences? The norms around privacy in the intimate partner context differ substantially from the parent-child case. In the intimate relationships, both people involved are adults with autonomy and reasonable expectation of privacy, and many of the suggested techniques seem to overstep social and legal norms among adults (especially without consent). Meanwhile, by biological, social, and legal norms, parents are responsible for the care of their children. So techniques for parental controls and surveillance fall within the norms for parenting, even if individual parents would disagree on the balance between control vs. autonomy, and safety

vs. privacy. Similarly, teenage children rebelling against parents is well within social norms, even if done in secret.

7.7.2 Gender

We observed that TikTok creators framed their videos from a femininized and heteronormative perspective. The videos we collected predominantly used feminine language and were targeted to a feminine audience. Given the limitations of our method, which is observational about TikTok videos, we refrain from assuming the gender identities of creators. Instead, we qualitatively discuss the *feminine* (as opposed to *masculine*) coding of the video content, in alignment with scholarship on gender performativity [85] and in particular, gendered language (e.g., [209, 390]).

Specifically, we observed that many creators in the intimate partner context used feminized language towards *themselves*, e.g., #crazygirlfriend, "she's back," and masculinized language to describe the *targets* of their strategies, e.g., "the boys aren't gonna like what I'm about to share with you" (TT23). Additional videos presumed the audience to be women in relationships with men: "ladies, the goal here is to manipulate the algorithm, sorta like the way men manipulate us" (TT39).

In the parent-child context, most creators used feminized language when referring to themselves, e.g., #momhack. One creator described using AirTags to track her daughter's location on the weekends when her ex-husband had custody of the daughter. Many implicitly associated their motherhood with the role of ensuring their children's safety, calling for other mothers (and not fathers) to follow their advice.

Why Feminine-Coded? We propose two explanations: First, society prescribes gendered dynamics for the relationships in which these tutorials exist (romantic relationships, parenting). Historical gender roles place significant burdens on women to do emotional labor in sustaining

hetereosexual relationships and to compromise or make behavioral changes whenever relationship issues arise [591]. Similarly, childcare and other domestic labor typically falls on mothers [42]. Further, the predominant motivations in these interpersonal contexts were to prevent cheating and ensure child safety, implying that if women did not carry out their gendered responsibilities, negative consequences should be blamed on the women (instead of on the men or children also in these relationships) or that men default to infidelity and children to danger.

Second, there could be selection bias in our data collection. It is possible that our search keywords or hashtags were somehow biased to mainly find videos containing gendered language or performative displays associated with women. However, even when we returned to data collection to find more videos containing gendered language or performative displays associated with men — to triangulate (see Section 7.4.1) this finding — we were not successful in surfacing them.

7.7.3 TikTok Culture

The aesthetics and substance of the videos in our dataset are strongly shaped by TikTok's attention economy dynamics: there is significant pressure to make viral content, optimized for TikTok's recommendation system.

Strong Emotional Appeals. The creators in our dataset tend to make the stakes or potential outcome of listening to their video clear from the very start of the video. On TikTok, getting to the next piece of content only takes one quick swipe, so creators very often say or show something engaging in the first few seconds of a video, e.g., "Think he's a cheater? I got u girlie" (TT6) or "PROTECT YOUR CHILDREN!!! ALWAYS WATCH THEIR LOCATION!" (TT65).

Controversial Content. Another established way to increase popularity is to be controversial, and indeed, the very nature of anti-privacy and anti-security advice is controversial. This can be

seen in the comments to videos we studied, where some disagreed with the creator, e.g., "not good in any way, this is super toxic" (comment to TT3) or otherwise passed judgement: "say you're controlling and have low self-esteem without actually saying it" (comment to TT5).

Multi-Modal Content. On TikTok broadly, as well as within the videos in our dataset, content is intensely multi-modal. Videos often have music and captions that support the overall message of the video, as well as concurrent audio speech and text overlaid on the screen. Anti-privacy and anti-security advice videos further contained screenshots and screen recordings, overlaid with annotations. This means that a viewer needs to take in multiple streams of content at once, sometimes watching the video multiple times to catch everything.

Subcommunities. Creators and influencers seek to cultivate a unique (and large) audience, which can lead to the development of subcommunities. For example, the creator of one series began the videos with, "Welcome to [name of video series]", asserting that the viewer had entered an established digital space. In another video, a creator referred to populations of their viewers: "junior toxics" who needed to learn from "senior toxics" about the "toxicity basics," because after all, the senior toxics had a "legacy to uphold." Unlike structured communities on platforms like Reddit or Facebook, TikTok subcommunities exist fluidly and organically, using the same hashtags, commenting on videos, and responding to each other (e.g., in the forms of TikTok "stitches" or "duets").

7.8 Discussion and Conclusion

Our work sheds light on a part of TikTok where creators give anti-privacy and anti-security advice around surveillance and control in interpersonal relationships. We believe that studying, documenting, and describing how people use (or misuse) technology today, and exploring ecosystems

like the ones we see here within TikTok, is intrinsically interesting and valuable. We also draw from our findings concrete implications for security and privacy research and practice.

7.8.1 Implications and Recommendations

The surveillance and control techniques used by stakeholders in our case studies show ways that existing solutions are insufficient for preventing harm. What can or should be done?

Designing for strong interpersonal adversaries with physical access. Our work provides additional evidence and concrete examples of how adversaries with physical access to devices are a realistic threat for regular people, occurring commonly in both contexts we studied. Threat models should take physical access seriously for assets like location and communications privacy — these are not just at risk for people who expect to be targeted by (for example) intelligence agencies.

To raise the bar for attacks relying on physical access, apps and operating systems could require additional authentication at privacy and security sensitive points, such as for data downloads. But while such mitigations may make some attacks more difficult — e.g., preventing "casual" or opportunistic surveillance — they do not address cases where interpersonal control or access goes further. For instance, password sharing is common in romantic relationships [435]. In more opportunistic surveillance contexts, audit logs may be helpful to surface unexpected activity, but in more extreme intimate partner abuse situations, the situation is likely more complex. As other work studying intimate partner surveillance has discussed as well, novel and thoughtful approaches are required.

Mitigating risks of location tracking hardware. Our work surfaces examples of real users openly discussing (surprisingly openly, to us) the abuse of location tracking hardware like AirTags to non-consensually track peoples' location. Though Apple has implemented some protections, including playing audible alerts if an AirTag has followed you for too long, our data and other

anecdotes suggest that these mitigations are insufficient. As of early 2022, Apple is designing modifications to make AirTags louder and improve the alerting system for unrecognized AirTags [80]. Is it possible to develop technologies or policies that prevent the use case of tracking individuals at all?

Anticipating deeply personal motivations. We note that the motivations for the surveillance and control techniques we see in our data are deeply personal and emotional (and common): romantic partners worried about their partners cheating, parents worried about their childrens' safety, and children wishing to assert their independence. The underlying social phenomena motivating people to "hack" others are thus unlikely to go away. Developers of any apps or hardware used in these interpersonal contexts must consider how their product might be used or misused for these reasons. Our work complements other work which seeks to draw attention to these motivations and challenges[336, 573, 565].

Monitoring TikTok by researchers and developers. Given the popularity and openness with which we found anti-security advice on TikTok, continued monitoring of TikTok for these topics (including comments left on these videos, which we did not investigate) might be useful for those researching or providing support to victims of intimate partner surveillance, as well as to the companies whose technologies are being potentially misused or exploited. Future research could also evaluate the risks posed by the advised techniques.

Managing problematic viral content. Finally, we draw attention to the potential for TikTok to virally spread anti-privacy and anti-security advice to large audiences. Unlike in other contexts, like forums discussing how to do intimate partner surveillance [573], the nature of TikTok is such that its users may not be searching for specific content but rather receive content pushed to their feeds by TikTok's recommendation algorithm. And unlike ethical security vulnerability reports, these videos explicitly suggest exploiting vulnerabilities to violate the security and privacy of

others (especially in the intimate partner context).

Thus, we must consider TikTok's role in moderating, recommending, and perhaps limiting the spread of this type of content. TikTok's community guidelines already forbid videos from providing instructions on how to conduct illegal activity [568], which may apply to some of the videos in our dataset. Even for content that should not directly be prohibited, there may be a role for TikTok to display additional information (e.g., pointers to resources for all parties in interpersonal relationships), similar to misinformation-related notices on social media platforms. Whether and how such notices should be designed to be helpful is a question for future work.

7.8.2 TikTok as a Qualitative Data Source

Benefits. Our work demonstrates how TikTok can be used as an alternative source of qualitative, observational data for security and privacy-related topics, especially in contexts where traditional usable security methods such as interviews and surveys might be challenging to recruit for or conduct. For instance, recruiting and asking people to discuss the techniques they use to surveil or control intimate partners may not have surfaced as rich results due to social desirability bias. TikTok's user and creator base also has different demographics (e.g., skewing younger) than other social media platforms commonly studied in research (e.g., Twitter, Reddit) [97].

TikTok videos contain rich information in a short video: individual videos in our dataset often contained a multi-modal combination of video of the creator, speech, music, or other audio, text overlaid on the video, and screenshots or screen recordings. Additional context is provided through the video's caption, which often includes hashtags.

Challenges. A major challenge we faced was identifying relevant TikTok videos to study. The utility of text-based search is limited, and the emergence of different subcommunities on the platform (e.g., "toxics") meant that we had to discover specific terminology to find additional

relevant videos.

We also could not easily investigate TikTok's features for remixing and responding to content. Creators can "duet" videos by adding their own video to an existing one, or "stitch" videos by clipping and integrating clips into their own video. Unfortunately for our data collection, TikTok's platform does not offer a feature to find all duets and stitches.

Future work. This chapter has just scratched the surface of the types of security and privacy questions that we might investigate via TikTok content. For example, future work might investigate *pro*-security advice on TikTok. Anecdotally, we have also observed rich content on the topic of "sharenting". There may also be other sub-communities of interest, such as people conducting more technically sophisticated exploits.

Acknowledgements

We thank our reviewers for their helpful feedback. We are grateful for the many insights of Chris Geeng 🤓, Kentrell Owens 🤠, Tina Yeung 🍪, Sudheesh Singanamalla 🦁, and Os Keyes 🖤 during this research. We thank Kaiming Cheng 🚀 for assisting with the screenshots. This work was supported in part by the U.S. National Science Foundation under Awards CNS-1565252 and CNS-2114230, and by a gift from Google.

Chapter 8

"We're utterly ill-prepared to deal with something like this": Teachers' Perspectives on Student Generation of Synthetic Nonconsensual Explicit Imagery

Synthetic nonconsensual explicit imagery, also referred to as "deepfake nudes", is becoming faster and easier to generate. In 2023 and 2024, synthetic nonconsensual explicit imagery was reported in at least fourteen US middle and high schools, generated by students of other students. Teachers are at the front lines of this new form of image abuse and have a valuable perspective on threat models in this context. In this chapter, we interviewed 17 US teachers to understand their opinions and concerns about synthetic nonconsensual explicit imagery in schools. No teachers knew of it happening at their schools, but most expected it to be a growing issue. Teachers proposed many interventions, such as improving reporting mechanisms, focusing on consent in sex education, and updating technology policies. However, teachers disagreed about appropriate consequences for

students who create such images. We unpack our findings relative to differing models of justice, sexual violence, and sociopolitical challenges within schools.

This chapter originally appeared as the paper "'We're utterly ill-prepared to deal with something like this': Teachers' Perspectives on Student Generation of Synthetic Nonconsensual Explicit Imagery" at the CHI Conference on Human Factors in Computing Systems in 2025 [599]. 'We' in this chapter refers to me and the co-authors: Christina Yeung, Franziska Roesner, and Tadayoshi Kohno.

Warning: This chapter includes text description and quotes about image-based sexual abuse and child abuse.

8.1 Introduction

Synthetic nonconsensual explicit imagery (SNCEI) is easier and faster to generate than ever before. Creation of realistic pictures or video is possible with a single image and knowledge of a website or app that provides cheap or free "nudify-as-a-service" operations, using AI to digitally "remove" clothing or to swap faces onto nude bodies. Prior research has investigated the specialized communities that discuss advanced machine learning techniques to develop the underlying AI models capable of generating explicit content [569, 604], but users no longer need to engage with those communities to create SNCEI. Lowering the technical barriers means that today, creating SNCEI has grown beyond highly technical communities, and instead is now readily available to a much broader online audience.

In late 2023 and early 2024, journalists began reporting that students in middle and high schools used "nudify" services to generate images. In one of the first criminal cases of its kind in the US, two boys in Florida were charged for creating AI-generated nude images of their

classmates [251]. Students have allegedly used similar platforms to create SNCEI of their classmates in at least 9 other US middle and high schools, including in California, Ohio, Alabama, Florida, and Washington [251, 536, 102, 297, 250, 226, 143, 530, 347, 256], as well as in schools in Spain and South Korea [571, 543]. The creators of the synthetic images were one or more boys who made images of their classmates who were girls without their knowledge or consent, where some of the victim-survivors depicted were as young as 12 years old. Demonstrating how simple it has become to find these "undress" services, some creators had discovered the tools on social media, including TikTok and Instagram [102].

There has been a notable lack of consensus on how schools should respond to students creating SNCEI. Though many of the 10 US schools suspended or expelled the students who created the images, in at least one case, school officials delayed reporting to law enforcement, citing confusion about their role as mandatory reporters and whether synthetic imagery fell into the same domain [102]. In another, a school superintendent said that their "hands were tied" in terms of the actions the school could take as it did not happen directly on school grounds, and that it therefore should have been considered a case between the relevant guardians and law enforcement [530] However, involving law enforcement in and of itself is an unclear solution. Regulation of SNCEI is an evolving area: though there are some proposed bills that have been brought in 2024, there are currently no federal laws in the US that directly address scenarios where youth create SNCEI.

Synthetic images are only going to become more common. Taking a proactive approach, we interviewed US middle and high school teachers — people with high levels of interaction with and deep knowledge of students — to understand teachers' knowledge and perspectives about student-generated synthetic nonconsensual synthetic explicit imagery (SNCEI). Specifically, we asked:

RQ1: Threat models. What motivations do teachers expect their students to have for creating SNCEI? Who do they think may become a perpetrator or victim-survivor? How easy do they think it is to learn about and use these technologies?

RQ2: Interventions. What kinds of interventions do teachers anticipate would be effective? How do schools currently handle incidents related to student safety? What resources would teachers want regarding SNCEI in the future?

RQ3: Broader sociopolitical context. How are teachers' opinions about SNCEI informed by the broader sociopolitical context in the US, including about justice, gender, sexual health, machine learning, and more?

We found that teachers broadly have heard about SNCEI, although no teacher mentioned it happening at their own school. Most teachers remarked that it could be possible, and one even suspected it was already happening, but they just did not know about it. Notably, teachers' understanding of how SNCEI would appear in schools revealed awareness of dynamics of gender-based violence, particularly that girls would be the most affected, although boys and girls could have motivations for creating. Teachers had significant concerns about the potential for SNCEI to worsen existing cyberbullying or sexual harassment, intuiting that new technologies would exacerbate existing avenues of interpersonal harm.

Most teachers emphasized, however, that kids may not understand the consequences of their actions and are still learning how to build healthy relationships. Teachers saw multiple opportunities for improved education as promising avenues for addressing the impending issues of SNCEI, including around sexual health, use of social media and technology, and emotional and social development. We discuss these opportunities, and others, for curtailing the harms of synthetic content.

8.2 Related Work

Our investigation of synthetic nonconsensual explicit imagery generation by students is situated within the broader landscape of image-based sexual abuse, as a form of technology-facilitated gender-based violence [261, 384]. We describe how our work builds on prior work within the computer security, privacy, and online safety literature, including on youth digital safety.

8.2.1 Image-Based Sexual Abuse

In 2017, McGlynn and Rackley conceptualized image-based sexual abuse (IBSA) as an umbrella term for a range of harms relating to the nonconsensual creation or distribution of private sexual images. They proposed image-based sexual abuse as a term that better situates harms like nonconsensual sharing of intimate images ("revenge porn") or upskirting as forms of sexual violence and thus part of a broader approach to respond to sexual violence [382, 358]. In subsequent years, research continued to explore more manifestations of IBSA, e.g., threats to distribute and financial sextortion [260, 423, 264, 138], as well as measure rates of victimization and perpetration [175, 495, 264].

While people who experience IBSA do not all experience the same consequences, these consequences frequently include serious emotional, social, financial, and physical impacts [260, 37, 381]. A study investigating the experiences of 75 victim-survivors of IBSA in the UK, Australia and New Zealand developed five phenomenological themes in victim-survivors' accounts of harm: immense *social rupture* that altered their sense of self and relationships with others, perceived *constancy* and *existential threat* of the harm, as well as consequential *isolation* and *constrained liberty* that radically changed their experience of the world [381]. In intimate relationships, abusers may use IBSA specifically as a tactic for gaining power and control, e.g., as part of emotional abuse or using coercion and threats [176].

Recourse for survivors of IBSA varies highly by location, identity, and other factors. Legal scholars have called for nonconsensual sharing of intimate images ("revenge porn") to be criminalized, given the grave harms that chill self-expression and devastating privacy invasions [114]. In the decade since, legal advocacy in the US have contributed to 49 states, DC, and two territories passing laws against nonconsensual distribution of intimate images [94], though these are only one of many forms of IBSA. Further, social stigma may lead people who experience IBSA to only seek informal help [596], or inhibit them from seeking help at all. Computing researchers are also exploring technical recourse for combating IBSA, e.g., proactive protection strategies [463], conceptual frameworks to identify intervention opportunities [465], ML-assisted detection [469]. We situate our work among this growing body of work on IBSA by affirming the broader understanding that abuse of intimate and sexual images is an extremely urgent and grave issue with significant variance by sociotechnical, political, and cultural dimensions depending on the specifics of each case.

Synthetic nonconsensual explicit imagery (SNCEI). Synthetic nonconsensual explicit imagery (SNCEI) is a specific form of image-based sexual abuse where the images used for abuse are synthetically created, whether through photoshopping [384] or generative AI tools [569, 578, 197]. Though image manipulation tools have existed for decades [184], using technology to nonconsensually create sexually explicit images drew public attention in 2017 when a Reddit user named "deepfakes" posted SNCEI (videos) of celebrities. In subsequent months, tens of thousands of users joined `r/deepfake`, a subcommunity dedicated to creating and sharing similar content. Reddit and other major social platforms have largely banned SNCEI [268], though it is still being produced on specialized forums [569, 604]. However, industry reports in the last two years increasingly highlight that AI-generated images are no longer restricted to niche underground forums, and are also widely available as part of a monetized online business model [327, 217].

A 2019 report found that essentially all SNCEI found online was sexually explicit and depicted ciswomen; similarly, a 2023 report confirmed these trends and that the number of deepfake videos online had increased 550% since 2019 [269]. People in the US have a strong opposition to the creation of SNCEI [320], though were less opposed to the seeking out or sharing of such content [76], aligning with research on perspectives in 10 countries that showed viewing of SNCEI of celebrities was more common than of non-celebrities [578]. Scholars have drawn attention to the uniquely harmful exploitations of deepfake technology, particularly for nonconsensual creation of sexual content of women, and called for additional regulation [112, 325, 320].

In this chapter, we focus specifically on SNCEI in school contexts, motivated by news reporting in late 2023 and early 2024 that identified SNCEI in schools as a newly prevalent concern. Most prior research focuses on SNCEI *by and of adults* in intimate partner relationships or online [569, 325, 596, 176, 381], but youth creation of SNCEI poses unique legal and social considerations. Laws are actively evolving about treating SNCEI that depicts people under 18 as child abuse or child sexual abuse material (CSAM) [415], though the FBI has issued an alert that it is [414]. The social context of a school is unlike most adult environments; students are required to attend and teachers have classroom authority. Further, teachers often have specialized training and are highly committed to student well-being, creating a unique opportunity to explore interventions that would be impossible for SNCEI in non-school contexts.

Terminology. Carceral responses to sexual violence tends to use "perpetrator" or "offender" to describe a person enacting violence, and "victim", "survivor", or "victim-survivor" to describe a person affected by the violence [586]. However, some advocacy organizations highlight that these terms reduce people's personhood to an identity related to one event, de-emphasizing their agency [124] and obscuring that someone who perpetrated violence often experienced violence themselves (e.g., substantial IBSA "victim-perpetrator" overlap [548]). These terms also carry

stigma and make people resistant to taking accountability [124, 585]. Given that perpetrator and victim-survivor are terms that are most commonly used in the security and privacy research community, we alternate between both sets of terms, i.e., perpetrator or creator of SNCEI; victim-survivor or subject of SNCEI.

8.2.2 Youth and Online Safety

Researchers in security, privacy, safety, and HCI have studied the digital safety of youth through multiple experiences of risk, including IBSA but also online (non-sexual) harassment, information breaches, financial fraud, misinformation [612, 201]. Researchers have looked proactively to understand the strategies taken by youth or adults who support youth [201], as well as the how youth respond to sexual risk experiences in private Instagram conversations [162, 472, 14, 17]. Experience of risk is not uniform: LGBTQ+ youth experience more high-risk online interactions than heterosexual youth [562]. In order to navigate such risks, youth often turn to their peers to learn more about safe sexting or other online safety risks [246, 281]. However, participatory approaches to online safety by working with youth, such as collaborative family-cenetered design for online safety [12] or co-management of online apps between parents and teenagers [12] emphasize that youth safety does not only fall to youth, but is instead a communal endeavor [93, 461].

Researchers in human-computer interaction have also studied the broader field of cyberbullying, including large-scale literature reviews to document its scale and history [282], youth peer help-seeking strategies [281], and bystander mitigation strategies on social media platforms [165]. Our work differs from prior research on cyberbullying as we focus on the potential impact of SNCEI, which involves gender or sexual abuse not found in all conversations about cyberbullying.

While SNCEI is not new, the use of modern generative AI tools to create it only become widely

accessible in the past few years. How youth might learn about new online safety risks is an important area of research, particularly regarding IBSA risks. In this work, we explore not only how youth may experience these risks, but also how youth are the perpetrators of this risk to other youth [449].

Policy and regulation about SNCEI and CSAM. Based on a report from August 2024, there are federal bills being reviewed to create civil penalties for creating synthetic images of someone against their will [403]. In particular, one bipartisan-supported proposed legislation called the "Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks (TAKE IT DOWN) Act" seeks to criminalize the intentional creation of SNCEI, and also mandates that social media platforms must remove content within 48 hours after being reported.

Additionally, 29 states have passed bills that explicitly discuss the creation of SNCEI [403]. Of these current bills, 18 address SNCEI as child sexual abuse material (CSAM), i.e., sexual content created about youth (individuals below the age of 18), by expanding "child pornography" laws to include digitally generated content. Other states take a different approach by amending "revenge pornography" laws to include SNCEI. Federal agencies, such as the FBI have also released statements clarifying that all CSAM, regardless of whether or not it was synthetically created, would be illegal to hold, create or distribute [414]. However, policies are still developing, and it is not yet clear how laws surrounding SNCEI and CSAM would be applied or enforced in school contexts [448].

8.3 Methods

We interviewed 17 middle and high school teachers in the US to understand their opinions and concerns about synthetic nonconsensual explicit imagery (SNCEI). Given that SNCEI is an emerging issue, we took an *proactive* approach in this study; we wanted to understand how

teachers expect that incidents will arise, as well as how they predict students and schools will respond. SNCEI incidents are possible in any school, so many teachers who encounter SNCEI may be encountering this issue for the first time.

8.3.1 Recruiting and Participants

We mainly recruited on Prolific, a crowdworking platform shown to be preferable to Amazon Mechanical Turk in terms of comprehension, attention, and honesty of participants [444]. We also shared the opportunity to participate with eligible individuals in our personal networks and on social media (see recruitment flyer in Appendix Figure C.1). All interested in participating were directed to a 3-minute screener survey to confirm their eligibility: located in the US, currently a teacher at a US middle or high school, and have at least two years of teaching experience. The screener survey also collected basic demographic information about the participant (age, gender, state), which were not used for recruitment criteria but were retained to report the diversity of our sample. Of the 375 survey responses we received, we incrementally invited eligible participants to schedule a 60-minute online interview until we anticipated that our data collection and analysis would reach conceptual depth, i.e., demonstrate a *wide*, *complex*, and *valid* range of concepts, with appropriate *subtlety* to meaning and *resonance* with the existing literature [409, 72]. Interviews were conducted in July and August 2024 and participants received $40. Details about our participants are shown in Table 8.1.

Of the 17 teachers interviewed, 5 were recruited through our personal networks and 12 through Prolific, representing 10 distinct states: FL (3), TX (3), GA (2), NY (2), WA (2), AZ, MA, MS, NH, TN. Participants' age ranged from 22 to 56, with an average age of 38; 12 were women, 4 were men, and 1 was a non-binary transmasculine[1] person. Most of the participants had between 11-20 years

[1]Transmasculine, sometimes abbreviated to transmasc, describes a transgender person who was assigned female at birth and whose gender is in some way aligned with masculinity.

of experience working with youth (7 participants) and the 17 participants covered 11 different teaching subjects, including English (4), Substitutes covering all subjects (3), Math (2), Chemistry, Freshman Seminar, German, History, Music, Science and Biology, Special education, and Speech Pathology. 12 participants provided the approximate size of the student body served by their school, ranging from an alternative school that had between 5 to 100 students at a time, up to 2,500 students. Participants described their respective student body as predominantly Black or Latino (6), from lower income families (5), socioeconomically diverse or a mix of backgrounds (4), or from wealthy backgrounds (2).

8.3.2 Interview Procedures

We developed our semi-structured interview script based on our research questions and refined the script through four pilot interviews. After reviewing our consent form, we began the interview by asking about teachers' role in their school and general characteristics about the school, and then about their awareness of SNCEI, in schools or more broadly. Next, we asked about teachers' opinions about which types of students might create SNCEI, their potential motivations, as well as who might be likely to be victim-survivors. After we had discussed various elements of their imagined threat models, we then shared a summary of the cases reported in local or national news about SNCEI in US schools. These selected cases reflected the focus of our study being on students creating SNCEI of other students. We solicited their impressions about these cases and whether something similar would be possible in their own schools.

In the second half of the interview, we explored potential interventions – what kinds of support they would like to see at an individual, school or administrative, and policy level. Finally, we concluded with some high-level questions about their fears and hopes for the future. We also shared our knowledge about terminology and provided an opportunity for participants to ask us

questions. The full interview protocol is in Appendix C.1.2.

8.3.3 Analysis Approach

We transcribed all interviews for qualitative analysis, which combined descriptive and interpretative approaches to thematic analysis (TA) [73]. Some of our research questions involved specific questions about teachers' imagined threat model of SNCEI generation, including potential motivations, perpetrators, and victim-survivors. For these, our analysis was more akin to codebook TA, as the interview script was more structured and participant responses were more predictable and descriptive. However, other parts of our interviewing and analysis relied more heavily on researcher interpretation, such as how teachers evaluated different interventions, which sociopolitical factors influenced their perspectives, and broader views of justice, education, and technology. For these, our analysis was more akin to a reflexive TA [73, 69, 74], where analysis occurred recursively through sustained engagement with the dataset. We use the six phases of reflexive thematic analysis [73, 69, 74] to describe the steps of our coding below, given that we conducted both types of analysis concurrently, i.e., codebook TA and reflexive TA.

To code the interviews, the lead researcher first reviewed all transcripts, including rewatching video recordings to minimize the risk of misinterpretation based on transcribed text (TA phase 1). Then, this researcher coded five interviews to develop a preliminary codebook as well as preliminary themes, which were discussed with the other members of the research team (TA phases 2-3). After initial discussion, the lead researcher continued to code all remaining interviews, iteratively updating the codeboook as necessary (TA phases 2-3). After coding all interviews, the other members of the research team reviewed the codebook again, suggesting new codes or merging multiple codes as necessary, and refining the codes and themes (TA phases 4-5). The final codebook had 119 codes, organized into 13 groups. To ensure analysis quality, a second coder

independently coded 4 interviews with the original codebook and added two codes, which were then propagated to the remainder of the interviews. The second coder then reviewed all codes made by the lead researcher and made changes as necessary.

8.3.4 Ethics

Our study was reviewed and approved by our institution's IRB. Given the sensitivity of this topic, at the beginning of this project, we also consulted experts at our institution on child abuse and child sexual abuse material (CSAM), human subjects research, interpersonal abuse, and counseling in education contexts. Throughout the research project, we carefully addressed ethical considerations related to mandatory reporting, participant anonymity, and participant and researcher well-being.

As employees of a public institution with mandatory reporting requirements for child abuse, we were especially cautious about how this could impact participant anonymity. Whether SNCEI is considered child abuse is an evolving legal area, and we recognized that making a report could introduce additional legal and personal risks to participants as well as their students. We disclosed our status as mandatory reporters to all participants, including the types of information that we would have to report if they disclosed abuse; we did not have to make any reports during the course of this research. Other measures we took to preserve participant anonymity included obtaining a waiver of documenting consent (we still obtained consent, but did not document it in a form that might otherwise de-anonymize participants) from our IRB, as well as a Certificate of Confidentiality[2] from the NIH.

All participants were informed about the nature of the study through a consent form (in text and verbally) before starting the interview, including the topics that would be discussed, and could

[2]NIH Certificates of Confidentiality (CoC) protect the privacy of participants enrolled in health-related research that use sensitive information, including in response to legal demands. See: https://www.era.nih.gov/erahelp/Coc_Ext/Content/A-Introduction/Introduction.htm

skip any question(s). We took additional measures to support the well-being of researchers in this research, including: weekly individual and group check-ins, meeting with trauma-informed experts, having access to therapists, and taking regular breaks. Before beginning the study, we drafted a document for researcher safety guidelines, which we periodically reviewed to ensure that we stayed attuned to the well-being of the researchers.

Positionality statement. The way we discuss and perceive technology, justice, and education in this work is informed by our particular social, cultural, political, and historical context. We are researchers who have predominantly lived and worked in the US, have English as a first language, and have attended US middle or high schools, though these periods were many years or decades ago. Though all co-authors have held or currently hold a role with some teaching responsibilities, most of these experiences are at the post-secondary level. Our motivation in conducting this research is to explore the landscape and different options for mitigating the harms of SNCEI, and particularly non-punitive approaches.

8.3.5 Limitations

As with all interview studies, our research was limited by who was motivated to participate in our study, as well as what participants were comfortable disclosing to us. In particular, we did not collect the names of participants' schools, and made it clear that we were only interested in their perspectives as a teacher, not on behalf of their school or school district. However, our participants may have been more tech-savvy than all teachers, as our participants voluntarily participated in a study about SNCEI (see definition of SNCEI in recruitment flyer in Appendix Figure C.1). We made clear at multiple points that there were no right or wrong answers to any questions, and were only interested in their opinions and perceptions. Additionally, who chose to participate may have been influenced by our recruitment material and process, which disclosed

our institution and the feminine names of the two interviewers.

Our recruitment criteria did not include having experience with handling SNCEI incidents given ethical considerations and potential risks to participants (see Section 8.3.4). Thus, this study is not suited to describe prevalence of SNCEI in schools nor the opinions of teachers who have dealt with SNCEI incidents. However, SNCEI is an emerging issue that many teachers will likely have to face for the first time, which is why we proactively study threat models and potential interventions. Studying abuse proactively, i.e., before it occurs, contributes a distinct yet valuable perspective compared to studies of abuse after it occurs.

Our use of the term "deepfake nudes" may have influenced the connotations that some participants brought to our interviews. However, these are the most commonly used terms in media, so we used them because we were interested in how participants would most likely discuss them in their schools or with their colleagues and students. When participants used different terms, we followed suit in the interviews to use whichever terms came most naturally to them. Though in theory "deepfake nudes" could be created consensually, our study was focused on nonconsensual cases, aligned with the terminology that public media tends to use when reporting about SNCEI. We also included a debrief at the end to explain why academics argue against using "pornography" to refer to nonconsensual content [382, 358], and informed participants that alternative words like "AI-generated" or "synthetic" are also common.

8.4 Results

8.4.1 Technology and Support Resources at School

To provide some context about the school environment, we briefly describe technology use and support resources mentioned by participants about their schools. Out of 17, nine teachers described

Table 8.1: Demographics and experiences of the 17 teachers interviewed for this study. Columns marked with * show abbreviated responses from participants, in their own words. N/A indicates that teachers did not give that information during the interview.

P#	Gender	Age	Years of Experience	How often heard of "deepfake nudes" or "AI porn"	Current Role Working with Youth*	# of Students in School	Student Characteristics and/or Background*
P1	Non-binary transmasc	22	4-5	Once or twice in the last 6 months	Math (9th - 12th grade)	N/A	Title 1 school, predominantly Black students
P2	Woman	30	11-20	Many times in the last 6 months	Freshman Seminar, high school basketball, volleyball and track coach	1,500	A lot of BIPOC students, many lower income families
P3	Woman	35	4-5	Once or twice in the last 6 months	English/Reading (9th & 11th grade)	N/A	Mix of Spanish speaking students, Latino, Black, and White students
P4	Woman	29	6-10	Not at all	Chemistry (9th - 12th grade), science department chair, environmental and crochet club sponsor	N/A	Predominantly Black school (~75%), White, Hispanic, Pacific Islander, Asian
P5	Woman	40	11-20	Once or twice in the last 6 months	High school substitute teacher	N/A	Predominantly Latino school, lower to middle class (~50%), increasing number of students who are learning English as a second language
P6	Woman	32	6-10	Not at all	Speech language pathologist with students aged 5 - 21	50	Legally blind, multiply disabled, different backgrounds
P7	Woman	56	6-10	Once or twice in the last 6 months	Substitute teacher (6th - 12th grade), all subjects	N/A	Students from wealthy backgrounds
P8	Woman	44	11-20	Once or twice in the last 6 months	Substitute teacher (6th - 8th grade), cooking club, STEM activities, walking club	300	Lower income, busy parents, some parents may be divorced
P9	Woman	33	2-3	Once or twice in the last 6 months	Math (11th - 12th grade), gaming club, crochet club	130	Title 1 school, lower income, 40% of students are bi or multiracial, significant number of Black, Latino, Pacific Islander students, large number of students (~50%) on queer or neurodivergent spectrum
P10	Man	42	21+	Once or twice in the last 6 months	Music theory and piano (9th - 12th grade)	2,500	Majority Black students, Hispanic students, minority White students
P11	Woman	42	11-20	Once or twice in the last 6 months	English Language and Arts (5th - 8th grade)	200	Title 1 school, lower income, 85% Black, also have Asian students, minority White students
P12	Man	33	6-10	Once or twice in the last 6 months	IB History, honors apUS history (11th - 12th grade), social studies honor society sponsor, founding member of social/emotional program for our IB students	2,000	Predominantly Black students (~65-75%), Hispanic students, White, minority Asian
P13	Woman	46	21+	Once or twice in the last 6 months	Science and biology (7th - 12th grade)	Varies, from ~5 or 6, to a max of 100	80% children of color, not much support at home
P14	Man	39	11-20	Many times in the last 6 months	English (8th grade), LEGO Club, National Junior Honor Society chapter, and the Gender-Sexuality Alliance (NJHS).	800 - 850	60% White, 40% not White, fair mix of immigrant and non-immigrant students, many English language learner students come from Dominican Republic, Puerto Rico, wide range of socioeconomic backgrounds
P15	Woman	38	6-10	Once or twice in the last 6 months	English (11th - 12th grade)	2,500	Socioeconomically diverse; 50% at risk, some college-bound
P16	Woman	44	11-20	Once or twice in the last 6 months	Special education (6th - 12th grade)	1,000	Lower income, students with special needs, some have ADHD, some students come from multiple households
P17	Man	40	11-20	Once or twice in the last 6 months	German (6th - 7th grade)	1,400	Higher than average socioeconomic status, about 90% Caucasian, as well as some African American, Asian, and Hispanic students

that their schools provided computing devices to each student (e.g., Chromebook or iPad). While teachers generally believed that school-provided technology was necessary for student learning, they were supportive of limitations that would reduce classroom distractions or opportunities for harm. Nine mentioned that their school provided WiFi with limitations, including blocking specific websites (mentioned 6 times), flagging content by keywords (3 mentions), or monitoring network traffic (3 mentions). Further, some teachers reported that there were existing school resources or policies that they would refer to if SNCEI incidents arose, but these varied greatly by school. For instance, some mentioned consulting school guidance counselors, code of conducts, policies that outlined the acceptable image sharing practices, or student digital safety courses; others mentioned they did not have these resources. The differences between each participants' school highlight one set of challenges in responding to SNCEI incidents: that each school will have different capacities and constraints to navigate, which may limit the effectiveness of generalized recommendations or interventions.

8.4.2 Perceptions of SNCEI

In this section, we describe what participants heard about SNCEI, and their opinions about its presence in schools.

Heard about SNCEI from news or social media. All of our participants had heard about SNCEI in some capacity. Multiple participants heard about SNCEI in the news, particularly as used against celebrities (e.g., Taylor Swift, Alexandria Ocasio-Cortez), or raising awareness about specific harm scenarios (e.g., financial sextortion, child sexual abuse material). Participants largely discussed students creating SNCEI of other students, though a few mentioned that students could also create SNCEI of teachers by mentioning a report of such in a recent New York Times article [535]. While no participants reported knowledge of actual cases at their own schools, and most had not heard

about news stories about it happening in schools, some participants reported it coming up in school trainings. One participant also mentioned hearing about it in a social media group where a mom was warning others because girls at her daughter's high school had created SNCEI of other girls. Some participants had familiarity with the broader use of generative AI, such as for political misinformation or academic dishonesty.

Participants did not mention hearing about SNCEI from family, friends, or students, though one had a conversation with a student about political deepfakes. A few had also discussed SNCEI with colleagues, and P13 shared the outcome of one such discussion: "our concern is that they could happen eventually. I think it's something that we'll have to face... this could happen here and it was a consensus, I guess between all of us, that yeah, that could."

Concerning, unsurprising, and likely already prevalent in schools. During our interviews, we shared a summary of news reports we had found about nine[3] cases of SNCEI in schools across the US, which was met by some teachers with notable concern, especially at the youth of the students involved. Remarking on age, P15 shared: "It's making me really sad. It's upsetting... I think that it's really hard to be a kid." Teachers immediately grasped the consequences, and P6 was very empathetic to victim-survivors in particular, reflecting: "I can't imagine being in a classroom and being that person that a nude was created of and feeling like everybody has seen you naked even though it's not what you might look like naked. I just feel for those people, it must be so awful." Other teachers were also concerned about what this would mean for people who had created SNCEI, and whether this would set them on a concerning path as an adult.

However, many teachers' reactions also conveyed a resigned lack of surprise and that the summary of cases matched their expectations almost exactly. Teachers particularly mentioned that the genders of the perpetrators (boys) and victim-survivors (girls), as well as the ages (middle

[3]At the time of the interviews, we had collected nine cases, but by the time of writing the paper, one additional case had been reported.

and high school aged), aligned with their experiences. P3 was not surprised that some students who created were in middle school: "middle schoolers today are scary, very scary... a lot of them have no feelings. They just do whatever, they don't think about anything outside of the action they want to commit in that moment."

When asked if similar situations could happen at their schools, most teachers responded yes. P15 shared that this was because it aligned with existing dynamics of harm: "I've seen bullying and I've seen cyberbullying. I've seen slut shaming, just in different forms, and now that the tools are available, I just don't see any reason why students wouldn't use those tools." Given that teachers were not aware of actual cases at their schools, multiple teachers hypothesized that it was actually happening but administrators or teachers were simply unaware, i.e., that "if the students are smart enough to create those kind of images then they're smart enough to hide it" (P17). Other teachers thought it was possible, but would be (or they hoped it would be) restricted to only a few students.

8.4.3 Teachers' Perceptions of SNCEI Threat Models

In this section, we describe teachers' perceptions about the process of creating or viewing SNCEI. We also summarize the possible scenarios in which teachers thought SNCEI could occur, exploring potential (and possibly overlapping) motivations, as well as possible or likely perpetrators and victim-survivors.

Perpetrator capabilities: Creation and viewing are trivial tasks. Teachers largely assumed that creating and viewing SNCEI would be simple, easy, and quick. When asked what someone would need (aside from a phone, laptop, or tablet and an internet connection) to create SNCEI, participants guessed that it would be only take one or a few photos and some type of specialized software, such as a "more upscale version of Photoshop" (P4). Reflecting on how widely accessible

other generative AI or face filtering tools were, many teachers thought that finding SNCEI tools would be relatively easy in a browser or smartphone app, though it might take more time or technical skill to make it look realistic. Most estimates of how long it would take to create ranged from a few seconds to minutes, with the longest estimate being two days. A few teachers mentioned that how-to guides would probably be readily accessible on major social media platforms.

In terms of viewing, teachers almost unanimously described that if students were to view SNCEI, it most likely be because it was shared with them, though some also guessed it would be easy to find through an online search or social media. Teachers described technosocial cultures of sharing that would escalate the spread of images, as multiple teachers predicted that the first thing that a student would do upon seeing SNCEI would be to share it with others. P13 recounted students' approach to media:

> "Share it with the whole world. Whoever they can get to look at it. These kids, in general when it comes to videos of any capacity, it's the number of likes, the number of followers... they thrive off of that." (P13)

Teachers attributed this to "students [being] very attuned to what other students are thinking or doing" (P7). And despite attempts to restrict the sharing of SNCEI, sharing could quickly get out of hand: "you assume your friend will keep a secret. But then that friend has a friend and they assume that they'll keep a secret. And that's how it eventually gets out" (P10). P9 remarked that one student might be in up to 70 group chats at once, so keeping secrets could be extremely difficult.

Only a few teachers discussed consequences for students who decided to share or distribute images that students did not originally create. P16 thought that students caught sharing SNCEI images ought to be educated on the impact to victim-survivors to evoke empathy, and prevent sharing in the future. One teacher (P11) believed that consequences for sharing SNCEI images

might be more severe than if a student was caught sharing non-digitally altered inappropriate images, as they thought that it implied a level of intention that was otherwise not present. Finally, a couple teachers believed that legal punishment through law enforcement would punish students for sharing SNCEI images (P1, P3).

Motivations: SNCEI for cyberbullying. Though SNCEI is by definition explicit, teachers often distinguished between motivations of a creator that were *not* sexualized (like cyberbullying described here) and those that were (like the sexual and gender abuse described below). Most participants described cyberbullying motivations as a prominent scenario for SNCEI, e.g., "trying to troll another student" (P8), "when they get angry... bringing up a lot of the tea on each other" (P4), or otherwise expressing anger by trying to hurt another student or damage their reputation. SNCEI would likely be perceived as a more serious form of bullying, and could be reserved for situations where someone was already being bullied but if "it wasn't getting to them the way they wanted to, creating a deepfake nude... [would go] that extra mile in order to hurt someone's reputation" (P13). In a number of situations, teachers described bullies feeling some kind of hurt themselves, which they then projected onto others: "they want somebody else to feel the way that they felt... the person that they feel deserves it" (P16).

While teachers generally described that students would bully others by generating SNCEI of the target of the bullying, one teacher also mentioned that cyberbullying could also be carried out by *falsely* accusing someone of generating SNCEI.

Gendered differences in SNCEI for cyberbullying: Some teachers noted that anyone could be mean or want to cause harm, while others believed that boys were much more likely to be the creators of SNCEI. P17 attributed this to seeing SNCEI as "an act of aggression, and for the most part, cisheterosexual men tend to be the most aggressive, when it comes to... getting their way."

Some teachers also mentioned students, likely boys, creating SNCEI to appear cool, funny, or

otherwise try to gain social status at the expense of others. Students could use a synthetic image to falsely claim they had a sexual encounter with someone else or to show off the fact that they could create SNCEI. P12 noted that the intent would be to "make laughs, just to make themselves the star of the show for a day," although this would probably be ineffective because "when those videos come out, nobody's talking about who created it. They're talking about who's in it." These motivations were not mentioned with respect to girls creating images, who teachers described as were more likely to create for cyberbullying in case of a friendship falling out.

While some teachers acknowledged that boys *could* be the subject of SNCEI, they generally believed that girls or other marginalized students would be the subject. No participants mentioned non-binary students in their threat model, either for creating or having images created of them. One teacher remarked that girls "have the most body parts to cover" (P3), aligning with another teacher who believed subjects might include "hijabi students, because they already have that mystery to the kids of what's underneath there, they might do something like that to stir up controversy" (P4).

Motivations: SNCEI for gender and sexual abuse. Teachers also described a range of gender and sexual abuse scenarios where the harms were rooted in systems of unequal power between (cis)women and (cis)men, including nonconsensual sexual behavior.[4] A particularly volatile and likely time for creating SNCEI was the moment of a relationship breakup. After a breakup, students could be so angry that "it becomes kind of a free-for-all" and prior respect or trust was abandoned (P13). These motivations were also described as "revenge" due to perceived rejections: because of a "spurned love" (P15) or because "[someone] went out with her and she wouldn't have sex" (P7). Teachers remarked that "revenge porn"[5] was a known issue among adults, so it made sense that

[4]No participants mentioned transgender people as creators or subjects of SNCEI in our study, so our results focus on cisheteronormative relations, though we note gender and sexual abuse also cause harm in contexts that are not cisheteronormative.

[5]We note that "revenge" is a false justification for abusive behavior and that "porn" should describe consensual

students would also have similar motivations.

Some scenarios of gender and sexual abuse that teachers speculated about did not involve a creator of SNCEI actively seeking to do harm. Students might also be motivated to create it because of a crush or fantasy, and might think that SNCEI for such purposes is "less exploiting because it's not really [that person]... Deepfakes for them might not see as crossing the line of consent" (P4). However, teachers repeatedly mentioned that despite students' intentions to keep things secret, youth sharing practices (described above) would make doing so nearly impossible.

Two teachers also mentioned SNCEI in the context of two other types of image-based sexual abuse. P13 imagined SNCEI could be generated to blackmail, i.e., non-financial sextortion, and to manipulate a partner: "I made this of you, and look at it. And now you're gonna need to do what I want you to do – be my girlfriend, be my boyfriend, whatever, or I will send it out to the world." P2 imagined that SNCEI might be created by a friend intending to help someone in a pressurized sexting situation who would not feel comfortable sending a genuine explicit image.

Gendered differences in SNCEI for gender and sexual abuse: Teachers again largely agreed that boys would be more likely to be creators of SNCEI, and girls to be the subject of it. Boys who spent too much time in "weird" (P1) or "unsavory" (P14) online spaces were considered to have patterns of problematic behaviors and be particularly prone to create SNCEI; this imagined type of student was also described as "sad" (P5), a "loner" (P8), "broken" (P10), "socially isolated" (P15), and "insecure" (P16). Teachers named specific misogynistic people or online spaces that would also motivate boys to engage in slut-shaming, particularly targeted at sexually active, popular, or attractive girls. The misogynistic motivations to target girls led teachers to nearly unanimously concur that girls would be the victim-survivors of SNCEI. P14 noted that "my girls are pretty socially aware, and are by and large feminists", which had created a culture among girls to be more protective of each other, though P9 noted that they could imagine girls creating SNCEI

sexual imagery [382, 358].

because "there's still intense social pressures at that age and I don't think they've internalized enough yet how harmful it is to not support other women."

Multiple teachers remarked on societal expectations that boys think often of sex and are rewarded for being sexual or "more pressing" (P10), and even if "boys might be secretly upset about [being the subject of SNCEI]... it could be spun [as a positive for them]" (P11). On the other hand, teachers noted girls faced stigma for even being associated with sex. P8 thought girls would be extremely unlikely to create SNCEI because "they would feel probably grossed out by it and repulsed and just not want anything to do with it."

Teachers also pointed out that boys were more likely to work together for cyberbullying, and individually for gender and sexual abuse motivations, since students might recognize that creating SNCEI for sexual gratification would be stigmatized.

Motivations: SNCEI for curiosity. Much less common than for cyberbullying or gender and sexual abuse, teachers suspected students might be motivated to create SNCEI out of curiosity about technology or to alleviate boredom. Some students might have "so much free time" (P7), or feel bored because classes were either not challenging enough or too challenging.

Finally, two teachers mentioned that students could be motivated to create SNCEI for profit. P1 based this guess on existing markets: "if there's a market for feet pics, I'm sure there's a market for deepfake porn, maybe even deepfake feet pic porn."

8.4.4 Potential Interventions

In this section, we describe the range of potential interventions that teachers suggested, in four broad and overlapping categories: reporting and school policies, proactive measures, incident response measures, and interventions beyond schools.

Reporting and school policies. Teachers described that one of the first actions they would take,

if they heard about students creating SNCEI, would be to report it to their administrators, law enforcement, and/or social services agencies. The distinctions between these three groups — the first being employed by a school, the second and third being employed by the state — were not always clear, but teachers generally erred on the side of reporting to as many relevant parties as they could think of. Depending on the structure of their school, teachers listed a range of supervisors or staff within their school that they would inform, including principals, assistant principals, guidance counselors, school psychologists, and a diocese contact (in the case of a Catholic school). Many teachers described reporting to law enforcement and social services[6] as part of their duties as a mandatory reporter. If a teacher's school had an assigned sworn law enforcement officer, i.e., a school resource officer[7] (SRO), teachers described reporting to them, while other teachers merely mentioned "getting police involved." Teachers took reporting responsibilities seriously, seeing themselves as first responders: "in my capacity as a teacher, it's not my job to investigate. It's my job to report" (P13). Some also saw reporting as the fastest way to intervene, saying that "something like that could make a child commit suicide and you want to stop that in this tracks" (P3).

Separate from teachers reporting SNCEI incidents, our participants also mentioned developing a robust system for students to report SNCEI. Some teachers considered themselves or guidance counselors on campus as safe resources for students to report concerns. This included reassuring students that their reports would only be communicated with appropriate parties. Teachers also recognized that students might feel more secure about reporting cases of SNCEI if the school had policies in place that kept their identities anonymous. P16 described this by saying, "Even though [the students] are comfortable with me, they don't tell me everything. I think having an anonymous place for them to go, maybe, to report that, it's very important for them. Not to be

[6]Depending on the state, these agencies may have different names, such as child protective services (CPS) or department of children and families (DCF).

[7]https://cops.usdoj.gov/supportingsafeschools

looked at as a tattletale or a rat between their peers."

Some teachers expressed concerns about what would happen after reporting. One teacher said they would only report to an SRO "who actually did their work because not all of them do" (P7). Each school's administration will inevitably have different procedures, and though some teachers described having positive relationships with their colleagues and supervisors, others were less positive. One teacher remarked that their school administrators might regard SNCEI as a "sweep under the rug issue", as "[administrators] would get the police involved eventually, but they're not the best at following through on things" (P12). Another teacher's faith in their administration handling SNCEI was low because they had found out about a "grossly mishandled" recent case of sexual assault. Many teachers said that after reporting, they might be asked to contribute to a written report, but then the incident would be out of their hands entirely. Teachers saw administrators as responsible for informing parents and guardians, though in a few cases teachers might also be asked to join those meetings.

Teachers named that existing school or school district policies about (cyber)bullying, harassment, threats, or pornography might be relevant for SNCEI, though the confidence they had in the policies varied. In a number of cases, these policies were shared with students in a code of conduct or handbook, potentially that they and their parents or guardians were required to sign. However, these policies were not always be followed — "we're supposed to [be zero tolerance] with bullying, but I feel like it still happens" (P8) — or even remembered: "I think there is a cyberbullying policy, but if I'm having trouble coming up with what it is, I guarantee you the kids have no idea what it is" (P12). Combining this less-than-solid faith in existing policies with the novelty of SNCEI tools, teachers expressed a need for expanding current policies or developing new policies entirely: "I would be surprised if schools had policies in place at all to handle something like that" (P9). P7 explained that there should be a standardized policy about what reporting processes and consequences would apply in a case of SNCEI for all schools in a given district, or even

potentially at a state level. Another benefit of creating new policies outlining consequences could be a deterrent effect.

Proactive educational measures against SNCEI. Teachers suggested a variety of measures that could be taken proactively against SNCEI. This included providing sex education where students would have a safe space to talk about sex, sexuality, and consent. In particular, teachers emphasized the importance of discussing consent with students, and connected SNCEI with how it violated an individual's ability to consent to images being made.

> "This is a form of getting consent from another person, and it's kind of like breaking that trust you have with that person. You are destroying an image of a person that gets spread around whether it was intended to harm them or not." (P4)

However, in order for sex education to be effective, teachers thought that it should be a mandatory part of students' curriculum. P4 observed that if it was not, students might not have as much time to internalize the meaning of consent, saying: "Because even though we talk about it [consent] during sex ed, a lot of my kids disappeared when they found out we're doing sex ed. They skipped."

Teachers also discussed providing digital safety education as a way for students to protect themselves from becoming victims of SNCEI. Strategies ranged from educating students on the types of information they put on the internet (P2, P8), to the longevity of posting content online and the difficulty of removing content once it exists on the web (P11, P13). P2 acknowledged how teenagers might experience social pressures to share images, and brought up how SNCEI perpetrators could create explicit images of students using seemingly "innocuous" pictures, saying:

> "I would just let them [students] know like, you always using Snapchat y'all are getting to this age where people are gonna start asking for pictures, and you're gonna feel compelled to do that, and knowing that even if you send them a picture where you are dressed regularly, or whatever. Somebody can still take your face and put it on

someone else's body and send it out there."

Teachers also believed that teaching students about how the consequences of their actions, should they create SNCEI, would be an effective deterrent. P3 summarized this:

> "I would have a conversation just let them [students] know about the repercussions because if no one tells them that there are repercussions for it, then they'll just keep doing it."

Other teachers thought that framing the creation of SNCEI as an activity that fell outside accepted social norms would discourage students from doing so. When asked how they might talk to students about SNCEI, teachers used adjectives such as "bad" (P13), "wrong" (P3, P4), and "negative" (P1) as ways to discourage students from creating it. For instance, P4 simply said, "some kids just need to be told: this is wrong, this is right, don't do this."

Nevertheless, teachers acknowledged that SNCEI might happen at their school, and wanted to be able to provide guidance to students to help them understand what to do in the aftermath. This included providing reporting resources, raising awareness among students by having open conversations with students (P5, P6, P13) and guardians (P6), and prominently posting contact information to school support groups (P6). P6 also mentioned a website called *Teachers Pay Teachers* as a way for teachers to create and share educational materials about SNCEI that could be used by peers.

Incident response measures. Beyond or after reporting, teachers discussed what types of punishments schools might impose on perpetrators in the immediate aftermath, as well as the role of mental health support networks for longer-term behavioral change.

Punitive measures: The overwhelming majority of teachers thought that students who created SNCEI would face disciplinary action — 16 of 17 teachers thought that their schools would either

expel, or suspend, or send students to alternative schools.[8] Teachers had varying opinions on what punitive measures would achieve: some thought that suspending students would be a clear way of communicating to students that schools perceived SNCEI as a serious infraction of rules. Using similar reasoning, participants talked about how the punishment itself would be used to deter perpetrators from engaging in harmful behavior in the future, where P11 said that students "should be aware that there's consequences, to try to cut that behavior off before it gets too far in the future where it spirals out of control", and P3 described sending students to alternative school as a way of telling perpetrators that "if they were an adult, this is the closest to jail that you can get."

Other teachers thought that expelling the perpetrator or sending them to an alternative school was a strategy to separate them from victim-survivors. P14 explained, "You can't have a successful school if there are students in it who were made to feel unsafe by the behavior of other kids." This separation could also be achieved without suspension or expulsion: one teacher mentioned a "stay-away order" that acted as an official notice to keep students separate, saying that it acted as a "report that pops up in the system that these two can't be in a room together" (P15), as well as other in-school contexts, such as lunches (P7).

Punishment-focused consequences also included removing the perpetrator's access to technology on campus. This ranged from students who would be "checking [their] phone in as soon as [they] get to school" (P2), to removing their ability to use school-provided laptops: "sometimes we take their Chromebooks away, and we freeze their accounts for a while" (P11).

Recruiting therapists, counselors, and psychologists: Beyond having clear consequences in place, teachers also wanted schools to have a support network that could address students' mental health needs: for victim-survivors, perpetrators, and any other affected students. Teachers emphasized

[8] Alternative schools are disciplinary programs where students are removed from their classrooms, and sent elsewhere to receive their education. Typically, they must stay at their alternative placement for a predetermined length of time before being allowed to return.

that this type of support would ideally come from staff or outside resources, such as therapists and psychologists who would be "specialized" and "well-trained" (P14).

In addition to thinking about mental health for victim-survivors, some discussed the importance of long-term rehabilitation of perpetrators, calling mental health counseling one of the "biggest things" (P13) that needed to happen in the direct aftermath, and hoping that they would be able to "find out what's really going on internally to make them do something like that" (P3). By identifying the underlying reasoning and motivation of the student, teachers and mental health experts could address not only the behaviors, but also their root causes and help the student avoid future harmful behaviors that went beyond a singular incident:

> "It needs to be child psych, a therapist involved to discover the why, so that way, we can help them have their whole behavior change, and not just target this one particular situation that you just happened to get caught at." (P10)

Evidence management: In the aftermath of an incident, teachers discussed balancing preserving enough evidence to support victim-survivors, with deleting the images and stopping the sharing and spread of harmful content. In support of preserving the images that perpetrators created, teachers were concerned that "if there's not enough evidence or documentation, then nothing happens" (P15), and that in order to do so, one of the first things they might do would be to prevent students from deleting the proof from their phones in order to show authorities (P8). Further, evidence might be used to support investigations, and identify the individual responsible for creating them, as well as anybody who shared them. On the other hand, other teachers believed that the best course of action was to remove evidence from not only the perpetrators' device, but also of any other students' who might have viewed the images, such as P16 who said "I think that making sure that all the material is erased. From the computer, and to whoever it's been shared with."

Interventions beyond the school: Technology. Teachers also discussed possible technical solutions that generative AI companies could integrate to prevent abuse. When discussing what AI companies could do to prevent SNCEI, teachers suggested methods that would limit youths' access to the tools, such as age restrictions or requiring valid forms of payment, indicating that there are "kids that aren't going to able to afford that 20 bucks, and that should eliminate the risk altogether" (P8).

In addition to preventing access, teachers thought that AI platforms should have built-in techniques for detecting and reporting inappropriate content, saying "any software that has artificial intelligence that can pick up that it's a child's face or body we're using and just shut it down before it can even be made" (P8). Others agreed that platforms were in unique positions, and could prevent people from even attempting to create face or body swapped images, or help investigations to identify the creator.

Lastly, teachers hoped that platforms would have clear indications that distinguished generated images from genuine pictures of people through mechanisms similar to fact-checking information found on social media pages. While teachers had many hopes for platforms' role in preventing SNCEI, we note that these solutions are less straightforward in practice, as we discuss in Section 8.5.

Interventions beyond the school: Policy. Beyond school policies, teachers wanted clear legislation on local and national levels. When considering SNCEI, teachers likened it to existing laws, such as "child pornography" and "revenge pornography," but was uncertain whether or not the synthetic nature of SNCEI meant it fell under the same policies, asking "Is this even considered pornography, even if it's a deepfake?" (P13). As such, many teachers hoped that creating clear legal policies would deter SNCEI, but acknowledged that getting consensus from all states to create a national policy might face political opposition or First Amendment concerns. In general, however, teachers were optimistic that creating laws around AI would promote ethical uses and

prevent abuse.

8.4.5 Broader Sociopolitical Context

In Section 8.4.4, we described specific interpersonal and institutional actions that teachers thought should or would happen if they found out about a student creating SNCEI. However, many teachers also referred—either implicitly or explicitly—to societal discourses about punishment and justice, child development, and educational institutions that informed their opinions about SNCEI. In this section, we describe areas of consensus or tension in specific teachers' opinions, relative to a broader sociopolitical context.

Differing conceptions of punishment and justice. Some teachers saw punishment as the only way to teach students about appropriate behavior or convey consequences. P8 suggested to "ban them from sports and dances, possibly for the rest of the school year, so they can feel as alienated as they made the person that they did this to" and P7 similarly remarked, "I would send them to some crappy school." Teachers spoke about wanting to dissuade students from taking harmful actions by outlining legal and carceral consequences, such as being registered as a sex offender, having distribution of "child pornography"[9] on your record, or being "labeled a pedophile" (P7). In short, these teachers saw the threat of prison as an effective form of deterrence, to "scare them into [doing the right thing]" (P8).

However, other teachers described why punishment would be ineffective. Students may see out-of-school suspension like a "vacation" and in-school suspension as a "fun" way to get attention from teachers (P8). P10, who had a master's degree in education, summarized that "all the graduate research has shown that suspensions don't help. They just don't." Expulsion was not seen as much better. P16 disliked expulsion because "you're really just taking away what [students] really need,

[9]Advocacy organizations use child sexual abuse material instead of "child pornography" to more accurately describe the abusive nature of this content. [467]

which is school." Though expulsion might remove a student from the place of harm and convey the message that they should not have created SNCEI, P10 predicted that "they're going to try something else in another context, because you haven't gotten to the root of the problem."

Skeptics of punishment also questioned whether the it was actually to motivate behavior change, or if it was only out of desire for retribution: "victims don't want rehabilitation [for a perpetrator]. They want the person to be punished." (P10) Rather than focusing on punishment, these other teachers discussed ways to get kids to understand the harms of their actions, either to themselves or subjects of SNCEI images. Four teachers each proposed using a conversational approach to asking questions that would get their students to put themselves in the shoes of others. For example, teachers said they would find news stories about SNCEI and ask, "How do you think the victim felt about this? How do you imagine that it impacted their life going forward?" (P15) or "What if that was your parent, what if that was your sister?" (P9). These approaches were favored as a way to help students grow and teach empathy.

This empathetic and pedagogically focused approach aligned particularly with the three teachers who detailed that restorative justice approaches were already being used for responding to conflicts between students in their schools. Restorative justice is a framework and ideology for repairing harm in relationships, often focused on community accountability [301, 135]. For example, P15 described restorative justice approaches as "a good move away from just simply punitive approaches to discipline which tend to alienate kids and oftentimes aren't applied fairly depending on the implicit biases of the administrator." However, teachers cautioned that students had to take this opportunity seriously: to genuinely see restorative justice meetings as a place for repair and healing. If students who had caused harm did not buy into the process and merely saw it as "a system they can continually abuse" (P9) to avoid consequences, restorative justice was doomed to fail. Two of the three teachers who had experience with restorative justice in their schools were dubious about using restorative justice in the aftermath of a case of SNCEI, based

on the reasoning that undoing the harm of SNCEI would be impossible and meetings would be re-traumatizing for the subject of the imagery. Ultimately, these teachers shared the notion that consequences were necessary to motivate behavior change:

> "Some of my kids hit police department, and then they finally get their first real consequence and then it packs them for the rest of their life. I could see slaps on the hand, slap on the hand, slap on the hand, talking to, discussion, and then a kid does a deepfake and is arrested and is like, well, I didn't know. And that's our failing for not having had any kind of stop along the way where they started to realize, I'm really pushing it too far." (P9)

In this way, teachers' varying perspectives on consequences — whether in a punitive or restorative justice framework — underlie not only how they might respond if a student created SNCEI, but also how to structure the broader systems of their schools.

Child development. Throughout the interviews, teachers frequently highlighted that students were still learning about life, themselves, and others. A common refrain was that kids "just don't think" and do not understand consequences: "for your average American student at this [age], I can almost guarantee they're not thinking one or two steps ahead" (P12). Particularly when it came to consequences relating to the internet, teachers reported that students were "under the impression that what they do online has no repercussions" (P1) and would do things online that they would not do in-person. Teachers described how this could manifest in students appearing to lack empathy. P17 described this specifically for groups of boys: "If they don't have somebody giving them, honestly, lessons on empathy and just self-awareness and being reflective about things, then they run wild with a lot of ideas."

In the context of romantic or sexual relationships, students' social and emotional development meant they were learning about consent and how to engage in safe and kind ways. Beyond the

range of topics covered in sex education courses, teachers also mentioned cultural influences on how students saw intimate relationships. On one hand, exposure to social media meant students were "far more aware of what abusive and toxic relationships look like" (P9), but growing up in the US meant "sex can feel more hush-hush... compared to Europe where sex is more open, and it's talked about, and it's not as big of a deal" (P14). School was seen as a place for students to develop relationships with others, figure out what they wanted, and have new experiences. However, some teachers discussed this going too far, and described a culture that encouraged sending explicit imagery. P14 hoped that schools could play a role in "having kids feel like they have enough self-worth and ownership of their body to comfortably decline and not feel like they're sacrificing something or giving some part of themselves to someone to impress them, or to build a relationship with someone." Ultimately, teachers saw the role of schools is to provide education in age-appropriate ways, informing their perspectives on consequences.

Institutional pressures. During interviews, some teachers alluded to how their schools were already strapped for resources and staff, influencing their ability to engage with preparing for or responding to SNCEI. Prioritizing between numerous issues at school was a challenge, as P2 described: "our counselors are so busy, and so many kids are going through something." Yet teachers were acutely aware that taking swift action was crucial because "other things will come up... it's just gonna keep being put to the side, it's just gonna become another folder on somebody's desk, and it may or may not get tended to" (P12). Institutional deficits could also be worsened by subpar leadership:"[handling cyberbullying] all depends on the school principal" (P7).

School policies about online incidents or off-campus incidents may have also reflected these institutional pressures. Typically, teachers described that their schools did not have any jurisdiction for issues outside of school, although some teachers wished that more could be done, recognizing that outside conflicts still affected students during school. However, one teacher noted that their

school administrators did address outside issues, including cyberbullying, "if it happens to be targeting a particular individual and their safety in particular" (P12).

Particularly when asked about whether they would talk to their students about SNCEI, teachers often mentioned ways that they were aware of the precarity of their employment. While some teachers said that they would talk to students about SNCEI — one teacher commented that they regarded the interview content was so interesting that they were going to bring it up in class the next day — others expressed concern. Among other reasons, P1 remarked, "I'm too worried about the concept of stirring the pot and creating any extra issues" and stated they would not bring up SNCEI. P5 saw SNCEI as outside the bounds of their work: "Your job as a teacher is, they want to try to keep your focus on what you teach, not on the personal lives of the students." Teachers also mentioned the politics of their state, e.g., being a conservative state with restrictions on discussing sex education, or their own gender, e.g., having discomfort discussing sensitive issues as a male teacher. The stability and confidence to which teachers had in their positions seemed to relate to the capacity they had to imagine different systems, and how much belief they had that they could act effectively in a situation involving SNCEI. Some teachers were additionally concerned about discussing SNCEI with their students, saying "the more you talk about it, the more it's out there, the more that certain students might have that idea now" (P12), while another teacher described the risk of encountering the "Streisand effect" (P14) where students might contrarily amplify actions that teachers wanted to prevent.

8.5 Discussion

Within the last two years, SNCEI has expanded beyond niche communities using specialized machine learning tools to create SNCEI of celebrities, now having reached US middle and high school students casually creating SNCEI of their classmates.

New technology, similar interpersonal harm. In this research, we interviewed teachers to shed light on their concerns for the imminent future of SNCEI in schools. New technologies tend to come with the hope of a better future; generative AI technologies were ostensibly intended to unlock a new era of human creativity. However, early reports show that youth in the US are not (only) using these generative AI technologies for ends beneficial to society. Synthesizing the SNCEI threat models that teachers described in this work, we describe how students' use of generative AI tools are likely to exacerbate existing interpersonal harms: cyberbullying and gender and sexual abuse. In the words of P14: "there are just bad actors and creeps who do things that they shouldn't, and that's kind of been true of human history before the advent of technology like the internet." Thus, technical interventions to mitigate SNCEI are urgently needed and can have significant value, but they are also inherently limited as only one of many approaches to mitigate harm. Still, given that interpersonal harm is a particularly entrenched problem in schools, the novel technical aspect of SNCEI may prove useful in drawing much needed resources — social, technical, or otherwise — to supporting teachers and schools in addressing them.

We're all in this together. Stepping back, teachers and school contexts are only one part of a larger system where interpersonal harms manifest. While the school context introduces unique challenges, e.g., that students tend to be young and are still learning about themselves and the world, it also offers unique opportunities. Middle and high schools can be places to introduce prosocial values and educational opportunities not feasible later in life. Schools are a nexus of cultural forces, from those within the purview of individuals to communities and broader society. Building on our key findings, we now discuss challenges and opportunities to address the harms of SNCEI sociotechnically, using a multi-pronged framework of individual, community and societal interventions.

8.5.1 Individual interventions

Teachers and school staff. Teachers are deeply passionate about supporting their students and are in classrooms with students every day. Resources and professional development courses about SNCEI for teachers could be a well-positioned intervention to provide immediate impact. Such resources could include prevalence statistics, technical descriptions of SNCEI tools, and ways to respond if teachers do suspect SNCEI.

Support for teachers regarding SNCEI can learn from existing efforts to respond to cyberbullying and gender and sexual abuse. About 1 in 6 high schoolers in the US reported cyberbullying in the last year, and 1 in 9 reported sexual violence [116]; to tackle these pervasive and serious issues, many efforts are being developed and evaluated in research outside of the computing literature [93]. Future work could, for example, compare warning signs of cyberbullying to warning signs of SNCEI specifically, updating existing trainings for teachers as appropriate.

Additionally, future work could also explore how to support or provide resources for school staff responsible for WiFi, student laptops and tablets, or other technical systems. Similarly, teachers often mentioned that they would turn to school counselors, psychologists, or social workers if they heard about SNCEI. While teachers usually have the most direct face time with students, all of these other school staff also play important roles in establishing schools as safe learning environments.

Parents, guardians, and caregivers. Though we did not interview parents, guardians, or caregivers in our study, teachers mentioned that these adults could either reinforce positive lessons from school or be adverse forces in students' lives. Future work could explore the perspectives of these adults about SNCEI, as well as how to communicate strategies for them to support youth who may be involved in incidents.

Students. During our interviews, some teachers also relayed anecdotes about specific students

who were positive influences at their schools. Given the deeply interpersonal and social nature of SNCEI dynamics explored in this work, peer-led outreach programs may be especially impactful in mitigating harm. Such programs could encourage students to share relevant information about SNCEI with each other and provide mutual support.

8.5.2 Community interventions

Setting prosocial norms in schools. School staff and school district administrators have meaningful ability to set the priorities and policies that govern school activities. In this way, such people could play an invaluable role in proactively developing mitigation strategies for SNCEI. The use of generative AI technologies for SNCEI reiterates the importance of school content that facilitates social and emotional learning, sex education, and online safety. During interviews, P15 spoke about how establishing a strong sense of community and belonging would be the most powerful against SNCEI:

> "I think that the best preventative is fostering a school culture where bullying and disrespectful behavior are not tolerated and not sanctioned. Trying to foster a culture where students have a sense of belonging... I want for students to feel that they have some kind of a sense of belonging within the classroom and students feel the need to watch out for each other." (P15)

Teachers in our interviews acknowledged the value of school environments for socializing students to being more thoughtful and empathetic, as well as building self-awareness and relationship skills. Teachers also believed that these skills would help students, and the future adults that they will grow to be, to not create SNCEI or inform adults if they did find out about SNCEI. Further, given the gendered and sexualized nature of the harms of IBSA, ensuring that students learn about the importance of sexual consent through sex education courses could also mitigate SNCEI creation.

Incorporating discussion of SNCEI within existing courses about social and emotional learning, sex education, or online safety would have the additional benefit of not drawing too much attention to the new capabilities of generative AI technologies. Teachers were worried about a possible "Streisand effect", i.e., that bringing up SNCEI to warn students not to create it would actually backfire and draw attention to tools that students otherwise did not know were available. With the proper framing within broader sex education or online safety content, SNCEI could be contextualized as a one instance of a broader context of serious harm, and reduce the attractive novel quality of a new technology.

Policies to facilitate accountability and repair. Finally, schools could continue to develop alternative justice models for facilitating accountability and repair in situations of harm, including for SNCEI. Restorative and transformative justice are alternative justice frameworks that have been practiced as grassroots community efforts to address societal harms [458, 135], including but not limited to domestic and sexual violence. Restorative justice practices had already begun to be a part of some of the schools that teachers worked at, though teachers expressed skepticism about whether a restorative justice approach would effectively address SNCEI, which they feared was too serious of an issue. However, some of the origins of restorative and transformative justice include Black and Indigenous communities that intended to address sexual and domestic violence [301]. While restorative justice focuses more on interpersonal relationships and transformative justice focus more on societal systems, both intend to find ways to respond to violence and harm without causing more violence and harm [301]. Challenges may arise in applying restorative justice frameworks in schools due to incomplete community buy-in to the process or imperfect attempts that dissuade individuals from investing in alternatives. However, we found that teachers' calls for nuanced and empathetic interventions for students who cause harm were well-aligned with restorative justice principles, for example, distinguishing between punishments ("inflicting cruelty,

pain, and suffering") and consequences ("being uncomfortable and losing some privileges") [301]. Therefore, much future work can be done to explore applying restorative and transformative approaches to addressing SNCEI in schools, including by drawing on existing toolkits [135, 458] and abolitionist teaching [350].

8.5.3 Societal interventions

During our interviews with teachers, many had hopes that solutions to preventing SNCEI at schools would come from either technical or legal domains. However, we note that both technical and legal solutions are still currently being developed, and face multiple challenges, such as how to balance policies that allow for the consensual (synthetic) sexual content while inhibiting actors who seek to create abusive material.

Technical solutions: self-regulation and deplatforming. When discussing generative AI platforms, most teachers had mental models of companies who acted in good faith and had incentives to prevent abuse. While there has been movement from larger entities to reduce image-based sexual abuse [601], many abusive platforms operate using the "nudify-as-a-service" model, whose singular purpose is to create explicit images. Soliciting voluntary cooperation from these parties may not be as straightforward as other companies have a more vested stake in maintaining their reputation. Technical studies to investigate such tools and groups that use them could shed light on effective ways to inhibit harmful outcomes.

Assuming that a hypothetical solution leads to deplatforming websites, it is still unclear who decides what content would be permissible, and which parties should be responsible for deplatforming. Though there has been some precedent in infrastructure providers deciding to withhold services to websites that host harmful content [54], whether or not this is a long-term solution, or even if the decision should fall on these entities is still an unresolved question.

Legal solutions and challenges. Some teachers used the threat of criminal consequences to deter students, using terms such as "child pornography" to impress upon students the gravity of creating SNCEI. However, whether this applies unilaterally across the US is an evolving matter. US states lack consensus on whether synthetic CSAM is legally equivalent to content not generated by generative AI tools. Legal scholars have pointed out that laws also have to contend with scenarios where individuals under the age of 18 create explicit images voluntarily of themselves (sometimes called "self-generated" or "voluntary" CSAM), such as youth exploring their sexuality and sending each other explicit material. Finally, yet-to-be-settled policies need to address whether youth who create synthetic CSAM should face the same legal ramifications as adults.

8.5.4 Future research

Further work by researchers from multiple backgrounds is needed to inform future interventions. In this work, we encountered challenges due to our status as mandatory reporters (see Sections 8.3.4 and 8.3.5), which required sensitivity and care to navigate. Research inevitably raises ethical considerations; navigating ethical and legal dilemmas in child abuse research is an open area of research [328, 235, 189], similar to ethical considerations of research with social media data [194, 508]. HCI researchers are well-positioned to navigate such challenges, particularly by bringing epistemic diversity, i.e., varied research methods, methodologies, researcher backgrounds, and research goals, to bear on complex sociotechnical issues.

Specifically, future work could investigate the perspectives of parents and guardians, law enforcement, school administrators, as well as students themselves. As SNCEI becomes more ubiquitous, follow-up studies can also explore teachers' direct experiences with SNCEI and how victimization or perpetration changes over time. Our study provides just one of many perspectives that will be needed to create robust and effective policies about SNCEI.

8.6 Conclusion

In this work, we conducted interviews with 17 teachers at US middle and high schools to use their thorough understanding of youth to construct threat models about SNCEI creation. We used a security and privacy lens to determine that while anyone could create SNCEI, teachers mostly anticipated that boys would create SNCEI of girls for cyberbullying and gender and sexual abuse. Teachers additionally described a multitude of interventions for mitigating the harms of SNCEI, calling for both proactive educational measures, e.g., sex education or digital safety education, as well as robust incident response measures. However, teachers' evaluation of these potential measures also depended on their conception of justice, as some teachers favored punitive approaches, while others advocated for restorative justice approaches. We synthesize our results to develop directions for teachers, schools, and technologists and policymakers to inform how to mitigate the harms of SNCEI and while also creating space for consensual online intimacy.

Acknowledgements

We thank our reviewers for their helpful feedback, Caroline Shelton for consultation about mandatory reporting, and to Galen Weld for insights about social media recruitment. We are deeply grateful to our pilot participants (Lauren Bricker, Jeremy Muench, John Rumney, and Gadiel Williams) as well as our study participants for sharing their perspectives and their dedication to their students. We acknowledge the UW Center for Studies in Demography and Ecology (CSDE) and Student Tech Fee for qualitative data analysis software access. This work was supported in part by the NSF under Award 2205171.

Part IV

Conclusion

Chapter 9

Conclusion

This dissertation has presented my work against technology-facilitated abuse and towards sociotechnical security and privacy. In Part I, I evaluated existing support for online abuse, taking an ecosystem-level perspective. I evaluated security advice to mitigate online hate and harassment as well as support sought and received on Reddit to mitigate image-based sexual abuse. In Part II, I mapped societal factors—sociodemographic factors and gender stereotypes—to technical security and privacy to develop conceptual tools that bridge the two. In Part III. I characterized emerging online abuse threats with particular attention to the technical and societal factors at play: techniques shared on TikTok for interpersonal surveillance and control and synthetic nonconsensual explicit imagery generated by students in schools. Completing this work further contributed demonstrations of creative ways to use information on social media, such as help-seeking posts (Chapter 4) or online discourse not immediately evident as related to security and privacy (Chapter 7).

Working from an interdisciplinary position within the technical fields of computer security and privacy and human-computer interaction, as well as drawing on social perspectives, I position and advance the study of online abuse as a key domain of S&P challenges. Further, I seek to create bridges for scholars in these fields to see their complementary skills and work together. I now

reflect on remaining challenges and future opportunities for research and beyond.

9.1 Fostering Nuanced Perspectives of Harm

Looking beyond the individual approach of security and privacy recognizes the systemic pressures on individuals to act in certain ways. This is not to say that experiencing systemic oppression will be a guarantee of causing harm, but rather a risk factor. Scholars in interpersonal violence have noted the "victim-offender overlap": i.e., the high likelihood of people who have perpetuated abuse to have themselves experienced abuse. This complexity suggests that binary distinctions of 'perpetrators' and 'victim-survivors' are not accurate. However, how to ensure people who cause harm are held accountable without using their prior experiences to minimize the harm they later caused, is an ongoing challenge in socio-legal contexts.

Designing and evaluating technologies with this complexity in mind will further be a difficult sociotechnical challenge. A necessary precondition is to resist simplistic, binary explanations for pheonmena or judgments of people. Further, the concept of defense in depth, i.e., the security strategy of employing many layered defensive measures, may yield transferable insights: multiple approaches that create accountability, not only at the individual or societal, but also at various communal levels may help. Systemic perspectives call on us to re-imagine the source of safety; rather than relying solely on institutional power and authority, how might technology facilitate other ways for people keep each other safe?

Collective and community approaches. Data privacy scholars have begun exploring collective approaches, including calling for collective data refusal [628, 627], collective harm reporting [615], and collective action [151]. Collective action could be generated through participatory systems that allow the public to "wage a coordinated digital protest campaign" through various tactics, e.g., "digital boycotts, data strikes, social media sit-ins" [151]. Similar tactics have been effective for

bringing about social change, giving reason to believe they can also be effective for sociotechnical change.

Social computing systems can also continue exploring smaller or decentralized subcommunities with community-based moderators, such as learning from and working with volunteer moderators on Reddit and Discord. Communities are also ideal places to test different interventions—though this should be done ethically and with the community's consent—such as account suspensions of different lengths [220], bans, or other types of digital punishments and accountability measures. Computer scientists are well-poised to design new affordances that offer communities different levers for change, beyond the now common platform block and mute features that are clearly still insufficient.

Navigating automation responsibly. Ultimately all computing systems are designed and deployed by people. Many forms of sociotechnical harm that are caused by automation [526], whether based on machine learning or algorithmic or labeled "artificial intelligence.' In addition to the disastrous environmental consequences of machine learning and widespread exploitation of labor that has contributed to the capabilities of machine learning algorithms today [45, 210], there remains another significant risk. The hype around "artificial intelligence" risks sidestepping the accountability of people and institutions who implement and maintain automated systems [168, 419], particularly for issues of security and safety. What does it mean for a computing system to be held accountable when it enables people to perpetuate online abuse and harm?

Researcher safety and ethical considerations when doing research. A major remaining challenge is how to ensure researcher safety when doing work on online abuse. In my personal experience, lots of self-education was required. S&P researchers are relatively new to studying interpersonal abuse and TFA, and actively updating best practices for doing so [44, 47]. Useful resources come from many fields, such as social work, journalism, HCI, social computing, and

more [366, 488, 169, 322, 192, 27, 515, 393, 149]. I especially recommend *Trauma Stewardship: An Everyday Guide to Caring for Self While Caring for Others* [346] and other resources on secondary or vicarious trauma. Above all: work together, ask how your colleagues are doing, and try to listen as non-judgmentally as possible.

9.2 Achieving Security and Liberation

Security and privacy researchers often come to believe that true security or privacy is elusive. Despite marketing materials that make superlative claims about how secure or private a technology is, e.g., VPNs [9] or secure messengers [8], the state of being completely protected from harm frequently seems unattainable. Adversaries and their capabilities evolve, which is why threat modeling is relied on as a structured process for prioritizing some of the most likely or damaging threats. Nevertheless, computer security and privacy researchers and practitioners continue to strive for progress and harm reduction [476]. Where might sociotechnical security, privacy, and safety efforts go next?

Learning from efforts for emancipation and liberation. Over thirty years ago, international relations scholar and political scientist Ken Booth theorized about the relationship of security with respect to emancipation in formal political and legal systems:

> 'Security' means the absence of threats. Emancipation is the freeing of people (as individual and groups) from those physical and human constraints which stop them carrying out what they would freely choose to do... Emancipation, not power or order, produces true security. Emancipation, theoretically, is security. [62]

As a theorist of international security, in the realm of nation-states and war, Booth laid out a vision of how people's social power could predict safety in a society: societies with less emancipation and greater subjugation would produce more global insecurity. Transferring these international

dynamics to interpersonal dynamics and in online environments, we can hypothesize that disempowerment and oppression may explain why some people pose harm to others. Indeed, creators of TikTok videos in Chapter 7 appealed to women, who have immense domestic expectations placed on them compared to men, and encouraged women to resort to surveillance; many online scams are perpetrated by people who experience economic scarcity [283, 580], potentially accounting for some of the financial sextortion studied in Chapter 4.

In this way, the ideal of emancipation—and even broader movements for liberation[1]—are deeply informative for computing researchers and practitioners concerned with advancing sociotechnical safety. Emancipation and liberation share with security and privacy the condition of having an elusive nature. It can be difficult to imagine a world without systemic oppression, a world liberated of sexism and gender oppression, racism and xenophobia, and countless other hegemonies; just as it can be difficult to articulate what a world with perfect security and privacy might be like. But struggles for emancipation and liberation are a guide to understanding the origins of security and privacy harms, and thus provide ideas about how to better address the root of the issue. A growing body of work already explores power relations in security and privacy with a liberatory aspirations, such as by studying carceral and surveillance technologies [432, 431] or in intimate communications [211], among many others.

For engineering security and privacy systems, Booth's perspective on international security also provides useful guidance:

> Emancipation should logically be given precedence in our thinking about security over the mainstream themes of power and order. The trouble with privileging power and order is that they are at somebody else's expense (and are therefore potentially unstable)... True (stable) security can only be achieved by people and groups if they

[1]An imperfect but clarifying distinction between emancipation and liberation is that "emancipation means becoming what one already is; liberation is becoming what one is not yet" [25].

> do not deprive others of it. [62]

This demonstrates how security and privacy systems that privilege power at others' expense, such as is often the case with surveillance technologies, are unlikely to be sustainable. Solutions that do not constrain or further disempower people are the way forward. Online platforms must design and implement interventions that enable people to depend on each other in pro-social ways, such as by setting community-driven expectations of behavior and accountability measures when someone causes harm. Prior measurement work in security and privacy has demonstrated a multitude of ways that harm manifests at scale, and sociotechnologists can build additional systems that facilitate further measurement work. Sociotechnologists could also build systems to enable experts already working on offline contexts, such as social workers or other service providers, to provide support for people who experience harmful sociotechnical interactions, such as by supplementing the help-seeking already happening on Reddit (Chapter 4). During any such efforts, new mitigations for reducing harm must reinforce people's autonomy and freedom.

Pluriversality toward safety. This dissertation advanced a sociotechnical approach to security and privacy, largely inspired by feminist STS and social theory. This is only one of many possible approaches that might work together to increase safety. Decolonial thought emphasizes that multiple perspectives reflecting situated knowledge and diverse narratives, also called pluriversality [539, 541], can be a method and practice for enacting liberatory values. The sociotechnical approach uniquely offers benefits in acknowledging the role of systems, particularly social systems and technical systems, to enable a richer, etiological view of security, privacy, and safety. This dissertation advocates for a sociotechnical approach in conjunction with, rather than in spite of others. Individual approaches can bring about more personalized or individually affirming interventions. Specific approaches focusing on technology, social groups, contexts, or other aspects can result in the beneficial advancement of different goals.

There are many opportunities for future work related to forms of justice alternative to punitive forms, which are at the expense of an individual's freedom. Social computing researchers are already exploring translations of restorative justice principles to content moderation, social media, and gaming contexts [619, 618, 617]. Similar approaches could also be promising for harm in interpersonal relationships, as structured ways to respond to interpersonal surveillance (as studied in Chapter 7) or image-based sexual abuse in schools (as studied in Chapter 8). However, one key tenet of restorative justice is centering the person who was harmed, and not all people who experience harm may choose a restorative justice process. Therefore, open questions remain for how to evaluate which contexts and cases are most suited.

Pluriversality also includes an epistemic plurality of methods for research and pathways to impact. In many cases, quantitative methods are used to justify the status quo [406], so quantitative methods should not be privileged over others. Researchers must continue to leverage qualitative, ethnographic, or other methods underrepresented in computer science, employing different theoretical frameworks, and even inventing new methods. For scientists, impact will require engaging with policymakers and the general public to ensure knowledge is not restricted to academic silos. For academics, training new students with pluriversality in mind will enable subsequent generations of scholars and practitioners to make progress toward the elusive ideals of security, privacy, safety, and liberation. Ultimately, nobody's safe until everyone's safe.

Bibliography

[1] Women, Minorities, and Persons with Disabilities in Science and Engineering: 2021. Technical report, National Center for Science and Engineering Statistics, 2021.

[2] Heartmob. https://iheartmob.org/, 2022.

[3] Crystal Abidin. Mapping Internet celebrity on TikTok: Exploring attention economies and visibility labours. *Cultural Science Journal*, 12(1):77–103, 2021.

[4] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. You get where you're looking for: The impact of information sources on code security. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2016.

[5] Mark S. Ackerman. The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. *Human–Computer Interaction*, 15(2-3):179–203, September 2000.

[6] Anne Adams and Martina Angela Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999.

[7] Sarah A. Aghazadeh, Alison Burns, Jun Chu, Hazel Feigenblatt, Elizabeth Laribee, Lucy Maynard, Amy L. M. Meyers, Jessica L. O'Brien, and Leah Rufus. *Online Harassment*, chapter GamerGate: A Case Study in Online Harassment, pages 179–207. Springer International Publishing, Cham, 2018.

[8] Omer Akgul, Ruba Abu-Salma, Wei Bai, Elissa M. Redmiles, Michelle L. Mazurek, and Blase Ur. From Secure to Military-Grade: Exploring the Effect of App Descriptions on User Perceptions of Secure Messaging. In *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society*, WPES '21, pages 119–135, New York, NY, USA, November 2021. Association for Computing Machinery.

[9] Omer Akgul, Richard Roberts, Moses Namara, Dave Levin, and Michelle L. Mazurek. Investigating influencer vpn ads on youtube. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2022.

[10] Devdatta Akhawe and Adrienne Porter Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *Proceedings of the USENIX Security Symposium*, 2013.

[11] Devdatta Akhawe and Adrienne Porter Felt. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *Proceedings of the USENIX Security Symposium*, 2013.

[12] Mamtaj Akter, Zainab Agha, Ashwaq Alsoubai, Naima Ali, and Pamela Wisniewski. Towards Collaborative Family-Centered Design for Online Safety, Privacy and Security, April 2024.

[13] Kendra Albert. Five Reflections from Four Years of FOSTA/SESTA. *Cardozo Arts & Entertainment Law Journal*, 2022.

[14] Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Joshua Gracie, Munmun De Choudhury, Pamela J. Wisniewski, and Gianluca Stringhini. Understanding the Digital Lives of Youth: Analyzing Media Shared within Safe Versus Unsafe Private Conversations on Instagram. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–14, New York, NY, USA, April 2022. Association for Computing Machinery.

[15] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Understanding the effect of deplatforming on social networks. In *13th ACM Web Science Conference 2021*, pages 187–195, 2021.

[16] Max Aliapoulios, Kejsi Take, Prashanth Ramakrishna, Daniel Borkan, Beth Goldberg, Jeffrey Sorensen, Anna Turner, Rachel Greenstadt, Tobias Lauinger, and Damon McCoy. A large-scale characterization of online incitements to harassment across platforms. In *Proceedings of the 21st ACM Internet Measurement Conference*, IMC '21, pages 621–638, New York, NY, USA, November 2021. Association for Computing Machinery.

[17] Ashwaq Alsoubai, Afsaneh Razi, Zainab Agha, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. Profiling the Offline and Online Risk Experiences of Youth to Develop Targeted Interventions for Online Safety. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1):114:1–114:37, April 2024.

[18] PEN America. Online Harassment Field Manual. https://onlineharassmentfieldmanual.pen.org.

[19] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. Social support, reciprocity, and anonymity in responses to sexual abuse disclosures on social media. *ACM Trans. Comput.-Hum. Interact.*, 25(5), Oct 2018.

[20] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 3906–3918, 2016.

[21] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. Sensitive self-disclosures, responses, and social support on instagram: The case of #depression. In *Proc. ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1485–1500, 2017.

[22] Bonnie Brinton Anderson, C. Brock Kirwan, David Eargle, Scott R. Jensen, and Anthony Vance. Neural correlates of gender differences and color in distinguishing security warnings and legitimate websites: a neurosecurity study. *Journal of Cybersecurity*, 2015.

[23] Ross Anderson. The Dependability of Complex Socio-Technical Infrastructure, June 2011.

[24] Ross Anderson. The Dependability of Complex Socio-technical Systems. In *Fundamental Approaches to Software Engineering*, pages 1–1. Springer, Berlin, Heidelberg, 2011.

[25] José Fernando Andrade Costa. Emancipation and liberation as normative horizons in critical theory. *Thesis Eleven*, 184-185(1):16–30, December 2024. Publisher: SAGE Publications Ltd.

[26] Manos Antonakakis, Tim April, Michael Bailey, Matt Bernhard, Elie Bursztein, Jaime Cochran, Zakir Durumeric, J Alex Halderman, Luca Invernizzi, Michalis Kallitsis, et al. Understanding the mirai botnet. In *Proceedings of the USENIX Security Symposium*, 2017.

[27] AoIR Risky Research Working Group. Risky Research: An AoIR Guide to Researcher Protection and Safety. Technical report, The Association of Internet Researchers, 2025.

[28] Apple. About sensitive content warning on Apple devices. https://support.apple.com/en-us/105071, 2024.

[29] David Armstrong, Ann Gosling, John Weinman, and Theresa Marteau. The Place of Inter-Rater Reliability in Qualitative Research: An Empirical Study. *Journal of Sociology*, 31(2):597–606, August 1997.

[30] ASA. Code of Ethics and Policies and Procedures of the ASA Committee on Professional Ethics, 1999.

[31] ASA. Topic: Harassment, July 2016.

[32] Jane Bailey, Nicola Henry, and Asher Flynn. Technology-Facilitated Violence and Abuse: International Perspectives and Experiences. In Jane Bailey, Asher Flynn, and Nicola Henry, editors, *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*, pages 1–17. Emerald Publishing Limited, January 2021.

[33] Shaowen Bardzell and Jeffrey Bardzell. Towards a Feminist HCI Methodology: Social Science, Feminism, and HCI. In *Proceedings of the 2011 CHI Conference on Human Factors in Computing Systems*, 2011.

[34] Ann Bartow. Copyright law and pornography. *Or. L. Rev.*, 91:1, 2012.

[35] Catherine Barwulor, Allison McDonald, Eszter Hargittai, and Elissa M Redmiles. "Disadvantaged in the American-dominated internet": Sex, Work, and Technology. In *Proceedings of the 2011 CHI Conference on Human Factors in Computing Systems*, 2021.

[36] Corey H. Basch, Grace C. Hillyer, and Christie Jaime. COVID-19 on TikTok: Harnessing an Emerging Social Media Platform to Convey Important Public Health Messages. *International Journal of Adolescent Medicine and Health*, 2020.

[37] Samantha Bates. Revenge Porn and Mental Health: A Qualitative Analysis of the Mental Health Effects of Revenge Porn on Female Survivors. *Feminist Criminology*, 12(1):22–42, January 2017.

[38] Maya A Beasley and Mary J. Fischer. Why they leave: the impact of stereotype threat on the attrition of women and minorities from science, math and engineering majors. *Social Psychology of Education*, 15:427–448, 2012.

[39] C. T. Begeny, M. K. Ryan, C. A. Moss-Racusin, and G. Ravetz. In some professions, women have become well represented, yet gender bias persists—perpetuated by those who think it is not happening. *Science Advances*, 6(26), 2020.

[40] Amy E. Bell, Steven J. Spencer, Emma Iserman, and Christine E. R. Logel. Stereotype Threat and Women's Performance in Engineering. *Journal of Engineering Education*, 2016.

[41] Valerie Bell, Craig Hemmens, and Benjamin Steiner. Up skirts and down blouses: A statutory analysis of legislative responses to video voyeurism. *Criminal Justice Studies*, 19(3):301–314, 2006.

[42] bell hooks. *Feminism is for Everybody*. Pluto Press, 2000.

[43] Rosanna Bellini, Emily Tseng, Nora McDonald, Rachel Greenstadt, Damon McCoy, Thomas Ristenpart, and Nicola Dell. "So-called privacy breeds evil": Narrative justifications for intimate partner surveillance in online forums. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW3), Jan 2021.

[44] Rosanna Bellini, Emily Tseng, Noel Warford, Alaa Daffalla, Tara Matthews, Sunny Consolvo, Jill Palzkill Woelfer, Patrick Gage Kelley, Michelle L. Mazurek, Dana Cuomo, Nicola Dell, and Thomas Ristenpart. Sok: Safer digital-safety research involving at-risk users. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, May 2024.

[45] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada, March 2021. ACM.

[46] Kristin Berdan. An evaluation of online security guides for journalists. Technical report, 2021.

[47] Rasika Bhalerao, Vaughn Hamilton, Allison McDonald, Elissa M Redmiles, and Angelika Strohmayer. Ethical practices for security research with at-risk populations. In *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 546–553. IEEE, 2022.

[48] Monica Biernat. Gender and height: Developmental patterns in knowledge and use of an accurate stereotype. *Sex Roles*, 29:691–713, 1993.

[49] Veroniek Binkhorst, Tobias Fiebig, Katharina Krombholz, Wolter Pieters, and Katsiaryna Labunets. Security at the end of the tunnel: The anatomy of VPN mental models among experts and Non-Experts in a corporate context. In *Proceedings of the 31st USENIX Security Symposium*, Boston, MA, August 2022. USENIX Association.

[50] Rena Bivens. The gender binary will not be deprogrammed: Ten years of coding gender on Facebook. *New Media & Society*, 2017.

[51] Amy M Blackstone. Gender roles and society. 2003.

[52] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and its consequences for online harassment: Design insights from heartmob. In *Proceedings of the ACM on Human-Computer Interaction*, 2017.

[53] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. In *Proc. CSCW*, 2017.

[54] The Cloudflare Blog. Blocking Kiwifarms. https://blog.cloudflare.com/kiwifarms-blocked, September 2022.

[55] Sarah Bloom. No vengeance for revenge porn victims: Unraveling why this latest female-centric, intimate-partner offense is still legal, and why we should criminalize it. *Fordham Urb. LJ*, 42:233, 2014.

[56] Nicole Bluett-Boyd, Bianca Fileborn, Antonia Quadara, and AD Moore. The role of emerging communication technologies in experiences of sexual violence: A new legal frontier? *J. of the Home Economics Institute of Australia*, 20(2):25–29, 2013.

[57] Alicia Blum-Ross and Sonia Livingstone. "Sharenting," parent blogging, and the boundaries of the digital self. *Popular Communication*, 15(2):110–125, 2017.

[58] Dannell D. Boatman, Susan Eason, Mary Ellen Conn, and Stephenie K. Kennedy-Rea. Human Papillomavirus Vaccine Messaging on TikTok: Social Media Content Analysis. *Health Promotion Practice*, 23(3):382–387, 2021.

[59] Emma Bond and Katie Tyrrell. Understanding revenge pornography: A national survey of police officers and staff in England and Wales. *J. of interpersonal violence*, 36(5-6):2166–2181, 2021.

[60] Bram Bonné, Sai Teja Peddinti, Igor Bilogrevic, and Nina Taft. Exploring decision making with Android's runtime permission dialogs using in-context surveys. In *Proc. SOUPS*, 2017.

[61] Joseph Bonneau. The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2012.

[62] Ken Booth. Security and Emancipation. *Review of International Studies*, 17(4):313–326, 1991.

[63] Elijah Robert 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*, 2023.

[64] Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and its Consequences*. MIT Press, 2000.

[65] Richard E Boyatzis. *Transforming qualitative information: Thematic analysis and code development*. Sage, 1998.

[66] Maia J. Boyd, Jamar L. Sullivan Jr., Marshini Chetty, and Blase Ur. Understanding the Security and Privacy Advice Given to Black Lives Matter Protesters. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–18, 2021.

[67] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[68] Virginia Braun and Victoria Clarke. *Successful Qualitative Research: A Practical Guide for Beginners*. SAGE, 2013.

[69] Virginia Braun and Victoria Clarke. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4):589–597, August 2019.

[70] Virginia Braun and Victoria Clarke. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy Research*, 21(1):37–47, 2021.

[71] Virginia Braun and Victoria Clarke. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18(3):328–352, 2021.

[72] Virginia Braun and Victoria Clarke. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative Research in Sport, Exercise and Health*, 13(2):201–216, March 2021.

[73] Virginia Braun and Victoria Clarke. Conceptual and design thinking for thematic analysis. *Qualitative Psychology*, 9(1):3–26, February 2022.

[74] Virginia Braun and Victoria Clarke. Toward good practice in thematic analysis: Avoiding common problems and be(com)ing a *knowing* researcher. *International Journal of Transgender Health*, 24(1):1–6, January 2023.

[75] Chris Brickell. The sociological construction of gender and sexuality. *The Sociological Review*, 54(1):87–113, 2006.

[76] Grace Brigham, Miranda Wei, Tadayoshi Kohno, and Elissa M. Redmiles. "Violation of my body:" Perceptions of AI-generated non-consensual (intimate) imagery. In *Symposium on Usable Privacy and Security (SOUPS)*, Philadelphia, PA, USA, 2024-08-11/2024-08-13.

[77] Anna Brosch. When the child is born into the Internet: Sharenting as a growing trend among parents on Facebook. *The New Educational Review*, 43:225–235, 2016.

[78] Anna Brosch. Sharenting: Why do parents violate their children's privacy? *The New Educational Review*, 4:75–85, 2018.

[79] Christia Spears Brown. Sexualized gender stereotypes predict girls' academic self-efficacy and motivation across middle school. *International Journal of Behavioral Development*, 43(6):523–529, 2019.

[80] Kellen Browning. Apple says it will make airtags easier to find after complaints of stalking. https://www.nytimes.com/2022/02/10/business/apple-airtags-safety.html, 2022.

[81] Amber M. Buck and Devon F. Ralston. I didn't sign up for your research study: The ethics of using "public" data. *Computers and Composition*, 61:102655, 2021. Rhetorics of Data: Collection, Consent, & Critical Digital Literacies.

[82] Elie Bursztein, Einat Clarke, Michelle DeLaune, David M Elifff, Nick Hsu, Lindsey Olson, John Shehan, Madhukar Thakur, Kurt Thomas, and Travis Bright. Rethinking the detection of child sexual abuse imagery on the internet. In *Proceedings of The Web Conference*, 2019.

[83] Tor Busch. Gender differences in self-efficacy and attitudes toward computers. *Journal of Educational Computing Research*, 12(2):147–158, 1995.

[84] Karoline Busse, Julia Schäfer, and Matthew Smith. Replication: No One Can Hack My Mind Revisiting a Study on Expert and Non-Expert Security Practices and Advice. In *Proc. SOUPS*, 2019.

[85] Judith Butler. *Gender Trouble.* Routledge, 1990.

[86] W. Carson Byrd and Matthew W. Hughey. Born that way? 'Scientific' racism is creeping back into our thinking. Here's what to watch out for. https://www.washingtonpost.com/news/monkey-cage/wp/2015/09/28/born-that-way-scientific-racism-is-creeping-back-into-our-thinking-heres-what-to-watch-out-for/, 2015.

[87] Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics. In *AIES*, 2022.

[88] Corey Call. Perceptions of Image-Based Sexual Abuse Among the American Public. *Criminology, Crim. Just. L & Soc'y*, 22:30, 2021.

[89] Ryan Calo. The scale and the reactor. *SSRN*, 2022.

[90] JC Campbell and Nancy Glass. Safety planning, danger, and lethality assessment. *Intimate partner violence: A health-based perspective*, pages 319–334, 2009.

[91] Julia K Campbell, Sydney McCartin Poage, Sophie Godley, and Emily F Rothman. Social anxiety as a consequence of non-consensually disseminated sexually explicit media victimization. *J. of Interpersonal Violence*, 37(9-10):NP7268–NP7288, 2022.

[92] M Ariel Cascio, Eunlye Lee, Nicole Vaudrin, and Darcy A Freedman. A team-based approach to open coding: Considerations for creating intercoder consensus. *Field Methods*, 31(2):116–130, 2019.

[93] Wanda Cassidy, Chantal Faucher, and Margaret Jackson. Cyberbullying among youth: A comprehensive review of current international research and its implications and application to policy and practice. *School Psychology International*, 34(6):575–612, December 2013.

[94] Cyber Civil Rights Initiative (CCRI). Existing laws on nonconsensual distribution of intimate images. https://cybercivilrights.org/nonconsensual-distribution-of-intimate-images/.

[95] CDC. About Intimate Partner Violence. https://www.cdc.gov/intimate-partner-violence/about/index.html, January 2025.

[96] Vanessa Ceia. Gender and Technology: A Rights-based and Intersectional Analysis of Key Trends. Technical report, Oxfam, 2021.

[97] Pew Research Center. Social media fact sheet. https://www.pewresearch.org/internet/fact-sheet/social-media/, 2021.

[98] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, (CSCW), 2017.

[99] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[100] Elizabeth Charters. The use of think-aloud methods in qualitative research: An introduction to think-aloud methods. *Brock Education*, 2003.

[101] Rahul Chatterjee, Periwinkle Doerfler, Hadas Orgad, Sam Havron, Jackeline Palmer, Diana Freed, Karen Levy, Nicola Dell, Damon McCoy, , and Thomas Ristenpart. The Spyware

Used in Intimate Partner Violence. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, SP '18. IEEE, 2018.

[102] Bridget Chavez. No charges as AI-generated nude pictures of female students circulate around Issaquah school. https://www.kiro7.com/news/local/no-charges-ai-generated-nude-pictures-female-students-circulate-around-issaquah-school/MCQTOKWRVREPTK3K2IAQWTRR6U/, November 2023.

[103] Brian J Chen and Jacob Metcalf. Explainer: A Sociotechnical Approach to AI Policy. Policy Brief, Data & Society, May 2024.

[104] Christine Chen, Nicola Dell, and Franziska Roesner. Computer security and privacy in the interactions between victim service providers and human trafficking survivors. In *Proceedings of the 28th USENIX Security Symposium*, 2019.

[105] Gina Masullo Chen, Paromita Pain, Victoria Y Chen, Madlin Mekelburg, Nina Springer, and Franziska Troger. 'you really have to have a thick skin': A cross-cultural perspective on how online harassment influences female journalists. *Journalism*, 21(7):877–895, 2020.

[106] Janet X Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. Trauma-Informed Computing: Towards Safer Technology Experiences for All. In *Proceedings of the USENIX Security Symposium*, 2020.

[107] Jay Chen, Michael Paik, and Kelly McCabe. Exploring Internet Security Perceptions and Practices in Urban Ghana. In *SOUPS*, 2014.

[108] Sapna Cheryan, Victoria C. Plaut, Paul G. Davies, and Claude M. Steele. Ambient belonging: how stereotypical cues impact gender participation in computer science. *J Pers Soc Psychol*, 96:1045–1060, 2009.

[109] Sonia Chiasson, Alain Forget, Elizabeth Stobert, P. C. van Oorschot, and Robert Biddle. Multiple password interference in text passwords and click-based graphical passwords. In *Proc. CCS*, 2009.

[110] Sonia Chiasson, Paul C van Oorschot, and Robert Biddle. A usability study and critique of two password managers. In *Proceedings of the USENIX Security Symposium*, volume 15, pages 1–16, 2006.

[111] Hichang Cho and Anna Filippova. Networked Privacy Management in Facebook: A Mixed-Methods and Multinational Study. In *Proc. CSCW*, 2016.

[112] Danielle Citron and Robert Chesney. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107(6):1753, December 2019.

[113] Danielle Keats Citron. *Hate Crimes in Cyberspace*. Harvard University Press, 2016.

[114] Danielle Keats Citron and Mary Anne Franks. Criminalizing Revenge Porn. *SSRN Scholarly Paper*, (2368946), May 2014.

[115] Elizabeth M Clancy, Megan K Maas, Evita March, Dominika Howard, and Bianca Klettke. Just checking it out? Motivations for and behavioral associations with visiting "slutpages" in the United States and Australia. *Frontiers in Psychology*, 12:671986, 2021.

[116] Heather B. Clayton. Dating Violence, Sexual Violence, and Bullying Victimization Among High School Students — Youth Risk Behavior Survey, United States, 2021. *MMWR Supplements*, 72, 2023.

[117] Camille Cobb and Tadayoshi Kohno. How Public Is My Private Life? Privacy in Online Dating. In *The World Wide Web Conference*, WWW '17, pages 1231–1240, 2017.

[118] Camille Cobb, Lucy Simko, Tadayoshi Kohno, and Alexis Hiniker. A Privacy-Focused Systematic Analysis of Online Status Indicators. *PoPETS*, 2020(3):384–403, 2020.

[119] Andy Cockburn, Carl Gutwin, and Alan Dix. HARK No More: On the Preregistration of CHI Experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[120] Support Ho(s)e Collective. Sex Worker Centered Guide for Academics. https://sxhxcollective.org/wp-content/uploads/2021/03/SxHx-academic-guide-final.pdf, 2021.

[121] Patricia Hill Collins. *Black Feminist Thought: Knowledge, Consciousness and the Politics of Empowerment*. Hyman, 1990.

[122] Patricia Hill Collins. *Fighting Words: Black Women & The Search for Justice*. Uniersity of Minnesota Press, 1998.

[123] Jessica Colnago, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Lorrie Cranor, and Nicolas Christin. "it's not actually that horrible": Exploring adoption of two-factor authentication at a university. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–11. ACM, 2018.

[124] Common Justice. Beyond Offender and Victim. Technical report.

[125] Daniel Le Compte and Daniel Klug. Poster: "It's Viral!" A Study of the Behaviors, Practices, and Motivations of TikTok Users and Social Activism. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '21, pages 108–111, 2021.

[126] Sunny Consolvo, Patrick Gage Kelley, Tara Matthews, Kurt Thomas, Lee Dunn, and Elie Bursztein. "why wouldn't someone think of democracy as a target?": Security practices & challenges of people involved with u.s. political campaigns. In *Proceedings of the 30th USENIX Security Symposium*, 2021.

[127] Dan Conway, Ronnie Taib, Mitch Harris, Kun Yu, Shlomo Berkovsky, and Fang Chen. A qualitative investigation of bank employee experiences of information security and phishing. In *Proc. SOUPS*, 2017.

[128] Kovila P.L. Coopamootoo. Usage patterns of privacy-enhancing technologies. In *Proc. CCS*, 2020.

[129] Kovila P.L. Coopamootoo, Maryam Mehrnezhad, and Ehsan Toreini. "I feel invaded, annoyed, anxious and I may protect myself": Individuals' Feelings about Online Tracking and their Protective Behaviour across Gender and Country. In *Proceedings of the USENIX Security Symposium*, 2022.

[130] Sasha Costanza-Chock. *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press, 2020.

[131] Catherine Courage, Jhilmil Jain, and Stephanie Rosenbaum. Best Practices in Longitudinal Research. In *CHI Extended Abstracts*, 2009.

[132] Benjamin F Crabtree. *Doing qualitative research*. Sage, Thousand Oaks, CA, 1999.

[133] Lorrie Faith Cranor, Adam L. Durity, Abigail Marsh, and Blase Ur. Parents' and Teens' Perspectives on Privacy In a Technology-Filled World. In *Symposium on Usable Privacy and Security*, SOUPS '14, pages 19–35, 2014.

[134] Kate Crawford and Tarleton Gillespie. What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3):410–428, 2016.

[135] Creative Interventions. Creative Interventions Toolkit: A Practical Guide to Stop Interpersonal Violence. Technical report, Creative Interventions, 2012.

[136] Kimberle Crenshaw. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6):1241, July 1991.

[137] Kimberlé W. Crenshaw. *On Intersectionality: Essential Writings*. The New Press, 2017.

[138] Cassandra Cross, Karen Holt, and Thomas J. Holt. To pay or not to pay: An exploratory analysis of sextortion in the context of romance fraud. *Criminology and Criminal Justice*, February 2023.

[139] Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y. Zhao. Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[140] Amy J. C. Cuddy, Susan T. Fiske, and Peter Glick. The BIAS map: behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, 92(4):631–648, 2007.

[141] Amy J. C. Cuddy, Susan T. Fiske, Virginia S. Y. Kwan, Peter Glick, Stéphanie Demoulin, Jacques-Philippe Leyens, Michael Harris Bond, Jean-Claude Croizet, Naomi Ellemers, Ed Sleebos, Tin Tin Htun, Hyun-Jeong Kim, Greg Maio, Judi Perry, Kristina Petkova, Valery Todorov, Rosa Rodríguez-Bailón, Elena Morales, Miguel Moya, Marisol Palacios, Vanessa Smith, Rolando Perez, Jorge Vala, and Rene Ziegler. Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, 48(1):1–33, 2009.

[142] Scott Cunningham. *Causal inference: The mixtape*. Yale University Press, 2021.

[143] Peter Curi and Dana Rebik. Explicit AI photos of Illinois students prompt investigation. *NewsNation*, May 2024.

[144] Alexei Czeskis, Ivayla Dermendjieva, Hussein Yapit, Alan Borning, Batya Friedman, Brian Gill, and Tadayoshi Kohno. Parenting from the Pocket: Value Tensions and Technical Directions for Secure and Private Parent-Teen Mobile Safety. In *Symposium on Usable Privacy and Security*, SOUPS '10, pages 1–15, 2010.

[145] Alaa Daffalla, Lucy Simko, Tadayoshi Kohno, and Alexandru G Bardas. Defensive technology use by political activists during the sudanese revolution. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2021.

[146] DAIP. The Development of the Power and Control and Equality Wheels. Technical report, Domestic Abuse Intervention Programs, March 2017.

[147] Duy Dang-Pham, Siddhi Pittayachawan, and Vince Bruno. Impacts of Security Climate on Employees' Sharing of Security Advice and Troubleshooting: Empirical Networks. *Business Horizons*, 59(6):571–584, 2016.

[148] Duy Dang-Pham, Siddhi Pittayachawan, and Vince Bruno. Why Employees Share Information Security Advice? Exploring the Contributing Factors and Structural Patterns of Security Advice Sharing in the Workplace. *Computers in Human Behavior*, 67:196–206, 2017.

[149] Dart Center. Working with Traumatic Imagery. https://dartcenter.org/content/working-with-traumatic-imagery, August 2014.

[150] Sauvik Das, Laura A. Dabbish, and Jason I. Hong. A Typology of Perceived Triggers for End-User Security and Privacy Behaviors. In *Symposium on Usable Privacy and Security*, SOUPS '19, pages 97–115, 2019.

[151] Sauvik Das, W. Keith Edwards, DeBrae Kennedy-Mayo, Peter Swire, and Yuxi Wu. Privacy for the People? Exploring Collective Action as a Mechanism to Shift Power to Consumers in End-User Privacy. *IEEE Security & Privacy*, 19(5):66–70, September 2021.

[152] Sauvik Das, Joanne Lo, Laura Dabbish, and Jason I. Hong. Breaking! A Typology of Security and Privacy News and How It's Shared. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[153] Jill Davies and Eleanor Lyon. *Domestic violence advocacy: Complex lives/difficult choices*. Sage Publications, 2013.

[154] Simone de Beauvoir. *The Second Sex*. Vintage, 1949.

[155] John P. DeCecco and John P. Elia. A Critique and Synthesis of Biological Essentialism and Social Constructionist Views of Sexuality and Gender. *Journal of Homosexuality*, 24(3-4):1–26, 1993.

[156] Deeptrace Labs. The State of Deepfakes: Landscape, Threats, and Impact. Technical report, September 2019.

[157] Martin Degeling, Christopher Lentzsch, Alexander Nolte, Thomas Herrmann, and Kai-Uwe Loser. Privacy by Socio-Technical Design: A Collaborative Approach for Privacy Friendly System Design. In *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*, pages 502–505, November 2016.

[158] Jasmine DeHart, Kamya Stell, and Christan Grant. Social Media and the Scourge of Visual Privacy. *Information*, 11(2):57, 2020.

[159] Walter S DeKeseredy and Martin D Schwartz. Thinking sociologically about image-based sexual abuse: The contribution of male peer support theory. *Sexualization, Media, & Society*, 2(4), 2016.

[160] John D. DeLamater and Janet Shibley Hyde. Essentialism vs. social constructionism in the study of human sexuality. *The Journal of Sex Research*, 35(1):10–18, 1998.

[161] Jayati Dev, Pablo Moriano, and Jean L. Camp. Lessons Learnt from Comparing WhatsApp Privacy Concerns Across Saudi and Indian Populations. In *Proc. SOUPS*, 2020.

[162] Prema Dev, Jessica Medina, Zainab Agha, Munmun De Choudhury, Afsaneh Razi, and Pamela J. Wisniewski. From Ignoring Strangers' Solicitations to Mutual Sexting with

Friends: Understanding Youth's Online Sexual Risks in Instagram Private Conversations. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW'22 Companion, pages 94–97, New York, NY, USA, November 2022. Association for Computing Machinery.

[163] Michael Devitt. Defending Intrinsic Biological Essentialism. *Philosophy of Science*, 88(1):67–82, 2021.

[164] Rachna Dhamija, J. D. Tygar, and Marti Hearst. Why Phishing Works. In *Proceedings of the 2006 CHI Conference on Human Factors in Computing Systems*, 2006.

[165] Dominic DiFranzo, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D Wherry, and Natalya N Bazarova. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[166] Michele Dillon. *Introduction to sociological theory: Theorists, concepts, and their applicability to the twenty-first century*. John Wiley & Sons, 2020.

[167] 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 Transactions on Computer-Human Interaction (TOCHI)*, 2021.

[168] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, Adrian Weller, and Alexandra Wood. Accountability of AI Under the Law: The Role of Explanation, December 2019.

[169] Sam Dubberley and Michelle Grant. Journalism and Vicarious Trauma: A Guide for Journalists, Editors and News Organizations. Technical report, First Draft, April 2017.

[170] Maeve Duggan. 1 in 4 black americans have faced online harassment because of their race or ethnicity. https://www.pewresearch.org/short-reads/2017/07/25/1-in-4-black-americans-have-faced-online-harassment-because-of-their-race-or-ethnicity/, 2017.

[171] Jared Duval, Ferran Altarriba Bertran, Siying Chen, Melissa Chu, Divya Subramonian, Austin Wang, Geoffrey Xiang, Sri Kurniawan, and Katherine Isbister. Chasing Play on TikTok from Populations with Disabilities to Inspire Playful and Inclusive Technology Design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–15, 2021.

[172] Andrea Dworkin. *Right-Wing Women*. Picador, New York, first picador paperback edition edition, 2025.

[173] Brianna Dym and Casey Fiesler. Ethical and privacy considerations for research using online fandom data. *Transformative Works and Cultures*, 33, 2020.

[174] Alice H. Eagly, Anne E. Beall, and Robert J. Sternberg. *The psychology of gender*. Guilford Press, 2005.

[175] Asia A. Eaton, Holly Jacobs, and Yanet Ruvalcaba. 2017 Nationwide Online Study of Nonconsensual Porn Victimization and Perpetration: A Summary Report. Technical report, Cyber Civil Rights Initiative, June 2017.

[176] Asia A. Eaton, Sofia Noori, Amy Bonomi, Dionne P. Stephens, and Tameka L. Gillum. Nonconsensual Porn as a Form of Intimate Partner Violence: Using the Power and Control Wheel to Understand Nonconsensual Porn Perpetration in Intimate Relationships. *Trauma, Violence, & Abuse*, 22(5):1140–1154, December 2021.

[177] Asia A Eaton, Divya Ramjee, and Jessica F Saunders. The relationship between sextortion during COVID-19 and pre-pandemic intimate partner violence: A large study of victimization among diverse us men and women. *Victims & Offenders*, 18(2):338–355, 2023.

[178] 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 2008 CHI Conference on Human Factors in Computing Systems*, 2008.

[179] Serge Egelman, Marian Harbach, and Eyal Peer. Behavior Ever Follows Intention?: A Validation of the Security Behavior Intentions Scale (SeBIS). In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

[180] Serge Egelman and Eyal Peer. Scaling the Security Wall: Developing a Security Behavior Intenteions Scale (SeBIS). In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*, 2015.

[181] Michael Fagan and Mohammad Maifi Hasan Khan. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 59–75, Denver, CO, June 2016. USENIX Association.

[182] Cori Faklaris, Laura Dabbish, and Jason I. Hong. A Self-Report Measure of End-User Security Attitudes (SA-6). In *Proc. SOUPS*, 2019.

[183] Erika Falk and Kate Kenski. Issue saliency and gender stereotypes: Support for women as presidents in times of war and terrorism. *Social Science Quarterly*, 87(1):1–18, 2006.

[184] Hany Farid. Creating, Using, Misusing, and Detecting Deep Fakes. *Journal of Online Trust and Safety*, 1(4), September 2022.

[185] Florian M Farke, David G Balash, Maximilian Golla, Markus Dürmuth, and Adam J Aviv. Are Privacy Dashboards Good for End Users? Evaluating User Perceptions and Reactions to Google's My Activity. In *Proceedings of the USENIX Security Symposium*, 2021.

[186] Lynn Farrell, Andy Cochrane, and Louise McHugh. Exploring attitudes towards gender and science: The advantages of an IRAP approach versus the IAT. *Journal of Contextual Behavioral Science*, 4(2):121–128, 2015.

[187] Lynn Farrell and Louise McHugh. Examining gender-STEM bias among STEM and non-STEM students using the Implicit Relational Assessment Procedure (IRAP). *Journal of Contextual Behavioral Science*, 6(1):80–90, 2017.

[188] Lynn Farrell and Louise McHugh. Exploring the relationship between implicit and explicit gender-stem bias and behavior among stem students using the implicit relational assessment procedure. *Journal of Contextual Behavioral Science*, 15:142–152, 2020.

[189] Jui-Ying Feng, Yi-Wen Chen, Susan Fetzer, Ming-Chu Feng, and Chiao-Li Lin. Ethical and

legal challenges of mandated child abuse reporters. *Children and Youth Services Review*, 34(1):276–280, January 2012.

[190] Mathieu Fenniak, Matthew Stamy, pubpub zz, Martin Thoma, Matthew Peveler, exiledkingcc, and pypdf Contributors. The pypdf library. https://pypi.org/project/pypdf/.

[191] Jennifer Fereday and Eimear Muir-Cochrane. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *Int. J. of Qualitative Methods*, 5(1):80–92, 2006.

[192] Jessica L. Feuston, Arpita Bhattacharya, Nazanin Andalibi, Elizabeth A. Ankrah, Sheena Erete, Mark Handel, Wendy Moncur, Sarah Vieweg, and Jed R. Brubaker. Researcher Wellbeing and Best Practices in Emotionally Demanding Research. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, pages 1–6, New York, NY, USA, April 2022. Association for Computing Machinery.

[193] Casey Fiesler, Michaelanne Dye, Jessica L. Feuston, Chaya Hiruncharoenvate, C.J. Hutto, Shannon Morrison, Parisa Khanipour Roshan, Umashanthi Pavalanathan, Amy S. Bruckman, Munmun De Choudhury, and Eric Gilbert. What (or Who) Is Public? Privacy Settings and Social Media Content Sharing. In *Proc. CSCW*, 2017.

[194] Casey Fiesler, Michael Zimmer, Nicholas Proferes, Sarah Gilbert, and Naiyan Jones. Remember the Human: A Systematic Review of Ethical Considerations in Reddit Research. *Proceedings of the ACM on Human-Computer Interaction*, 8(GROUP):1–33, February 2024.

[195] Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. A Model of (Often Mixed) Stereotype Content: Competence and Warmth Respectively Follow From Perceived Status and Competition. *Journal of Personality and Social Psychology*, 82(6):898–902, 2002.

[196] Asher Flynn, Elena Cama, Anastasia Powell, and Adrian J Scott. Victim-blaming and image-based sexual abuse. *J. of Criminology*, 56(1):7–25, 2023.

[197] Asher Flynn, Anastasia Powell, Adrian J Scott, and Elena Cama. Deepfakes and Digitally Altered Imagery Abuse: A Cross-Country Exploration of an Emerging form of Image-Based Sexual Abuse. *The British Journal of Criminology*, 62(6):1341–1358, November 2022.

[198] Canadian Centre for Child Protection (C3P). An analysis of financial sextortion victim posts published on r/sextortion. Technical report, C3P, 2022.

[199] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *AAAI International Conference On Web and Social Media*, 2018.

[200] Zak Franklin. Justice for revenge porn victims: Legal theories to overcome claims of civil immunity by operators of revenge porn websites. *Calif. L. Rev.*, 102:1303, 2014.

[201] Diana Freed, Natalie N. Bazarova, Sunny Consolvo, Eunice J Han, Patrick Gage Kelley, Kurt Thomas, and Dan Cosley. Understanding Digital-Safety Experiences of Youth in the U.S. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–15, New York, NY, USA, April 2023. Association for Computing Machinery.

[202] Diana Freed, Sam Havron, Emily Tseng, Andrea Gallardo, Rahul Chatterjee, Thomas Ris-

tenpart, and Nicola Dell. "Is My Phone Hacked?" Analyzing clinical computer security interventions with survivors of intimate partner violence. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[203] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. Digital Technologies and Intimate Partner Violence: A Qualitative Analysis with Multiple Stakeholders. In *CSCW*. ACM, 2017.

[204] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. Digital technologies and intimate partner violence: A qualitative analysis with multiple stakeholders. *PACM: Human-Computer Interaction: Computer-Supported Cooperative Work and Social Computing (CSCW)*, Vol. 1(No. 2):Article 46, 2017.

[205] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. "A Stalker's Paradise": How Intimate Partner Abusers Exploit Technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–13, 2018.

[206] Feminist Frequency. Speak Up & Stay Safe(r). https://onlinesafety.feministfrequency.com/en/.

[207] Andrea Gallardo, Hanseul Kim, Tianying Li, Lujo Bauer, and Lorrie Cranor. Detecting iphone security compromise in simulated stalking scenarios: Strategies and obstacles. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*. USENIX Association, August 2022.

[208] Ellen Garbarino and Michal Strahilevitz. Gender differences in the perceived risk of buying online and the effects of receiving a site recommendation. *Journal of Business Research*, 2004.

[209] Danielle Gaucher, Justin Friesen, and Aaron C. Kay. Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality. *Journal of Personality and Social Psychology*, 101(1):109, 2011.

[210] Timnit Gebru and Émile P. Torres. The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*, April 2024.

[211] Chris Geeng. *Analyzing Usable Security, Privacy, and Safety Through Identity-Based Power Relations*. PhD Thesis, University of Washington, 2022.

[212] 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 *Proceedings of the USENIX Security Symposium*, 2022.

[213] R. Stuart Geiger. Bot-based collective blocklists in twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society*, 19(6):787–803, 2016.

[214] Gendered Innovations in Science, Health & Medicine, Engineering, and Environment. Stereotypes, 2021. https://genderedinnovations.stanford.edu/terms/stereotypes.html.

[215] Jens Gerken. *Longitudinal Research in Human-Computer Interaction*. PhD thesis, Universität Konstanz, 2011.

[216] Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J. LaViola Jr., and Pamela J. Wisniewski. Safety vs. Surveillance: What Children Have to Say about Mobile Apps for Parental Control. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–14, 2018.

[217] Cassidy Gibson, Daniel Olszewski, Natalie Grace Brigham, Anna Crowder, Kevin R. B. Butler, Patrick Traynor, Elissa M. Redmiles, and Tadayoshi Kohno. Analyzing the AI Nudification Application Ecosystem, November 2024.

[218] Eric Gilbert, Karrie Karahalios, and Christian Sandvig. The network in the garden: an empirical analysis of social media in rural life. In *Proceedings of the 2008 CHI Conference on Human Factors in Computing Systems*, 2008.

[219] Meira Gilbert, Miranda Wei, and Lindah Kotut. "tiktok, do your thing"': User reactions to social surveillance in the public sphere. In *Proc. SOUPS*, 2025.

[220] Jeffrey Gleason, Alex Leavitt, and Bridget Daly. In Suspense About Suspensions? The Relative Effectiveness of Suspension Durations on a Popular Social Platform. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–12, New York, NY, USA, April 2025. Association for Computing Machinery.

[221] Peter Glick and Susan T. Fiske. Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of Women Quarterly*, 21(1):119–135, 1997.

[222] Peter Glick and Susan T. Fiske. Ambivalent sexism revisited. *Psychology of Women Quarterly*, 35(3):530–535, 2011.

[223] Matt Goerzen, Elizabeth Anne Watkins, and Gabrielle Lim. Entanglements and Exploits: Sociotechnical Security as an Analytic Framework. 2019.

[224] Maximilian Golla, Grant Ho, Marika Lohmus, Monica Pulluri, and Elissa M. Redmiles. Driving 2FA Adoption at Scale: Optimizing Two-Factor Authentication Notification Design Patterns. In *Proceedings of the USENIX Security Symposium*, 2021.

[225] Maximilian Golla, Miranda Wei, Juliette Hainline, Lydia Filipe, Markus Dürmuth, Elissa Redmiles, and Blase Ur. What was that site doing with my facebook password?: Designing password-reuse notifications. In *Proceedings of the ACM Conference on Computer and Communications Security*, 2018.

[226] David Gonzalez. Laguna Beach HS investigating incident involving AI-generated nude photos of students. https://abc7.com/laguna-beach-high-school-investigating-incident-involving-ai-generated-nude-photos-of-students/14603765/, April 2024.

[227] Google Cloud. PaLM2 for Text. https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/text, 2024.

[228] Marleen Gorissen. It's just a distance thing: Affordances and decisions in online disclosure of sexual violence victimization. *J. of interpersonal violence*, page 08862605241246800, 2024.

[229] Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke, Christian Stransky, Sebastian Möller, Yasemin Acar, and Sascha Fahl. Developers Deserve Security Warnings, too: On the Effect of Integrated Security Advice on Cryptographic API Misuse. In *Proc. SOUPS*, 2018.

[230] Margaret Gratian, Sruthi Bandi, Michel Cukier, Josiah Dykstra, and Amy Ginther. Correlating human traits and cyber security behavior intentions. *Computers & Security*, 73:345–358, 2018. Publisher: Elsevier.

[231] Jeremy A Greene, Niteesh K Choudhry, Elaine Kilabuk, and William H Shrank. Online social networking by patients with diabetes: a qualitative evaluation of communication with facebook. *J. of General Internal Medicine*, 26:287–292, 2011.

[232] Sheila Greene. *Biological Determinism and Essentialism*, chapter 2, pages 13–34. John Wiley & Sons, Ltd, 2020.

[233] Anthony G. Greenwald, Debbie E. McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464, 1998.

[234] Tamy Guberek, Allison McDonald, Sylvia Simioni, Abraham H Mhaidli, Kentaro Toyama, and Florian Schaub. Keeping a low profile? technology, risk and privacy among undocumented immigrants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[235] Katherine Guttmann, Michelle Shouldice, and Alex V. Levin. *Ethical Issues in Child Abuse Research.* Springer International Publishing, Cham, 2019.

[236] Hana Habib, Jessica Colnago, Vidya Gopalakrishnan, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, and Lorrie Faith Cranor. Away From Prying Eyes: Analyzing Usage and Understanding of Private Browsing. In *Proc. SOUPS*, 2018.

[237] Hana Habib, Neil Shah, and Rajan Vaish. Impact of Contextual Factors on Snapchat Public Sharing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[238] Tzipora Halevi, James Lewis, and Nasir Memon. A Pilot Study of Cyber Security and Privacy Related Behavior and Personality Traits. In *Proc. WWW*, 2013.

[239] Julie M Haney and Wayne G Lutters. "It's Scary... It's Confusing... It's Dull": How Cybersecurity Advocates Overcome Negative Perceptions of Security. In *Proc. SOUPS*, pages 411–425, 2018.

[240] Donna Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 1988.

[241] Marian Harbach, Alexander De Luca, Nathan Malkin, and Serge Egelman. Keep on Lockin' in the Free World: A Multi-National Comparison of Smartphone Locking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

[242] Sandra G Harding. *The science question in feminism.* Cornell University Press, 1986.

[243] Eszter Hargittai and Yuli Patrick Hsieh. Succinct survey measures of web-use skills. *Social Science Computer Review*, 30(1):95–107, 2012.

[244] Eszter Hargittai and Eden Litt. New Strategies for Employment? Internet Skills and Online Privacy Practices during People's Job Search. *IEEE Security & Privacy*, 11(3):38–45, 2013.

[245] Sasha Harris-Lovett. In survey, 88% of U.S. adults said they had sexted and 96% of

them endorsed it. https://www.latimes.com/science/sciencenow/la-sci-sn-sexting-sexual-satisfaction-20150807-story.html, 2015.

[246] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. Safe Sexting: The Advice and Support Adolescents Receive from Peers regarding Online Sexual Risks. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1):42:1–42:31, April 2021.

[247] Rakibul Hasan, Bennett I. Bertenthal, Kurt Hugenberg, and Apu Kapadia. Your Photo is so Funny that I don't Mind Violating Your Privacy by Sharing it: Effects of Individual Humor Styles on Online Photo-sharing Behaviors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[248] Ayako A. Hasegawa, Daisuke Inoue, and Mitsuaki Akiyama. How WEIRD is usable privacy and security research? In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 3241–3258, Philadelphia, PA, August 2024. USENIX Association.

[249] Ayako A. Hasegawa, Naomi Yamashita, and Nitsuaki Akiyama. Why They Ignore English Emails: The Challenges of Non-Native Speakers in Identifying Phishing Emails. In *Proc. SOUPS*, 2021.

[250] Josh Haskell. Calabasas teen says classmate not disciplined for sharing real and fake nude images of her. https://abc7.com/calabasas-high-school-student-accuses-classmate-sharing-real-and-fake-nude-photos/14521422/, March 2024.

[251] Caroline Haskins. Florida Middle Schoolers Arrested for Allegedly Creating Deepfake Nudes of Classmates. *Wired*, 2024.

[252] Andrea Havercamp. The Complexity of Nonbinary Gender Inclusion in Engineering Culture. In *2018 ASEE Annual Conference & Exposition*, 2018.

[253] Sam Havron, Diana Freed, Rahul Chatterjee, Damon McCoy, Nicola Dell, and Thomas Ristenpart. Clinical Computer Security for Victims of Intimate Partner Violence. In *Proceedings of the USENIX Security Symposium*, 2019.

[254] Lauren Hawthorne, Shannon K. McCoy, Ellen E. Newell, Amy Blackstone, and Susan K. Gardner. The Role of Sex and Gender Identification in STEM Faculty's Work-Related Stress And Emotional Well-Being. *Journal of Women and Minorities in Science and Engineering*, 24(4):325–337, 2018.

[255] Weijia He, Maximilian Golla, Roshni Padhi, Jordan Ofek, Markus Dürmuth, Earlence Fernandes, and Blase Ur. Rethinking Access Control and Authentication for the Home Internet of Things (IoT). In *Proceedings of the USENIX Security Symposium*, pages 255–272, 2018.

[256] Jon Healey. Beverly Hills school district expels 8th graders involved in fake nude scandal. https://www.latimes.com/california/story/2024-03-07/beverly-hills-school-district-expels-8th-graders-involved-in-fake-nude-scandal, March 2024.

[257] Nicola Henry and Asher Flynn. Image-based sexual abuse: Online distribution channels and illicit communities of support. *Violence against women*, 25(16):1932–1955, 2019.

[258] Nicola Henry, Asher Flynn, and Anastasia Powell. Policing image-based sexual abuse: Stakeholder perspectives. *Police practice and research*, 19(6):565–581, 2018.

[259] Nicola Henry, Asher Flynn, and Anastasia Powell. Technology-facilitated domestic and sexual violence: A review. *Violence against women*, 26(15-16):1828–1854, 2020.

[260] Nicola Henry, Clare McGlynn, Asher Flynn, Kelly Johnson, Anastasia Powell, and Adrian J. Scott. *Image-Based Sexual Abuse: A Study on the Causes and Consequences of Non-Consensual Nude or Sexual Imagery*. Routledge, Abingdon, Oxon, 2022.

[261] Nicola Henry and Anastasia Powell. Beyond the 'sext': Technology-facilitated sexual violence and harassment against adult women. *Australian & New Zealand Journal of Criminology*, 48(1):104–118, March 2015.

[262] Nicola Henry and Anastasia Powell. Technology-facilitated sexual violence: A literature review of empirical research. *Trauma, Violence, & Abuse*, 19(2):195–208, 2018.

[263] Nicola Henry, Anastasia Powell, and Asher Flynn. AI can now create fake porn, making revenge porn even more complicated. https://theconversation.com/ai-can-now-create-fake-porn-making-revenge-porn-even-more-complicated-92267, March 2018.

[264] Nicola Henry and Rebecca Umbach. Sextortion: Prevalence and correlates in 10 countries. *Computers in Human Behavior*, 158:108298, September 2024.

[265] Cormac Herley. So long, and no thanks for the externalities: The rational rejection of security advice by users. In *New Security Paradigms Workshop (NSPW)*, 2009.

[266] Cormac Herley. More is not the answer. *IEEE Security and Privacy*, January 2014.

[267] Cormac Herley. Unfalsifiability of security claims. *Proceedings of the National Academy of Sciences*, 113(23):6415–6420, 2016.

[268] Alex Hern. Reddit bans 'deepfakes' face-swap porn community. https://www.theguardian.com/technology/2018/feb/08/reddit-bans-deepfakes-face-swap-porn-community, February 2018.

[269] Security Hero. 2023 State Of Deepfakes: Realities, Threats, And Impact. Technical report, Security Hero.

[270] Rosanna Hertz. *Reflexivity and Voice*. SAGE, 1997.

[271] Paul Hitlin. Research in the Crowdsourcing Age, a Case Study. Technical report, Pew Research Center, July 2016.

[272] Curt Hoffman and Nancy Hurst. Gender stereotypes: Perception or rationalization? *Journal of Personality and Social Psychology*, 58(2):197–208, 1990.

[273] Anna Lauren Hoffmann. Terms of inclusion: Data, discourse, violence. *New Media & Society*, 23(12):3539–3556, 2021.

[274] Andrew Gary Darwin Holmes. Researcher positionality - a consideration of its influence and place in qualitative research - a new researcher guide. *International Journal of Education*, 8(4):1–10, 2020.

[275] Karen Holtzblatt and Hugh Beyer. *Contextual Design: Evolved*, chapter Consolidation and Ideation: The Bridge to Design, pages 21–52. Morgan & Claypool Publishers, 2014.

[276] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. Do platform migrations compromise

content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–24, 2021.

[277] Hayward Derrick Horton. Critical demography: The paradigm of the future? *Sociological Forum*, 14:363–367, 1999.

[278] Roberto Hoyle, Srijita Das, Apu Kapadia, Adam J. Lee, and Kami Vaniea. Viewing the Viewers: Publishers' Desires and Viewers' Privacy Concerns in Social Networks. In *Proc. CSCW*, 2017.

[279] Antoinette Huber. 'A shadow of me old self': The impact of image-based sexual abuse in a digital society. *Int. Rev. of Victimology*, 29(2):199–216, 2023.

[280] Antoinette Raffaela Huber. Image-based sexual abuse: Online communities and the broader misogynistic landscape. *British J. of Criminology*, 63(4):967–983, 2023.

[281] Jina Huh-Yoo, Afsaneh Razi, Diep N. Nguyen, Sampada Regmi, and Pamela J. Wisniewski. "Help Me:" Examining Youth's Private Pleas for Support and the Responses Received from Peers via Instagram Direct Messages. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–14, New York, NY, USA, April 2023. Association for Computing Machinery.

[282] Netta Iivari, Leena Ventä-Olkkonen, Sumita Sharma, Tonja Molin-Juustila, and Essi Kinnunen. Chi against bullying: Taking stock of the past and envisioning the future. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[283] Aleksandar Ilievski and Bernik Igor. Social-Economic Aspects of Cybercrime. *Innovative Issues and Approaches in Social Sciences*, 9(3), September 2016.

[284] Jane Im, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, and Eric Gilbert David Jurgens and. Still out there: Modeling and identifying russian troll accounts on twitter. *WebSci*, 2020.

[285] Jane Im, Sarita Schoenebeck, Gabriel Grill Marilyn Iriarte, Daricia Wilkinson, Amna Batool, Rahaf Alharbi, Audrey N. Funwie, Tergel Gankhuu, Eric Gilbert, and Mustafa Naseem. Women's perspectives on harm and justice after online harassment. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):32, 2018.

[286] Amnesty International. Troll patrol. Technical report, 2018.

[287] Privacy International. No body's business but mine: How menstruation apps are sharing your data. Technical report, 2019.

[288] Michael Inzlicht, Alexa M. Tullett, and Jennifer N. Gutsell. Stereotype threat spillover: The short- and long-term effects of coping with threats to social identity. In *Stereotype threat: Theory, process, and application*, pages 107–123. Sage Publications Ltd., 2011.

[289] Iulia Ion, Rob Reeder, and Sunny Consolvo. "...No one Can Hack My Mind": Comparing Expert and Non-Expert Security Practices. In *SOUPS*. USENIX Association, 2015.

[290] Mansoor Iqbal. TikTok Revenue and Usage Statistics. https://www.businessofapps.com/data/tik-tok-statistics/, 2022.

[291] L. Iyadurai, S. E. Blackwell, R. Meiser-Stedman, P. C. Watson, M. B. Bonsall, J. R. Geddes, A. C. Nobre, and E. A. Holmes. Preventing intrusive memories after trauma via a brief intervention involving Tetris computer game play in the emergency department: A proof-of-concept randomized controlled trial. *Molecular Psychiatry*, 23(3):674–682, March 2018.

[292] Catherine Jennifer, Fatemeh Tahmasbi, Jeremy Blackburn, Gianluca Stringhini, Savvas Zannettou, and Emiliano De Cristofaro. Feels bad man: Dissecting automated hateful meme detection through the lens of facebook's challenge. *CySoc*, 2022.

[293] Sarah Jeong. The internet of garbage. https://www.theverge.com/2018/8/28/17777330/internet-of-garbage-book-sarah-jeong-online-harassment, August 2018.

[294] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. In *Proceedings of the ACM Transactions on Computer-Human Interaction*, 2018.

[295] Haiyan Jia, Pamela J. Wisniewski, Heng Xu, Mary Beth Rosson, and John M. Carroll. Risk-taking as a Learning Process for Shaping Teen's Online Information Privacy Behaviors. In *Proc. CSCW*, 2015.

[296] Jigsaw. Designing for an at-risk world. https://medium.com/jigsaw/designing-for-an-at-risk-world-75081c6fa061, 2016.

[297] Karin Johnson. Butler County teen pushing for change after photo was manipulated into AI-generated nude pic. https://www.wlwt.com/article/butler-county-ai-generation-nude-picture-teenager-change/61691002, July 2024.

[298] Maritza Johnson, Serge Egelman, and Steven M Bellovin. Facebook and privacy: it's complicated. In *Proc. SOUPS*, pages 1–15, 2012.

[299] Edward E Jones and Richard E Nisbett. The actor and the observer: Divergent perceptions of the causes of behavior. In *Perceiveing the causes of behavior*. Lawrence Erlbaum Associates, Inc., 1987.

[300] John T. Jost and Jojanneke van der Toorn. System justification theory. In *Handbook of Theories of Social Psychology*, pages 313–343. Sage Publications Ltd., 2012.

[301] Mariame. Kaba, Naomi Murakawa, and Tamara K. Nopper. *We do this 'til we free us : abolitionist organizing and transforming justice*. The Abolitionist papers series. Haymarket Books, Chicago, Illinois, 2021 - 2021.

[302] Mudasir Kamal and William J Newman. Revenge pornography: Mental health implications and related legislation. *J. of the American Academy of Psychiatry and the Law Online*, 44(3):359–367, 2016.

[303] Nur Shazwani Kamarudin, Vineeth Rakesh, Ghazaleh Beigi, Lydia Manikouda, and Huan Liu. A Study of Reddit-User's Response to Rape. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 591–592, August 2018.

[304] Minna Kamppuri, Roman Bednarik, and Markku Tukiainen. The Expanding Focus of HCI: Case Culture. In *Proc. NordiCHI*, 2006.

[305] Naveena Karusala, Apoorva Bhalla, and Neha Kumar. Privacy, Patriarchy, and Participation on Social Media. In *Proc. DIS*, 2019.

[306] Mannat Kaur, Michel van Eeten, Marijn Janssen, Kevin Borgolte, and Tobias Fiebig. Human factors in security research: Lessons learned from 2008-2018.

[307] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*, 2015.

[308] D. Bondy Valdovinos Kaye, Aleesha Rodriguez, Katrin Langton, and Patrik Wikstrom. You Made This? I Made This: Practices of Authorship and (Mis) Attribution on TikTok. *International Journal of Communication*, 15:3195–3215, 2021.

[309] Joseph 'Jofish' Kaye. Self-reported password sharing strategies. In *Proceedings of the 2011 CHI Conference on Human Factors in Computing Systems*, 2011.

[310] Os Keyes. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *CSCW*, 2018.

[311] Os Keyes, Chandler May, and Annabelle Carrell. You Keep Using That Word: Ways of Thinking about Gender in Computing Research. *CSCW*, 2021.

[312] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J Wisniewski, and Munmun De Choudhury. A human-centered systematic literature review of cyberbullying detection algorithms. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–34, 2021.

[313] Anne M Koenig and Alice H Eagly. Evidence for the social role theory of stereotype content: observations of groups' roles shape stereotypes. *Journal of personality and social psychology*, 107(3):371, 2014.

[314] Marlene Kollmayer, Barbara Schober, and Christiane Spiel. Gender stereotypes in education: Development, consequences, and interventions. *European Journal of Developmental Psychology*, 15(4):361–377, 2018.

[315] Nikolaos Koukopoulos, Madeleine Janickyj, and Leonie Maria Tanczer. Defining and Conceptualizing Technology-Facilitated Abuse ("Tech Abuse"): Findings of a Global Delphi Study. *Journal of Interpersonal Violence*, page 08862605241310465, January 2025.

[316] Rachel Kowert. Dark participation in games. *Frontiers in Psychology*, 11, 2020.

[317] Kristen Bialik Kristi Walker and Patrick van Kessel. Strong men, caring women: How Americans describe what society values (and doesn't in each gender). https://www.pewresearch.org/social-trends/interactives/strong-men-caring-women/, 2018.

[318] Ivar Krumpal. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity*, 47(4):2025–2047, 2013.

[319] John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[320] Matthew B. Kugler and Carly Pace. Deepfake Privacy: Attitudes and Regulation. *SSRN Electronic Journal*, 2021.

[321] Neha Kumar and Naveena Karusala. Braving Citational Justice in Human-Computer Interaction. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing*

Systems, pages 1–9, Yokohama Japan, May 2021. ACM.

[322] Smita Kumar and Liz Cavallaro. Researcher Self-Care in Emotionally Demanding Research: A Proposed Conceptual Framework. *Qualitative Health Research*, 28(4):648–658, March 2018.

[323] 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 *Proc. SOUPS*, 2009.

[324] Lynette Kvasny, KD Joshi, and Eileen Trauth. The influence of self-efficacy, gender stereotypes and the importance of IT skills on college students' intentions to pursue IT careers. In *Proc. iConference*, pages 508–513. 2011.

[325] Jennifer Laffier and Aalyia Rehman. Deepfakes and Harm to Women. *Journal of Digital Life and Learning*, 3(1):1–21, June 2023.

[326] Elizabeth Laird, Maddy Dwyer, and Kristin Woelfel. In Deep Trouble: Surfacing Tech-Powered Sexual Harassment in K-12 Schools. Technical report, CDT, September 2024.

[327] Santiago Lakatos. AI-Generated 'Undressing' Images Move from Niche Pornography Discussion Forums to a Scaled and Monetized Online Business. Technical report, Graphika, December 2023.

[328] Aviv Y Landau, Susi Ferrarello, Ashley Blanchard, Kenrick Cato, Nia Atkins, Stephanie Salazar, Desmond U Patton, and Maxim Topaz. Developing machine learning-based models to help identify child abuse and neglect: Key ethical challenges and recommended solutions. *Journal of the American Medical Informatics Association*, 29(3):576–580, March 2022.

[329] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction*, 2, November 2018.

[330] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research Methods in Human-Computer Interaction*. Morgan Kaufmann, 2017.

[331] Cambell Leaper and Christine R. Starr. Helping and Hindering Undergraduate Women's STEM Motivation: Experiences With STEM Encouragement, STEM-Related Gender Bias, and Sexual Harassment. *Psychology of Women Quarterly*, 43(2):165–183, 2018.

[332] Tama Leaver. Intimate Surveillance: Normalizing Parental Monitoring and Mediation of Infants Online. *Social Media & Society*, 3(2), 2017.

[333] Jooyoung Lee, Sarah Rajtmajer, Eesha Srivatsavaya, and Shomir Wilson. Digital Inequality Through the Lens of Self-Disclosure. In *Proc. PETS*, 2021.

[334] Amanda Lenhart, Michelle Ybarra, and Myeshia Price-Feeney. Nonconsensual image sharing. Technical report, 2016.

[335] Pedro Leon, Blase Ur, Richard Shay, Yang Wang, Rebecca Balebako, and Lorrie Cranor. Why Johnny can't opt out: A usability evaluation of tools to limit online behavioral advertising. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 589–598, Austin Texas USA, May 2012. ACM.

[336] Karen Levy and Bruce Schneier. Privacy Threats in Intimate Relationships. *Journal of Cybersecurity*, pages 1–13, 2020.

[337] Karen E.C. Levy. Intimate Surveillance. *Idaho Law Review*, 51:679, 2014.

[338] Colleen M. Lewis, Niral Shah, and Katrina Falkner. Equity and diversity. In Sally A. Fincher and Anthony V.Editors Robins, editors, *The Cambridge Handbook of Computing Education Research*, Cambridge Handbooks in Psychology, pages 481–510. Cambridge University Press, 2019.

[339] Ruth Lewis and Sundari Anitha. Upskirting: A systematic literature review. *Trauma, Violence, & Abuse*, 24(3):2003–2018, 2023.

[340] Tom De Leyn, Ralf De Wolf, Mariek Vanden Abeele, and De Lieven Marez. In-Between Child's Play and Teenage Pop Culture: Tweens, TikTok & Privacy. *Journal of Youth Studies*, 25(8):1108–1125, 2022.

[341] Hanlin Li, Brent Hecht, and Stevie Chancellor. All That's Happening behind the Scenes: Putting the Spotlight on Volunteer Moderator Labor in Reddit. *Proc. of the Int. AAAI Conference on Web and Social Media*, 16:584–595, May 2022.

[342] Yachao Li, Mengfei Guan, Paige Hammond, and Lane E. Berrey. Communicating COVID-19 Information on TikTok: A Content Analysis of TikTok Videos From Official Accounts Featured in the COVID-19 Information Hub. *Health education research*, 36(3):261–271, 2021.

[343] Calvin Liang, Sean A. Munson, and Julie A. Kientz. Embracing Four Tensions in Human-Computer Interaction Research with Marginalized People. *TOCHI*, 28(2), 2021.

[344] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. How WEIRD is CHI? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[345] Richard A Lippa, Kathleen Preston, and John Penner. Women's representation in 60 occupations from 1972 to 2010: More women in high-status jobs, few women in things-oriented jobs. *PloS one*, 9(5):e95960, 2014.

[346] Laura van Dernoot Lipsky and Connie Burk. *Trauma Stewardship: An Everyday Guide to Caring for Self While Caring for Others*. Berrett-Koehler Publishers, San Francisco, 2009.

[347] Tiffany Llou. 'I don't want to live in fear anymore': North Texas girl victimized with deepfake nudes pushes for federal law. https://www.wfaa.com/article/news/local/north-texas-girl-victimized-with-deepfake-nudes-pushes-for-federal-law/287-7cf49849-1b27-4a21-864c-d69cf1644e9c, June 2024.

[348] Johnny Long. *No Tech Hacking: A Guide to Social Engineering, Dumpster Diving, and Shoulder Surfing*. Syngress, 2008.

[349] Antonio López Martínez, Manuel Gil Pérez, and Antonio Ruiz-Martínez. A Comprehensive Review of the State-of-the-Art on Security and Privacy Issues in Healthcare. *ACM Comput. Surv.*, 55(12):249:1–249:38, March 2023.

[350] B.L. Love. *We Want to Do More Than Survive: Abolitionist Teaching and the Pursuit of Educational Freedom*. Beacon Press, 2019.

[351] Deborah Lupton and Ben Williamson. The Datafied Child: The Dataveillance of Children and Implications For Their Rights. *New Media & Society*, 19(5):780–794, 2017.

[352] Kim Lyons. Tiktok says it has passed 1 billion users. https://www.theverge.com/2021/9/27/22696281/tiktok-1-billion-users, 2021.

[353] Megan K Maas, Kyla M Cary, Elizabeth M Clancy, Bianca Klettke, Heather L McCauley, and Jeff R Temple. Slutpage use among us college students: the secret and social platforms of image-based sexual abuse. *Archives of Sexual Behavior*, 50:2203–2214, 2021.

[354] Dominique Machuletz, Stefan Laube, and Rainer Böhme. Webcam Covering as Planned Behavior. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[355] David MacPhee, Samantha Farro, and Silvia Sara Canetto. Academic Self-Efficacy and Performance of Underrepresented STEM Majors: Gender, Ethnic, and Social Class Patterns. *Analyses of Social Issues and Public Policy*, 13(1):347–369, 2013.

[356] Mary Madden. Privacy management on social media sites. *Pew Internet Report*, 24:1–20, 2012.

[357] Mary Madden. Privacy, security, and digital inequality. https://datasociety.net/wp-content/uploads/2017/09/DataAndSociety_PrivacySecurityandDigitalInequality.pdf, 2017.

[358] Sophie Maddocks. From Non-consensual Pornography to Image-based Sexual Abuse: Charting the Course of a Problem with Many Names. *Australian Feminist Studies*, 33(97):345–361, July 2018.

[359] Kaitlin Mahar, Amy X. Zhang, and David Karger. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 586:1–586:13. ACM, 2018.

[360] William R. Marczak and Vern Paxson. Social Engineering Attacks on Government Opponents: Target Perspectives. *Proceedings on Privacy Enhancing Technologies*, 2017(2):172–185, April 2017.

[361] William R Marczak, John Scott-Railton, Morgan Marquis-Boire, and Vern Paxson. When governments hack opponents: a look at actors and technology. In *Proceedings of the USENIX Security Symposium*, 2014.

[362] Alison J. Marganski, Lisa A. Melander, and Walter S. DeKeseredy. Single, repeat, and poly intimate partner violence victimization among women at a college campus: Extending research through the inclusion of technology-facilitated violence and examining key social determinants for intimate partner violence prevention. *Violence Against Women*, 28(12-13):3013–3036, 2022.

[363] Annette Markham. Fabrication as ethical practice: Qualitative inquiry in ambiguous internet contexts. *Information, Communication & Society*, 15(3):334–353, 2012.

[364] David F Marks and Lucy Yardley. *Research methods for clinical and health psychology*. Sage, 2004.

[365] Alice Marwick. Privacy Without Power: What Privacy Research Can Learn from Surveillance Studies. *Surveillance & Society*, 20(4):397–405, December 2022.

[366] Alice E Marwick, Lindsay Blackwell, and Katherine Lo. Best Practices for Conducting Risky Research and Protecting Yourself from Online Harassment. Technical report, Data & Society, New York, 2016.

[367] Allison Master, Sapna Cheryan, and Andrew N. Meltzoff. Computing whether she belongs: Stereotypes undermine girls' interest and sense of belonging in computer science. *Journal of Educational Psychology*, 108(3), 2016.

[368] Allison Master, Sapna Cheryan, Adriana Moscatelli, and Andrew N. Meltzoff. Programming experience promotes higher stem motivation among first-grade girls. *Journal of Experimental Child Psychology*, 160:92–105, 2017.

[369] Allison Master, Andrew N. Meltzoff, and Sapna Cheryan. Gender stereotypes about interests start early and cause gender disparities in computer science and engineering. *PNAS*, 118(48), 2021.

[370] Arunesh Mathur and Marshini Chetty. Impact of User Characteristics on Attitudes Towards Automatic Mobile Application Updates. In *Proc. SOUPS*, 2017.

[371] Arunesh Mathur, Jessica Vitak, Arvind Narayanan, and Marshini Chetty. Characterizing the Use of Browser-Based Blocking Extensions To Prevent Online Tracking. In *Proc. SOUPS*, 2018.

[372] J. Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. Reporting, reviewing, and responding to harassment on twitter, 2015.

[373] Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F. Churchill, and Sunny Consolvo. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2189–2201. ACM, 2017.

[374] Scott E Maxwell, Michael Y Lau, and George S Howard. Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, 70(6):487, 2015.

[375] Michelle L Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur. Measuring Password Guessability for an Entire University. In *Proc. CCS*, 2013.

[376] Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur. Measuring Password Guessability for an Entire University. In *Proc. CCS*, 2013.

[377] Kelly S McClure. Identifying and defining the constructs and variables to measure. In *Selecting and Describing Your Research Instruments*. American Psychological Association, 2020.

[378] Anthony McCosker. Making sense of deepfakes: Socializing AI and building data literacy on GitHub and YouTube. *New Media & Society*, 26(5):2786–2803, 2022.

[379] Allison McDonald, Catherine Barwulor, Michelle L. Mazurek, Florian Schaub, and Elissa M. Redmiles. "it's stressful having all these phones": Investigating sex workers' safety goals, risks, and practices online. In *Proceedings of the 30th USENIX Security Symposium*, 2021.

[380] Clare McGlynn and Kelly Johnson. *Cyberflashing: Recognising harms, reforming laws*. Policy Press, 2021.

[381] Clare McGlynn, Kelly Johnson, Erika Rackley, Nicola Henry, Nicola Gavey, Asher Flynn, and Anastasia Powell. 'It's Torture for the Soul': The Harms of Image-Based Sexual Abuse. *Social & Legal Studies*, 30(4):541–562, August 2021.

[382] Clare McGlynn and Erika Rackley. Image-based sexual abuse. *Oxford Journal of Legal Studies*, 37(3):534–561, 2017.

[383] Clare McGlynn, Erika Rackley, and Ruth Houghton. Beyond 'revenge porn': The continuum of image-based sexual abuse. *Feminist legal studies*, 25:25–46, 2017.

[384] Clare McGlynn, Erika Rackley, Kelly Johnson, Nicola Henry, Asher Flynn, Anastasia Powell, Nicola Gavey, and Adrian Scott. Shattering Lives and Myths: A Report on Image-Based Sexual Abuse. Technical report, July 2019.

[385] Susan McGregor and Elizabeth Anne Watkins. 'security by obscurity': Journalists' mental models of information security. *International Symposium on Online Journalism*, 6, 2016.

[386] Susan E. McGregor, Polina Charters, Tobin Holliday, and Franziska Roesner. Investigating the computer security practices and needs of journalists. In *Proceedings of the 24th USENIX Security Symposium*, 2015.

[387] Susan E. McGregor, Franziska Roesner, and Kelly Caine. Individual versus organizational computer security and privacy concerns in journalism. In *Proceedings on Privacy Enhancing Technologies*, 2016.

[388] Susan E. McGregor, Elizabeth Anne Watkins, Mahdi Nasrullah Al-Ameen, Kelly Caine, and Franziska Roesner. When the weakest link is strong: secure collaboration in the case of the panama papers. In *Proceedings of the 26th USENIX Conference on Security Symposium*, SEC'17, pages 505–522, USA, August 2017. USENIX Association.

[389] Maryam Mehrnezhad and Teresa Almeida. Caring for intimate data in fertility technologies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[390] Michael A. Messner, Margaret Carlisle Duncan, and Kerry Jensen. Separating The Men From The Girls: The Gendered Language of Televised Sports. *Gender & Society*, 7(1):121–137, 1993.

[391] Microsoft. Civility, safety, and interaction online. https://www.microsoft.com/en-us/digital-skills/digital-civility, 2019.

[392] Kimberly J. Mitchell, Anna Segura, Lisa M. Jones, and Heather A. Turner. Poly-victimization and peer harassment involvement in a technological world. *J. of Interpersonal Violence*, 33(5):762–788, 2018.

[393] Wendy Moncur. The emotional wellbeing of researchers: Considerations for practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1883–1890, New York, NY, USA, April 2013. Association for Computing Machinery.

[394] E. R. Mondschein, K. E. Adolph, and C. S. Tamis-LeMonda. Gender bias in mothers' expectations about infant crawling. *J Exp Child Psychol*, 77(4):304–316, 2000.

[395] David Moore, Geoffrey M. Voelker, and Stefan Savage. Inferring Internet denial of service activity. In *Proceedings of the USENIX Security Symposium*, 2001.

[396] Rosetta Moors and Ruth Webber. The dance of disclosure: Online self-disclosure of sexual assault. *Qualitative Social Work: Research and Practice*, 12(6):799–815, 2013.

[397] David L. Morgan and Andreea Nica. Iterative Thematic Inquiry: A New Method For Analyzing Qualitative Data. *International Journal of Qualitative Methods*, 19, 2020.

[398] Rachel E. Morgan and Jennifer L. Truman. Stalking Victimization. Technical Report NCJ 301735, U.S. Department of Justice, Office of Justice Programs Bureau of Justice Statistics, February 2022.

[399] Michael Morris. Standard white: Dismantling white normativity. *California Law Review*, 2016.

[400] Janice M. Morse. "Perfectly Healthy, But Dead": The Myth of Inter-Rater Reliability. *Qualitative Health Research*, 7(4):445–447, November 1997.

[401] Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479, 2012.

[402] Corinne A. Moss-Racusin, Aneta K. Molenda, and Charlotte R. Cramer. Can Evidence Impact Attitudes? Public Reactions to Evidence of Gender Bias in STEM Fields. *Psychology of Women Quarterly*, 39(2):194–209, 2015.

[403] Multistate. Deepfakes in Elections. https://www.multistate.ai/deepfakes-sexual.

[404] Collins W. Munyendo, Peter Mayer, and Adam J. Aviv. "I just stopped using one and started using the other": Motivations, Techniques, and Challenges When Switching Password Managers. In *Proc. CCS*, 2023.

[405] Mary C. Murphy, Claude M. Steele, and James J. Gross. Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychological Science*, 19(10):879–885, 2007.

[406] Arvind Narayanan. The limits of the quantitative approach to discrimination, October 2022.

[407] NCMEC. Take it down. https://takeitdown.ncmec.org/, 2024.

[408] Anton J Nederhof. Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, 15(3):263–280, 1985.

[409] James Nelson. Using conceptual depth criteria: Addressing the challenge of reaching saturation in qualitative research. *Qualitative Research*, 17(5):554–570, October 2017.

[410] Mark W. Newman, Debra Lauterbach, Sean A. Munson, Paul Resnick, and Margaret E. Morris. It's not that I don't have problems, I'm just not putting them on Facebook: Challenges and opportunities in using online social networks for health. In *Proc. of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, pages 341–350, 2011.

[411] James Nicholson, Lynne Coventry, and Pamela Briggs. "If It's Important It Will Be A Headline": Cybersecurity Information Seeking in Older Adults. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019.

[412] Helen Nissenbaum, Batya Friedman, and Edward Felten. Computer Security: Competing Concepts, October 2001. arXiv:cs/0110001.

[413] Borke Obada-Obieh, Yue Huang, Lucrezia Spagnolo, and Konstantin Beznosov. Sok: The dual nature of technology in sexual abuse. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, pages 2320–2343, 2022.

[414] Federal Bureau of Investigation. Child Sexual Abuse Material Created by Generative AI and Similar Online Tools is Illegal. https://www.ic3.gov/Media/Y2024/PSA240329.

[415] U.S. House of Representatives. 18 U.S. Code § 2256.

[416] Executive Board of the American Anthropological Association. AAA Statement on Race. https://www.americananthro.org/ConnectWithAAA/Content.aspx?ItemNumber=2583.

[417] United Nations Human Rights Office of the High Commissioner. Gender stereotyping. https://www.ohchr.org/en/issues/women/wrgs/pages/genderstereotypes.aspx, 2021.

[418] Kate O'Flaherty. Apple AirTags get important anti-stalking upgrade. https://www.forbes.com/sites/kateoflahertyuk/2022/12/22/apple-airtags-get-important-anti-stalking-upgrade/, 2022.

[419] Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–29, Yokohama Japan, April 2025. ACM.

[420] Chidera Okolie. Artificial intelligence-altered videos (deepfakes), image-based sexual abuse, and data privacy concerns. *J. of Int. Women's Studies*, 25(2):11, 2023.

[421] Alannah Oleson, Benjamin Xie, Jean Salac, Jayne Everson, F Megumi Kivuva, and Amy J Ko. A Decade of Demographics in Computing Education Research: A Critical Review of Trends in Collection, Reporting, and Use. In *Proc. ICER*, 2022.

[422] Kenneth Olmstead and Aaron Smith. What the public knows about cybersecurity. *Pew Research Center*, 22, 2017.

[423] Roberta Liggett O'Malley. Short-Term and Long-Term Impacts of Financial Sextortion on Victim's Mental Well-Being. *Journal of Interpersonal Violence*, 38(13-14):8563–8592, July 2023.

[424] Roberta Liggett O'Malley and Karen M Holt. Cyber sextortion: An exploratory analysis of different perpetrators engaging in a similar crime. *J. of Interpersonal Violence*, 37(1-2):258–283, 2022.

[425] Kaan Onarlioglu, Utku Ozan Yilmaz, Engin Kirda, and Davide Balzarotti. Insights into User Behavior in Dealing with Internet Attacks. In *NDSS*, 2012.

[426] Tully O'Neill. 'Today I Speak': Exploring How Victim-Survivors Use Reddit. *Int. J. for Crime, Justice & Soc. Democracy*, 7(1):44–59, March 2018.

[427] Isabelle Oomen and Ronald Leenes. Privacy risk perceptions and privacy protection strategies. *Policies and research in identity management*, pages 121–138, 2008.

[428] Martin T Orne. On the social psychology of the psychological experiment: With particular

reference to demand characteristics and their implications. *American psychologist*, 17(11):776, 1962.

[429] Arnisson Andre C Ortega. Toward critical demography 2.0. *Human Geography*, 2023.

[430] Jahna Otterbacher, Jo Bates, and Paul Clough. Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[431] Kentrell Owens, Yael Eiger, Basia Radka, Tadayoshi Kohno, and Franziska Roesner. Understanding experiences with compulsory immigration surveillance in the U.S. In *ACM Conference on Fairness, Accountability, and Transparency*, Athens, Greece, June 2025.

[432] Kentrell Owens, Franziska Roesner, and Tadayoshi Kohno. Electronic Monitoring Smartphone Apps: An Analysis of Risks from Technical, Human-Centered, and Legal Perspectives. In *Proceedings of the 31st USENIX Security Symposium*, pages 4077–4, Boston, MA, USA, August 2022. USENIX.

[433] Stefan Palan and Christian Schitte. Prolific.ac — A Subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[434] Christina Pan, Sahil Yakhmi, Tara Iyer, Evan Strasnick, Amy Zhang, and Michael Bernstein. Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proc. ACM Hum.-Comput. Interact.*, (CSCW), October 2022.

[435] Cheul Young Park, Cori Faklaris, Siyan Zhao, Alex Sciuto, Laura Dabbish, and Jason Hong. Share and Share Alike? An Exploration of Secure Behaviors in Romantic Relationships. In *Symposium on Usable Privacy and Security*, SOUPS '18, 2018.

[436] Yong Jin Park. Do Men and Women Differ in Privacy? Gendered Privacy and (In)equality in the Internet. *Computers in Human Behavior*, 50:242–248, 2015.

[437] Kim Parker, Juliana Horowitz, and Renee Stepler. On gender differences, no consensus on nature vs. nuture. Technical report, Pew Research Center, December 2017.

[438] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*, 2016.

[439] Patricia Davis. Financial sextortion: "a growing crisis". https://www.missingkids.org/blog/2023/financial-sextortion-growing-crisis, 2023.

[440] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 2009.

[441] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[442] Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. Let's Go in for a Closer Look: Observing Passwords in Their Natural Habitat. In *CCS*, CCS, pages 295–310, New York, NY, USA, 2017. Association for Computing Machinery. event-place: Dallas, Texas, USA.

[443] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Why people (don't) use password managers effectively. In *SOUPS*, 2019.

[444] Eyal Peer, David Rothschild, Zak Evernden, Andrew Gordon, and Ekaterina Damer. Mturk, prolific or panels? choosing the right audience for online research. *SSRN Electronic Journal*, 01 2021.

[445] Alan Peshkin. The Goodness of Qualitative Research. *Educational Researcher*, 22(2):23–29, March 1993.

[446] PEW Research Center. Online harassment 2017. https://www.pewinternet.org/2017/07/11/online-harassment-2017/, 2017.

[447] Katharina Pfeffer, Alexandra Mai, Edgar Weippl, Emilee Rader, and Katharina Krombholz. Replication: Stories as informal lessons about security. In *Eighteenth Symposium on Usable Privacy and Security*, 2022.

[448] Riana Pfefferkorn. Addressing Computer-Generated Child Sex Abuse Imagery: Legal Framework and Policy Implications. https://www.lawfaremedia.org/article/addressing-computer-generated-child-sex-abuse-imagery-legal-framework-and-policy-implications.

[449] Riana Pfefferkorn. Teens Are Spreading Deepfake Nudes of One Another. It's No Joke. https://www.scientificamerican.com/article/teens-are-spreading-deepfake-nudes-of-one-another-its-no-joke/, June 2024.

[450] Katrina Piatek-Jimenez, Jennifer Cribbs, and Nicole Gill. College students' perceptions of gender stereotypes: making connections to the underrepresentation of women in STEM fields. *International Journal of Science Education*, 40:1432–1454, 2018.

[451] Dudley L. Poston, editor. *Handbook of population*. Springer, 2nd edition, 2019.

[452] Anastasia Powell and Nicola Henry. Sexual violence and harassment in the digital era. *The Palgrave handbook of Australian and New Zealand criminology, crime and justice*, pages 205–220, 2017.

[453] Anastasia Powell, Nicola Henry, and Asher Flynn. Image-based sexual abuse. *Routledge Handbook of Critical Criminology*, 2, 2018.

[454] Anastasia Powell, Nicola Henry, Asher Flynn, and Adrian J Scott. Image-based sexual abuse: The extent, nature, and predictors of perpetration in a community sample of Australian residents. *Computers in Human Behavior*, 92:393–402, 2019.

[455] Ahmad R. Pratama and Firman M. Firmansyah. Until you have something to lose! Loss aversion and two-factor authentication adoption. *Applied Computing and Informatics*, 2021.

[456] Samuel H. Preston, Patrick Heuveline, and Michel Guillot. *Demography: measuring and modeling population processes*. Blackwell, 2001.

[457] 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), 2021.

[458] Project Nia. A Restorative Conversation Toolkit. Technical report, Project Nia, May 2021.

[459] Prolific. Why prolific? https://www.prolific.co/prolific-vs-mturk/, 2020. Accessed: 2021-07-06.

[460] Helen C Purchase. *Experimental human-computer interaction: a practical guide with visual examples*. Cambridge University Press, 2012.

[461] Sarvech Qadir, Andy Niser, Xavier V Caddle, Ashwaq Alsoubai, Jinkyung Katie Park, and Pamela J. Wisniewski. Towards a Safer Digital Future: Exploring Stakeholder Perspectives on Creating a Sustainable Youth Online Safety Community. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, pages 1–10, New York, NY, USA, May 2024. Association for Computing Machinery.

[462] Lucy Qin, Vaughn Hamilton, Yigit Aydinalp, Marin Scarlett, Sharon Wang, and Elissa M. Redmiles. Towards safer intimate futures: Recommendations for tech platforms to reduce image based sexual abuse. https://www.eswalliance.org/toward_safer_intimate_futures_recommendations_tech_platforms_reduce_image_based_abuse, 2024.

[463] Lucy Qin, Vaughn Hamilton, Sharon Wang, Yigit Aydinalp, Marin Scarlett, and Elissa M. Redmiles. "Did They F***ing Consent to That?": Safer Digital Intimacy via Proactive Protection Against Image-Based Sexual Abuse. In *Proceedings of the 33rd USENIX Security Symposium*, pages 55–72. USENIX Association, 2024.

[464] Li Qiwei, Francesca Lameiro, Shefali Patel, Cristi-Isaula-Reyes, Eytan Adar, Eric Gilbert, and Sarita Schoenebeck. Feminist Interaction Techniques: Deterring Non-Consensual Screenshots with Interaction Techniques, April 2024. arXiv:2404.18867 [cs].

[465] Li Qiwei, Allison Mcdonald, Oliver L Haimson, Sarita Schoenebeck, and Eric Gilbert. The Sociotechnical Stack: Opportunities for Social Computing Research in Non-consensual Intimate Media. In *Proceedings of the ACM on Human-Computer Interaction*, volume 8, pages 1–21, 2024.

[466] Emilee Rader, Rick Wash, and Brandon Brooks. Stories as informal lessons about security. In *Eighth Symposium on Usable Privacy and Security*, 2012.

[467] RAINN. What is Child Sexual Abuse Material (CSAM) | RAINN. https://rainn.org/news/what-child-sexual-abuse-material-csam, August 2022.

[468] Reethika Ramesh, Anjali Vyas, and Roya Ensafi. Security at the end of the tunnel: The anatomy of VPN mental models among experts and Non-Experts in a corporate context. In *Proceedings of the USENIX Security Symposium*. USENIX Association, August 2023.

[469] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H. Sung. Deepfake Detection: A Systematic Literature Review. *IEEE Access*, 10:25494–25513, 2022.

[470] Yolanda A. Rankin and Jakita O. Thomas. Straighten up and fly right: Rethinking intersectionality in HCI research. *interactions*, 26(6):64–68, October 2019.

[471] Yasmeen Rashidi, Tousif Ahmed, Felicia Patel, Emily Fath, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. "you don't want to be the next meme": College students' workarounds to manage privacy in the era of pervasive photography. In *Proc. SOUPS*, 2018.

[472] Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1):89:1–89:29, April 2023.

[473] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J. Wisniewski. A human-centered systematic literature review of the computational approaches for online sexual risk detection. *Proceedings of the ACM on Human-Computer Interacttion*, 5, 2021.

[474] Afsaneh Razi, John S. Seberger, Ashwaq Alsoubai, Nurun Naher, Munmun De Choudhury, and Pamela J. Wisniewski. Toward Trauma-Informed Research Practices with Youth in HCI: Caring for Participants and Research Assistants When Studying Sensitive Topics. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1):134:1–134:31, April 2024.

[475] Elissa M Redmiles. Net Benefits: Digital Inequities in Social Capital, Privacy Preservation, and Digital Parenting Practices of US Social Media Users. In *AAAI*, 2018.

[476] Elissa M. Redmiles. Friction Matters: Balancing the Pursuit of Perfect Protection With Target Hardening. *IEEE Security & Privacy*, 22(1):76–75, January 2024.

[477] Elissa M. Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L. Mazurek. A Summary of Survey Methodology Best Practices for Security and Privacy Researchers. Technical report, Technical Reports of the Computer Science Department, 2017.

[478] Elissa M. Redmiles, Mia M. Bennett, and Tadayoshi Kohno. Power in Computer Security and Privacy: A Critical Lens. *IEEE Security & Privacy*, March/April 2023.

[479] Elissa M Redmiles, Jessica Bodford, and Lindsay Blackwell. "I just want to feel safe": A Diary Study of Safety Perceptions on Social Media. In *Proc. ICWSM*, 2019.

[480] Elissa M Redmiles, Neha Chachra, and Brian Waismeyer. Examining the Demand for Spam: Who Clicks? In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[481] 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 *ACM Conference on Computer and Communications Security*, CCS '16, pages 666–677, 2016.

[482] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. Where is the digital divide? a survey of security, privacy, and socioeconomics. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017.

[483] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2019.

[484] Elissa M. Redmiles, Amelia R. Malone, and Michelle L. Mazurek. I Think They're Trying to Tell Me Something: Advice Sources and Selection for Digital Security. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, SP '16, pages 272–288. IEEE, 2016.

[485] 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 *Proceedings of the USENIX Security Symposium*, pages 89–108, August 2020.

[486] 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, pages 1–13, New York, NY, USA, April 2018. Association for Computing Machinery.

[487] Robert W. Reeder, Iulia Ion, and Sunny Consolvo. 152 Simple Steps to Stay Safe Online: Security Advice For Non-Tech-Savvy Users. *IEEE Security & Privacy*, 15(5):55–64, 2017.

[488] Gavin Rees. Handling Traumatic Imagery: Developing a Standard Operating Procedure. https://dartcenter.org/resources/handling-traumatic-imagery-developing-standard-operating-procedure, April 2017.

[489] Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A.F. Almeida, and Wagner Meira Jr. Characterizing and detecting hateful users on twitter. In *AAAI International Conference on Web and Social Media*, 2018.

[490] Jessica Ringrose, Kaitlyn Regehr, and Betsy Milne. Understanding and combatting youth experiences of image-based sexual harassment and abuse. Technical report, Department of Education, Practice and Society, UCL Institute of Education, 2021.

[491] Kate Roberts, Anthony Dowell, and Jing-Bao Nie. Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. *BMC Medical Research Methodology*, 19:1–8, 2019.

[492] Phillip Rogaway. Radical CS. Technical report, University of California, Davis, June 2024.

[493] Fabio Rojas. *Theory for the working sociologist.* Columbia University press, New York, 2017.

[494] Kevin Roundy, Paula Mendelberg, Nicola Dell, Damon McCoy, Daniel Nissani, Thomas Ristenpart, and Acar Tamersoy. The Many Kinds of Creepware Used for Interpersonal Attacks. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, SP '20, pages 626–643. IEEE, 2020.

[495] Yanet Ruvalcaba and Asia A. Eaton. Nonconsensual pornography among U.S. adults: A sexual scripts framework on victimization, perpetration, and health correlates for women and men. *Psychology of Violence*, 10(1):68–78, January 2020.

[496] Mohammad Hammas Saeed, Shiza Ali, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Trollmagnifier: Detecting state-sponsored troll accounts on reddit. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, pages 2161–2175. IEEE, 2022.

[497] Rogelio Sáenz and Maria Cristina Morales. Demography of race and ethnicity. In *Handbook of Population*. Springer, 2nd edition, 2019.

[498] Angela Saini. *Superior: The Return of Race Science.* Penguin Random House, 2019.

[499] Jerome H Saltzer and Michael D Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975.

[500] Sonam Samat and Alessandro Acquisti. Format vs. Content: The Impact of Risk and Presentation on Disclosure Decisions. In *Proc. SOUPS*, 2017.

[501] Nithya Sambasivan, Nova Ahmed, Amna Batool, Elie Bursztein, Elizabeth Churchill, Laura Sanely Gaytan-Lugo, Tara Matthews, David Nemar, Kurt Thomas, and Sunny Con-

solvo. Toward Gender-Equitable Privacy and Security in South Asia. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2019.

[502] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. "They don't leave us alone anywhere we go": Gender and digital abuse in south asia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 2019.

[503] Nithya Sambasivan, Garen Checkley Checkley, Nova Ahmed, Amna Batool, David Nemer, Laura Sanely Gaytán-Lugo, Tara Matthews, Sunny Consolvo, and Elizabeth Churchill. "privacy is not for me, it's for those rich women": Performative privacy practices on mobile phones by women in south asia. In *Fourteenth Symposium on Usable Privacy and Security*, 2018.

[504] SAMHSA. Trauma and Violence - What Is Trauma and Its Effects?, November 2024.

[505] SAMHSA. Trauma-Informed Approaches and Programs. https://www.samhsa.gov/mental-health/trauma-violence/trauma-informed-approaches-programs, Tue, 12/03/2024 - 3:58 pm.

[506] Shruti Sannon and Andrea Forte. Privacy Research with Marginalized Groups: What We Know, What's Needed, and What's Next. *CSCW*, 2022.

[507] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. Self-confidence trumps knowledge: A cross-cultural study of security behavior. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2202–2214, 2017.

[508] Joseph Schafer, Brett A. Halperin, Sourojit Ghosh, and Julie Vera. To Screenshot or Not to Screenshot? Tensions in Representing Visual Social Media Platform Posts. http://spir.aoir.org, October 2024.

[509] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. Safe spaces and safe places: Unpacking technology-mediated experiences of safety and harm with transgender people. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 2018.

[510] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction*, 2021.

[511] Morgan Klaus Scheuerman, Katta Spiel, Oliver L. Haimson, Foad Hamidi, and Stacy M. Branham. HCI Gender Guidelines. https://www.morgan-klaus.com/gender-guidelines.html, 2020.

[512] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *CSCW*, 2020.

[513] Ari Schlesinger, W Keith Edwards, and Rebecca E Grinter. Intersectional HCI: Engaging identity through gender, race, and class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[514] Bruce Schneier. The Hacking of Culture and the Creation of Socio-Technical Debt, June 2024.

[515] Philipp Schulz, Anne-Kathrin Kreft, Heleen Touquet, and Sarah Martin. Self-care for gender-based violence researchers – Beyond bubble baths and chocolate pralines. *Qualitative Research*, 23(5):1461–1480, October 2023.

[516] Carol F Scott, Gabriela Marcu, Riana Elyse Anderson, Mark W Newman, and Sarita Schoenebeck. Trauma-Informed Social Media: Towards Solutions for Reducing and Healing Online Harm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, Hamburg Germany, April 2023. ACM.

[517] James C. Scott. *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, 2020.

[518] Silvia Semenzin and Lucia Bainotti. The use of Telegram for non-consensual dissemination of intimate images: Gendered affordances and the construction of masculinities. *Social Media + Society*, 6(4), 2020.

[519] Sensity. The State of Deepfakes 2024. Technical report, 2024.

[520] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. Dancing to the Partisan Beat: A First Analysis of Political Communication on TikTok. In *ACM Conference on Web Science*, pages 257–266, 2020.

[521] Reena Shah and Jane Ogden. 'What's in a face?' The role of doctor ethnicity, age and gender in the formation of patients' judgements: an experimental study. *Patient Education and Counseling*, 60(2):136–141, 2006.

[522] Jenessa R. Shapiro and Amy M. Williams. The role of stereotype threats in undermining girls' and women's performance and interest in stem fields. *Sex Roles*, 66:175–183, 2012.

[523] Mahmood Sharif, Jumpei Urakawa, Nicolas Christin, Ayumu Kubota, and Akira Yamada. Predicting Impending Exposure to Malicious Content from User Behavior. In *Proc. CCS*, 2018.

[524] Richard Shay, Saranga Komanduri, Patrick Gage Kelley, Pedro Giovanni Leon, Michelle L. Mazurek, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Encountering Stronger Password Requirements: User Attitudes and Behaviors. In *Proc. SOUPS*, 2010.

[525] Kim Bartel Sheehan. An investigation of gender differences in on-line privacy concerns and resultant behaviors. *Journal of Interactive Marketing*, 1999.

[526] Renee Shelby, Shalaleh Rismani, Kathryn Henne, Ajung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 723–741, Montréal QC Canada, August 2023. ACM.

[527] 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 2010 CHI Conference on Human Factors in Computing Systems*, 2010.

[528] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Cranor, Jason Hong, and Elizabeth Nunge. Anti-Phishing Phil: The design and evaluation of

a game that teaches people not to fall for phish. In *ACM International Conference Proceeding Series*, volume 229, pages 88–99, 2007.

[529] Max Sheridan. Doxxing Statistics in 2024. Technical report, SafeHome.org, August 2024.

[530] Austen Shipley. Demopolis middle school shaken by AI-generated porn scandal. https://yellowhammernews.com/demopolis-middle-school-shaken-by-ai-generated-porn-scandal/, December 2023.

[531] Patrick E Shrout and Joseph L Rodgers. Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology*, 69:487–510, 2018.

[532] Wendy Sigle. Demography's theory and approach: (How) has the view from the margins changed? *Population Studies*, 75(sup1):235–251, 2021.

[533] Lucy Simko, Ada Lerner, Samia Ibtasam, Franziska Roesner, and Tadayoshi Kohno. Computer security and privacy for refugees in the united states. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, pages 409–423. IEEE, 2018.

[534] Peter Simons. "Tetris for Trauma" Viral Twitter Thread: A Master Class in Misleading Psych Research. https://www.madinamerica.com/2021/10/tetris-trauma-viral-twitter-thread-master-class-misleading-psych-research/, October 2021.

[535] Natasha Singer. Students Target Teachers in Group TikTok Attack, Shaking Their School. https://www.nytimes.com/2024/07/06/technology/tiktok-fake-teachers-pennsylvania.html, July 2024.

[536] Natasha Singer. Teen Girls Confront an Epidemic of Deepfake Nudes in Schools. https://www.nytimes.com/2024/04/08/technology/deepfake-ai-nudes-westfield-high-school.html, April 2024.

[537] Mohit Singhal, Chen Ling, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. Sok: Content moderation in social media, from guidelines to enforcement, and research to practice. *arXiv*, 2022.

[538] Audrey Smedley and Brian D Smedley. Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race. *American Psychologist*, 60(1):16, 2005.

[539] Linda Tuhiwai Smith. *Decolonizing Methodologies: Research and Indigenous Peoples*. Zed, second edition edition.

[540] Mat Smith. Japan's noisy iPhone problem. https://www.engadget.com/2016-09-30-japans-noisy-iphone-problem.html, 2016.

[541] Rachel Charlotte Smith, Heike Winschiers-Theophilus, Daria Loi, Rogério Abreu De Paula, Asnath Paula Kambunga, Marly Muudeni Samuel, and Tariq Zaman. Decolonizing Design Practices: Towards Pluriversality. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–5. ACM.

[542] Peter Snyder, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2017.

[543] Shin So-yoon. Reports of deepfakes at Korean schools surge to 434 in span of two weeks. https://english.hani.co.kr/arti/english_edition/e_national/1157908.html.

[544] Sichao Song, Jun Baba, Junya Nakanishi, Yuichiro Yoshikawa, and Hiroshi Ishiguro. Mind The Voice!: Effect of Robot Voice Pitch, Robot Voice Gender, and User Gender on User Perception of Teleoperated Robots. In *Proc. CHI EA*, 2020.

[545] 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 *Proc. SOUPS*, 2011.

[546] Clare Southerton. Lip-Syncing and Saving Lives: Healthcare Workers on TikTok. *International Journal of Communication*, 15, 2021.

[547] Brandon Sparks. A snapshot of image-based sexual abuse (IBSA): Narrating a way forward. *Sexuality Research & Soc. Policy*, 19(2):689–704, 2022.

[548] Brandon Sparks, Skye Stephens, and Sydney Trendell. Image-based sexual abuse: Victim-perpetrator overlap and risk-related correlates of coerced sexting, non-consensual dissemination of intimate images, and cyberflashing. *Computers in Human Behavior*, 148:107879, 2023.

[549] Steven J. Spencer, Claude M. Steele, and Diane M. Quinn. Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, 35(1):4–28, 1999.

[550] Robert E. Stake. *The Art of Case Study Research.* SAGE, 1995.

[551] Christine R Starr. "I'm not a science nerd!" STEM stereotypes, identity, and motivation among undergraduate women. *Psychology of Women Quarterly*, 42(4):489–503, 2018.

[552] Valerie Steeves and Owain Jones. Surveillance, Children and Childhood. *Surveillance & Society*, 7(3/4):187–191, 2010.

[553] Sophie Stephenson, Christopher Nathaniel Page, Miranda Wei, Apu Kapadia, and Franziska Roesner. Sharenting on TikTok: Exploring Parental Sharing Behaviors and the Discourse Around Children's Online Privacy. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17, Honolulu HI USA, May 2024. ACM.

[554] StopNCII. How StopNCII.org works. https://stopncii.org/how-it-works/, 2024.

[555] Scott R Stroud. The dark side of the online self: A pragmatist critique of the growing plague of revenge porn. *J. of Mass Media Ethics*, 29(3):168–183, 2014.

[556] Fred Stutzman and Jacob Kramer-Duffield. Friends Only: Examining a Privacy-Enhancing Behavior in Facebook. In *Proceedings of the 2010 CHI Conference on Human Factors in Computing Systems*, 2010.

[557] Mario I Suárez and Patrick Slattery. Resisting erasure: Transgender, gender nonconforming, and nonbinary issues in curriculum studies. *Journal of Curriculum and Pedagogy*, 15(3):259–262, 2018.

[558] Marit Sukk and Andra Siibak. Caring Dataveillance and the Construction of "Good Parenting": Estonian Parents' and Pre-teens' Reflections on the Use of Tracking Technologies. *Communications*, 46(3):446–467, 2021.

[559] Saguy Tamar, Reifen-Tagar Michal, and Joel Daphna. The gender-binary cycle: the perpetual relations between a biological-essentialist view of gender, gender ideology, and gender-labelling and sorting. *Phil. Trans. R. Soc. B*, (1822), 2021.

[560] Wai Yen Tang and Jesse Fox. Men's harassment behavior in online video games: Personality traits and game factors. *Aggressive Behavior*, 42(6):513–521, 2016.

[561] Cara Tannenbaum, Robert P. Ellis, Friederike Eyssel, James Zou, and Londa Schiebinger. Sex and gender analysis improves science and engineering. *Nature*, 575(7781):137–146, 2019.

[562] Tangila Islam Tanni, Mamtaj Akter, Joshua Anderson, Mary Jean Amon, and Pamela J. Wisniewski. Examining the Unique Online Risk Experiences and Mental Health Outcomes of LGBTQ+ versus Heterosexual Youth. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–21, New York, NY, USA, May 2024. Association for Computing Machinery.

[563] Una Tellhed, Martin Bäckström, and Fredrik Björklund. Will I Fit in and Do Well? The Importance of Social Belongingness and Self-Efficacy for Explaining Gender Differences in Interest in STEM and HEED Majors. *Analyses of Social Issues and Public Policy*, 13(1):347–369, 2013.

[564] David R. Thomas. A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2):237–246, 2006.

[565] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. SoK: Hate, Harassment, and the Changing Landscape of Online Abuse. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, SP '21, pages 247–267. IEEE, 2021.

[566] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. "It's common and a part of being a content creator": Understanding How Creators Experience and Cope with Hate and Harassment Online. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022.

[567] Sigal Tifferet. Gender differences in privacy tendencies on social network sites: A meta-analysis. *Computers in Human Behavior*, 93:1–12, 2019.

[568] TikTok. Community guidelines. https://www.tiktok.com/community-guidelines, 2022.

[569] Brian Timmerman, Pulak Mehta, Progga Deb, Kevin Gallagher, Brendan Dolan-Gavitt, Siddharth Garg, and Rachel Greenstadt. Studying the Online Deepfake Community. *Journal of Online Trust and Safety*, 2(1), September 2023.

[570] Patricia Godeke Tjaden. *Extent, nature, and consequences of intimate partner violence.* US DOJ, Office of Justice Programs, National Institute of Justice, 2000.

[571] Alex Trelinski. School children who 'created AI porn images of teenagers' are quizzed by police in eastern Spain. https://www.theolivepress.es/spain-news/2024/02/19/school-children-who-created-ai-porn-images-of-teenagers-are-quizzed-by-police-in-eastern-spain/, February 2024.

[572] Eric Trist. The evolution of socio-technical systems. Ontario Quality of Working Life Centre,

June 1981.

[573] Emily Tseng, Rosanna Bellini, Nora McDonald, Matan Danos, Rachel Greenstadt, Damon McCoy, Nicola Dell, and Thomas Ristenpart. The Tools and Tactics Used in Intimate Partner Surveillance: An Analysis of Online Infidelity Forums. In *Proceedings of the USENIX Security Symposium*, 2020.

[574] Emily Tseng, Diana Freed, Kristen Engel, Thomas Ristenpart, and Nicola Dell. A Digital Safety Dilemma: Analysis of Remote Computer-Mediated Computer Security Interventions During COVID-19. In *ACM Conference on Human Factors in Computing Systems*, CHI '21, pages 1–17, 2021.

[575] Emily Tseng, Mehrnaz Sabet, Rosanna Bellini, Harkiran Kaur Sodhi, Thomas Ristenpart, and Nicola Dell. Care Infrastructures for Digital Security in Intimate Partner Violence. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022.

[576] Carolyn A Uhl, Katlin J Rhyner, Cheryl A Terrance, and Noël R Lugo. An examination of nonconsensual pornography websites. *Feminism & Psychology*, 28(1):50–68, 2018.

[577] Jodie B Ullman and Peter M Bentler. Structural equation modeling. *Handbook of Psychology, Second Edition*, 2, 2012.

[578] Rebecca Umbach, Nicola Henry, Gemma Faye Beard, and Colleen M. Berryessa. Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–20, New York, NY, USA, May 2024. Association for Computing Machinery.

[579] Scientific United Nations Educational and Cultural Organization. International Standard Classification of Education (ISCED) 2021. https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf.

[580] UNODC. Comprehensive Study on Cybercrime. Draft, United Nations Office on Drugs and Crime, Vienna, February 2013.

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

[582] Blase Ur, Jaeyeon Jung, and Stuart Schechter. Intruders Versus Intrusiveness: Teens' and Parents' Perspectives on Home-Entryway Surveillance. In *UbiComp*, pages 129–139, 2014.

[583] Dirk Van Bruggen, Shu Liu, Mitch Kajzer, Aaron Striegel, Charles R. Crowell, and John D'Arcy. Modifying Smartphone User Locking Behavior. In *Proc. SOUPS*, 2013.

[584] VAWnet. Violence against trans and non-binary people. https://vawnet.org/sc/serving-trans-and-non-binary-survivors-domestic-and-sexual-violence/violence-against-trans-and.

[585] Victoria State Government. Foundation Knowledge Guide: Guidance for professionals working with child or adult victim survivors, and adults using family violence. Technical report, July 2021.

[586] Virginia Sexual and Domestic Violence Action Alliance. "Perpetrator" vs. "Victim" and the Impact of Carceral Logic – Virginia Sexual & Domestic Violence Action Al-

liance. https://vsdvalliance.org/press_release/perpetrator-vs-victim-and-the-impact-of-carceral-logic/, June 2024.

[587] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2017.

[588] Kandrea Wade, Jed R. Brubaker, and Casey Fiesler. Protest Privacy Recommendations: An Analysis of Digital Surveillance Circumvention Advice During Black Lives Matter Protests. In *Extended Abstracts of the Conference on Human Factors in Computing Systems*, CHI EA '21, pages 1–6, 2021.

[589] Eli Wald. Glass Ceilings and Dead Ends: Professional Ideologies, Gender Stereotypes, and the Future of Women Lawyers at Large Law Firms. *Fordham L. Rev.*, 78(5):2245–2288, 2010.

[590] Ge Wang, Jun Zhao, Max Can Kleek, and Nigel Shadbolt. Protection or Punishment? Relating the Design Space of Parental Control Apps and Perceptions About Them to Support Parenting for Online Safety. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW):1–26, 2021.

[591] Jane Ward. *The Tragedy of Heterosexuality*. New York University Press, 2020.

[592] 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 *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2022.

[593] Rick Wash. How experts detect phishing scam emails. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28, 2020.

[594] Rick Wash and Emilee Rader. Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users. In *SOUPS*, 2015.

[595] Ruth Webber. Sexual assault in relationships: Seeking help on a Q&A website. *Australian Social Work*, 67(3):363–376, 2014.

[596] 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 *Proceedings of the USENIX Security Symposium*, Philadelphia, PA, USA, August 2024. USENIX.

[597] Miranda Wei, Sunny Consolvo, Patrick Gage Kelley, Tadayoshi Kohno, Franziska Roesner, and Kurt Thomas. "There's so much responsibility on users right now:" Expert advice for staying safer from hate and harassment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 2023.

[598] Miranda Wei, Pardis Emami-Naeini, Franziska Roesner, and Tadayoshi Kohno. Skilled or Gullible? Gender Stereotypes Related to Computer Security and Privacy. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2023.

[599] Miranda Wei, Christina Yeung, Franziska Roesner, and Tadayoshi Kohno. "we're utterly

ill-prepared to deal with something like this": Teachers' perspectives on student generation of synthetic nonconsensual explicit imagery. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025.

[600] 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 *Proc. Symposium on Usable Privacy and Security*, SOUPS '18. USENIX, August 2022.

[601] White House. White House Announces New Private Sector Voluntary Commitments to Combat Image-Based Sexual Abuse | OSTP. https://www.whitehouse.gov/ostp/news-updates/2024/09/12/white-house-announces-new-private-sector-voluntary-commitments-to-combat-image-based-sexual-abuse/, September 2024.

[602] Robin Whittemore, Susan K. Chase, and Carol Lynn Mandle. Validity in Qualitative Research. *Qualitative Health Research*, 11(4):522–537, July 2001.

[603] Alma Whitten and J. D. Tygar. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In *Proceedings of the USENIX Security Symposium*, 1999.

[604] David Gray Widder, Dawn Nafus, Laura Dabbish, and James Herbsleb. Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2035–2046, Seoul Republic of Korea, June 2022. ACM.

[605] Lauren Wilcox, Renee Shelby, Rajesh Veeraraghavan, Oliver L Haimson, Gabriela Cruz Erickson, Michael Turken, and Rebecca Gulotta. Infrastructuring Care: How Trans and Non-Binary People Meet Health and Well-Being Needs through Technology. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[606] Montrischa M. Williams and Casey George-Jackson. Using And Doing Science: Gender, Self-Efficacy, and Science Identity of Undergraduate Students in STEM. *Journal of Women and Minorities in Science and Engineering*, 20(2):99–126, 2014.

[607] Sue Winkle Williams, Shirley M Ogletree, William Woodburn, and Paul Raffeld. Gender roles, computer attitudes, and dyadic computer interaction performance in college students. *Sex Roles*, 29(7):515, 1993.

[608] Hugh Wimberly and Lorie M Liebrock. Using fingerprint authentication to reduce system security: An empirical study. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2011.

[609] Langdon Winner. Do artifacts have politics? *Daedalus*, pages 121–136, 1980.

[610] Rachel Winter and Anastasia Salter. Deepfakes: Uncovering hardcore open source on GitHub. *Porn Studies*, 7(4):382–397, 2020.

[611] Pamela Wisniewski, Arup Kumar Ghosh, Heng Xu, Mary Beth Rosson, and John M. Carroll. Parental Control vs. Teen Self-Regulation: Is There a Middle Ground for Mobile Online Safety? In *ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 51–69, 2017.

[612] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, Daniel F. Perkins, and John M. Carroll.

Dear Diary: Teens Reflect on Their Weekly Online Risk Experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3919–3930, San Jose California USA, May 2016. ACM.

[613] Jacob O. Wobbrock and Julie A. Kientz. Research Contributions in Human-Computer Interaction. *Interactions*, 23(3):38–44, 2016.

[614] World Population Review. Reddit Users by Country 2024. https://worldpopulationreview.com/country-rankings/reddit-users-by-country, 2024.

[615] Yuxi Wu, William Agnew, W. Keith Edwards, and Sauvik Das. Design(ing) Fictions for Collective Civic Reporting of Privacy Harms. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–26, May 2025.

[616] Yuxi Wu, Panya Gupta, Miranda Wei, Yasemin Acar, Sascha Fahl, and Blase Ur. Your secrets are safe: How browsers' explanations impact misconceptions about private browsing mode. In *Proc. WWW*, 2018.

[617] Sijia Xiao, Shagun Jhaver, and Niloufar Salehi. Addressing Interpersonal Harm in Online Gaming Communities: The Opportunities and Challenges for a Restorative Justice Approach. *ACM Transactions on Computer-Human Interaction*, 30(6):1–36, December 2023.

[618] Sijia Xiao, Haodi Zou, Amy Mathews, Jingshu Rui, Coye Cheshire, and Niloufar Salehi. SnuggleSense: Empowering Online Harm Survivors Through a Structured Sensemaking Process, April 2025. arXiv:2504.19158 [cs].

[619] Sijia Xiao, Haodi Zou, Alice Qian Zhang, Deepak Kumar, Hong Shen, Jason Hong, and Motahhare Eslami. What Comes After Harm? Mapping Reparative Actions in AI through Justice Frameworks, June 2025. arXiv:2506.05687 [cs].

[620] Chaeyoon Yoo and Paul Dourish. Anshimi: Women's Perceptions of Safety Data and the Efficacy of a Safety Application in Seoul. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[621] Rachel Younger. AirTags become 'weapon of choice of stalkers and abusers' as cases rocket, April 2024. Section: national.

[622] YouTube. Creator Safety Center. https://www.youtube.com/creators/safety/.

[623] Jing Zeng, Mike S .Schäfer, and Joachim Allgaier. Reposting "Till Albert Einstein is TikTok Famous": The Memetic Construction of Science on TikTok. *International Journal of Communication*, 15:3216–3247, 2020.

[624] Renwen Zhang, Natalya N. Bazarova, and Madhu Reddy. Distress Disclosure across Social Media Platforms during the COVID-19 Pandemic: Untangling the Effects of Platforms, Affordances, and Audiences. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[625] Leah Zhang-Kennedy, Christine Mekhail, Yomna Abdelaziz, and Sonia Chiasson. From Nosy Little Brothers to Stranger-Danger: Children and Parents' Perception of Mobile Threats. In *International Conference on Interaction Design and Children*, pages 388–399, 2016.

[626] Wenqi Zheng, Emma Walquist, Isha Datey, Xiangyu Zhou, Kelly Berishaj, Melissa Mcdonald, Michele Parkhill, Dongxiao Zhu, and Douglas Zytko. "It's Not What We Were Trying to

Get At, but I Think Maybe It Should Be": Learning How to Do Trauma-Informed Design with a Data Donation Platform for Online Dating Sexual Violence. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–15, New York, NY, USA, May 2024. Association for Computing Machinery.

[627] Jonathan Zong. From individual consent to collective refusal: Changing attitudes toward (mis)use of personal data. *XRDS: Crossroads, The ACM Magazine for Students*, 27(2):26–29, December 2020.

[628] Jonathan Zong and J. Nathan Matias. Data Refusal from Below: A Framework for Understanding, Evaluating, and Envisioning Refusal as Design. *ACM Journal on Responsible Computing*, 1(1):1–23, March 2024.

[629] Yixin Zou, Allison McDonald, Julia Narakornpichit, Nicola Dell, Thomas Ristenpart, Kevin Roundy, Florian Schaub, and Acar Tamersoy. The Role of Computer Security Customer Support in Helping Survivors of Intimate Partner Violence. In *Proceedings of the USENIX Security Symposium*, 2021.

[630] Yixin Zou, Kevin Roundy, Acar Tamersoy, Saurabh Shintre, Johann Roturier, and Florian Schaub. Examining the Adoption and Abandonment of Security, Privacy, and Identity Theft Protection Practices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[631] Liza Zvi and Mally Shechory Bitton. Perceptions of victim and offender culpability in non-consensual distribution of intimate images. *Psychology, Crime & Law*, 27(5):427–442, 2021.

[632] Liza Zvi and Mally Shechory-Bitton. Police officer perceptions of non-consensual dissemination of intimate images. *Frontiers in Psychology*, page 2148, 2020.

Part V

Appendices

Appendix A

Help-Seeking and Help-Giving on Reddit for IBSA

A.1 Additional Methodological Details

A.1.1 Subreddits used

The subreddits we manually selected, as described in Section 4.3.1, were: r/advice, r/techsupport, r/legaladvice, r/legaladviceuk, r/legaladviceindia, r/legaladvicecanada, r/legaladviceeurope, r/legaladvicegerman, r/legaladviceNZ, r/legaladviceireland, r/legaladviceeu, r/relationship_advice, r/dating_advice, r/relationships, r/dating, r/marriage, r/sex, r/online_dating_advice, r/indian_datingadvice, r/femaledatingstrategy, r/onlinedating, r/tinder, r/bumble, r/hingeapp, r/grindr, r/onlyfansadvice, r/fansly_advice, r/camgirlproblems, r/antipornography, r/loveafterporn, r/pornfreerelationships, r/pornfree, r/sexualassault, r/sexualassaultsurvivor, r/creepyPMs, r/twoxchromosomes, r/thegirlsurvivalguide, r/askwomen, r/askwomennsfw, r/askmen, r/askgaybros, r/sextortion, r/scams.

A.1.2 Filtering: Keywords vs. LLM Prompts

As described in Section 4.3.1, we initially explored using keywords chosen from prior work, manual searches, discussion with experts, and our own domain expertise. We collected 60+ keywords and developed a three-part formula for surfacing relevant posts: >1 media keyword (e.g., *image, nude, screenshot, personal vid*), >1 IBSA keyword (e.g., *revenge porn, cyberflashing, without consent, sextortion, leak, threat, coerce*), >1 help keyword (e.g., *help, support, what should i do*). However, keyword searches introduced a prohibitive number of false positives to manually review, e.g., discussing IBSA in the news, posts that were not help-seeking.

Therefore, we developed the LLM-based filtering and categorizing approach described in Section 4.3.1; figures A.1 and A.2 show additional prompts. All posts were subsequently validated through manual review.

Title: [TITLE]
Article: [ARTICLE]

Task: Categorize the Title and Article as either
- 1) Sextortion
- 2) Nude image taken without consent
- 3) Cyberflashing or receiving unwanted nudity
- 4) Manipulated media like deepfakes, photoshopped images, or cheapfakes
- 5) Pressuring someone into sending nude images
- 6) Unwanted attention
- 7) Recording of sexual assault
- 8) Other

Categorization:"""

Figure A.1: Categorization prompt designed to identify the most relevant type of IBSA discussed in a Reddit post. IBSA type names and description differ from types shown in Table 4.1 due to experimental iteration with LLM prompts.

A.1.3 Codebook

The full list of codes applied to the 261 posts in our post dataset, and the 160 associated threads, with abbreviated definitions for each code. Codes were not mutually exclusive.

1) Codes about the Nature of IBSA
Perpetrator: current partner, ex-partner, friend, family members, colleague (work or school), stranger (unknown to victim-survivor prior to IBSA), other (none of the above), unspecified
Perpetrator(s)' and victim-survivors' gender, age, location: in terms used by post
Method of distribution

- Threat: Perpetrator threatens to share or distribute the image
- Possession: Perpetrator has possession of the image without threatening to or actually sharing
- Social media: Perpetrator posted the image to social media
- Website: Perpetrator posted the image to a website or elsewhere online
- Messaging: Perpetrator directly messaged the image to others
- N/A (no content): Perpetrator pressured the VS to send an image, but VS did not
- Unspecified

Image origin

- Consented: Shared for a different reason or in a prior context, but consent has been revoked
- Coerced: Sent or taken under coercion from the perpetrator including: under false pretenses; non-secret, non-consensual image taking; soliciting face images for abuse
- Coerced (not sent): VS has not sent image, but perpetrator attempted coercion
- Created: Synthetically created
- Hacked: Obtained through hacking or unauthorized access
- Sold: Available for sale by VS, but not for sharing

You are a content moderator for safety policies. Your job is to carefully inspect articles and answer each question below using the single word "Yes" or "No".

Title: [TITLE]
Article: [ARTICLE]

Question 1: Does the title or article discuss a threat to expose nude images unless a payment is made?
Question 2: Does the title or article discuss a threat to expose nude images unless the person sends more nude images or stays in a relationship?
Question 3: Does the title or article discuss taking, recording, or leaking a nude image without consent?

Your answer should be in json format seen below, where your answer should replace the "‘ :
{
"Question 1": "‘,
"Question 2": "‘,
"Question 3": "‘
}

Figure A.2: Categorization prompt to determine whether a post discussed a specific type of IBSA (here, nonfinancial sextortion). A Python parser validated the sequence of Yes and No answers necessary for a category label to apply.

- Secretly recorded: Recording not known to VS when recorded
- Unconsented exposure: VS nonconsensually exposed to someone else's image
- Unspecified

2) Codes about Help-Seeking

Self-help: Yes (poster is the person experiencing the abuse), No (poster is NOT the person experiencing the abuse)

Strategies attempted

- Support: Talking with others in a support network, online or offline; does not include the Reddit post itself
- Platform report: Using online platform reporting mechanisms
- Police report: Contacting the police or making a police report
- Workplace report: Contacting HR or workplace supervisor
- Engage: Negotiating conditions of abuse with the perpetrator
- Disengage: Stopping contact or other means of engagement with the perpetrator to try to mitigate the abuse
- Securing accts: Taking actions to secure online accounts
- Evidence: Recording evidence of the abuse
- Delete content: Deleting images, even if not all copies
- Other
- Unspecified

Help sought: See Table 4.5.

When poster sought help: Based on the user states framework [296].

- Prevention: Aiming to prevent future exposure to digital-safety risks; often but not always during relatively calm state of mind.
- Monitoring: Watching for digital-safety events to quickly respond; often with low to moderate stress.
- Crisis / Active event: Actively experiencing a digital-safety event (possibly hours/days); likely want to stop event; often very high stress.
- Recovery: Digital-safety event stopped or it's been determined stopping it may not be possible; ready to address damage; possibly much later; often moderate to high stress.
- Multiple events: Complex cases with multiple or overlapping digital-safety events; often heightened stress and trauma.

3) Codes about Help-Giving: See Table 4.6.

A.2 Additional Results

Tables A.1 and A.2 show, respectively and for each IBSA type, strategies previously attempted by victim-survivors and platform(s) involved.

Table A.1: Strategies attempted before posting on Reddit.

Strategy	FS	NFS	NCSEI	PS	CF	NCEI	RSA
Disengage	22	18	22	18	10	8	12
Engage	16	10	11	22	8	15	6
Securing accounts	22	9	9	0	2	2	0
Social support	5	5	2	2	0	7	5
Police report	10	0	4	0	2	4	5
Platform report	9	1	6	1	1	2	1
Evidence	7	0	6	0	1	2	1
Delete content	2	1	0	1	0	4	2
Workplace report	0	1	1	2	4	0	0
Other	2	1	2	0	1	1	0
Unspecified	9	14	13	8	17	18	17

Table A.2: Platforms involved in the IBSA. Messaging also includes messages or images otherwise described in posts as "sent."

Platform	FS	NFS	NCSEI	PS	CF	NCEI	RSA
Messaging (apps, DMs)	42	26	30	31	32	15	8
Mainstream social media	31	16	23	3	5	10	4
Device	2	11	9	3	0	18	27
Dating app	11	2	5	7	6	0	0
Adult content (websites, apps)	0	1	2	0	0	3	4
Social discovery app	1	1	0	0	1	2	0
Financial / commerce	3	0	1	0	1	1	0
Website	0	1	0	0	0	2	0
Unspecified	0	4	1	7	2	1	1

Appendix B

Anti-Privacy and Anti-Security Advice on TikTok

B.1 Summary Table

Table B.1: A summary of all of the motivations, goals and techniques we observed in our dataset, across two interpersonal contexts: intimate partner relationships and parent-child relationships. We identify what goals were sought for what motivations, with which techniques.

	Goal (what?)	**Motivation** (why?)	**Techniques** (how?)
Intimate Partner Context	*Instigator Perspective*		
	Surveil digital communications	Detect cheating	Use data downloads to obtain message history (and other metadata)
			Check recently used emojis for sexually explicit emojis
			Takeover Snapchat account with 2FA vulnerability
	Stalk on social media	Detect cheating	Find public conversations between target and suspected affair partner
	Surveil dating app usage	Detect cheating	Use 3rd party site to see if on dating app
			See if email address already exists on dating app
			Create fake account to see if on dating app
	Surveil other digital activities	Detect cheating	Look at photo metadata to determine when it was originally taken
			Get physical access to data on phone: explicit photos, vault apps that could hide explicit photos, porn websites in browsing history, dating apps, emails from hookup sites
	Surveil physical world	Detect cheating	Use AirTags/AirPods to track target's location
			Use monitoring apps (Life360)
			Get physical access to view location on phone or in accounts (Google Maps, iOS Significant Locations)
			Abuse accessibility features to listen (Live Listen, auto-answer calls)
	Stalk on social media	Arbitrary surveillance	Use 3rd party site to anonymously view target's Instagram stories or display photo
			See order of who target recently followed on Instagram website
			Use app to detect when target is signing on/off WhatsApp
			Use app to see searched/clicked/viewed your Instagram
			Create fake account to view Instagram story
			Keep track of Snapchat score to see if mass sending
	Manipulate social media	Exert control	Restrict account on Instagram, sends DM to message requests to evade read receipts and get more time to respond
			Change phone time to delete previously sent WhatsApp message
			Create fake tag in Instagram story using poll feature and see who clicks
			Hide and unhide story so instigator's Instagram story appears first
	Text someone who blocked you	Exert control	Message from email (does not work)
	Target Perspective		
	Detect call surveillance	Evade surveillance	Check carrier settings for call forwarding or redirection
Parent-Child Context	*Parent Perspective*		
	Surveil physical world	Child safety	Hide AirTag in bag, clothing, or car
			Give AirTag bracelet or keychain
			Install tracking app (Life360)
	Surveil digital world	Child safety	Sync iCloud messages
			Use text forwarding
	Restrict content and usage	Exert control	Locked down smartphone
			Parental control apps (Bark, FamiSafe)
	Child Perspective		
	Evade location tracking app	Location privacy	Disable app tracking cellular data permissions
			Put phone on Do Not Disturb
			Install app on another device
	Evade digital surveillance	Device privacy	Hide home screen pages
	Evade parental controls	Autonomy	Brute force passcode by detecting fingerprints on screen
			Use different VPN
			Sign out of app store and use new Apple ID

Appendix C

Teachers' Perspectives on Student Generation of SNCEI

C.1 Study materials

C.1.1 Recruitment materials

The flyer used to recruit participants for our interviews is shown in Figure C.1.

C.1.2 Semi-structured interview protocol

Our full interview protocol included the following questions. Note that since this was a semi-structured interview, the interviewer may have asked slightly altered or additional questions depending on the particular conversation.

Introduction
Thanks so much for joining, I'm [name of primary interviewer], this is [name of second interviewer], we're both graduate students at [institution name]. I'll be leading the interview and [name of second interviewer] will be taking notes. I have a few things to get through before we start the interview. And we should be done in about an hour. Sound good?

Did you get a chance to look at the consent form before we begin? Do you have any questions or does everything look okay to you? *If not, link to consent form and ask participant to read it.*

This isn't part of the study, but just as a check for us: What school do you work at?

And could you show a teacher ID, school T-shirt, notebook, or something else to confirm that you work for that school?

- If yes, or plausible reason about going to get it: Ah okay, perfect, I believe you! This was mainly to confirm participants are actually teachers, we've had some folks say they were but upon being asked for ID, they hung up.
- If no: Alright, please reschedule for a different time and have something ready so we can confirm your status as a teacher. Thanks!

Figure C.1: Recruitment flyer used to solicit participants for our interview study.

I have a fine-print paragraph I need to read for you, if you could just bear with me:
As researchers at a public university, I am a mandatory reporters of child abuse. This means that if you reveal identifiable information (e.g., name, organization, or location) of someone who has created or is in possession of pornographic depictions, real or synthetic, of someone under 18, I would be required to pass on this information to the appropriate authorities. For the purposes of this study, I'm only interested in your general knowledge and opinions and suggest that you do not reveal unnecessary identifiable information about yourself or others. Sound good? Any questions?

And last thing before we begin, would it be okay with you if we record this interview? This is mainly for the transcript so we can make sure we're catching what you said accurately.

Context Setting
In what context do you typically work with youth?

- How old are the youth?
- How long have you been working in this context?

How long in working with youth, overall, in any capacity?

How big is your school overall?

Could you describe the youth you typically work with – what kind of background or any other characteristics that stick out to you?

General Perceptions and Experiences Related to Deepfake Nudes
There's no right or wrong answers for any questions that I ask in this interview –= my main goal is to understand what your thoughts and opinions are. When I say 'deepfake nudes', what comes to mind for you?

Yup, seems like we're on the same page. So for the purposes of this study, we are most interested in learning about…
[[Depending on what they say, also remind them of the following details:]]

- Images generated/edited by a computer, whether through photoshop, AI tools, or other means (e.g., faceswap, undressing)
- Images that are sexually explicit or depict nudity
- Images that where the subject did NOT consent to their creation or sharing

In the last 6 months, how often have you seen any news stories about synthetic/deepfake nudes in schools? What do you remember about them?

In the last 6 months, how often have you talked to colleagues or students about synthetic/deepfake nudes, including overhearing conversations? What do you remember those conversations?

If someone had a laptop/tablet/phone and an internet connection, what else do you think they would need in order to **create** deepfake nudes of someone specific that they know?

And similarly, if someone had a laptop/tablet/phone and an internet connection, what else do you think they would need in order to **view** deepfake nudes / synthetic content of someone specific that they know?

You've said that [details from participants]. Do you think these factors would be true if it was specifically the students at your school that were creating or viewing synthetic content?

If you had to take a guess, do you think it's more likely a student would find it themselves vs. see it because others shared it with them?

Threat Model
What are some situations in which you could see a student creating deepfakes of someone else?

- For what reasons do you think a student would create deepfake nudes of someone else?
- Once an image is created, what do you think the student would do with it?

Speaking in general terms, what kind of students do you think might **create** deepfake nudes?

- Do you think students would work alone or together?

Speaking in general terms, what kind of students do you think might **be a victim** of deepfake nudes?

[Optional] And do you think people of certain genders would be more likely to create or have images created of them?

Case Studies
So now I'd like to share a little bit of what we've learned while conducting this research. We've been searching online and have found some stories local and national news, that…

- This has happened at least 9[1] schools in the US in different states
- In pretty much all cases, it was one or more boys, creating images of multiple female classmates
- Though the articles didn't always include all details, many described that students were using AI tools or apps, and in some cases, sharing the images through Snapchat
- Students who created the images were between 13 and 17. In some cases, the people in the images were between 12 and 15

What are your first impressions?

Do you think something similar is possible in your context?

What would you do if this happened in your context?

- [Optional] Would you report to law enforcement? [if asked, assure them you're not testing them, just curious how they think about it]
- [Optional] Do you think there are existing legal policies to handle this?

What do you think the appropriate response would be or what the consequences should be (if any) for a student who creates images of other students?

[Optional] What do you think other students would do if they received deepfake nude images from someone else, and showing someone else in the images?

Potential Interventions
What kinds of additional resources would be helpful for you to figure out what to do?

If you wanted to create policies or tools to prevent deepfake nudes, what would you do?

- [Optional] If you were asked to talk to your students at the beginning of the school year about deepfake nudes, what would you say to them?

What kinds of policies that exist for other related concerns might or might not apply to deepfake nudes? -OR- Does your school have content or courses about sex ed for your students?

Additional optional questions about prevention:

- [Optional] How does your school handle incidents that happen outside of school hours?
- [Optional] How well do you feel like your school responds to online safety concerns or incidents?

[1] At the time of the interviews, we had collected nine cases, but by the time of writing the paper, one additional case had been reported.

Deepfakes in the Future [Optional, if time]
We're almost at the end of the interview, just some final high-level questions before we wrap up.

With respect to deepfake nudes and students, what are your biggest fears in the next five years?

With respect to deepfake nudes and students, what are your biggest hopes in the next five years?

Conclusion and Wrap-up
Just a last thing to mention: We've been using the term 'deepfake nudes.' But of course this is still an emerging area of research and you might see people use many different terms. Some say 'deepfake porn', although there are academic researchers who argue that calling this type of content "pornography" is wrong because pornography should refer to consensual content and the person did not consent. Also, academics tend to call this content 'synthetic' or 'AI-generated' because those are more accurate terms than 'deepfake' is a reference to a specific computer science technique for creating fake images.

Do you have any questions for me? Or any questions I can answer about what we talked about today?

Thank you so much for your time today. As our research progresses, would you like to be kept up to date as we reach more milestones, such as reviewing a copy of the transcript from this interview, reviewing a draft of the paper, or being contacted about future research opportunities?

Appendix D

Gender Stereotypes Related to Computer S&P

D.1 Survey Instrument

[Consent form] We are researchers at the University of Washington (UW) studying security and privacy in human contexts.

This study was reviewed by the UW Institutional Review Board (IRB) and deemed exempt because it involves no more than minimal risk and meets other criteria. Your responses to this survey will be anonymized. Data from this survey will be stored securely and kept confidential. Your participation in this study is voluntary. You may withdraw your participation at any time. If you have questions about this study, you may contact Miranda Wei (PhD student at UW) at weimf@cs.washington.edu. You may also contact the UW Human Subjects Division (HSD), which manages IRB review, at hsdinfo@uw.edu.

I am at least 18 years old, I have read and understood this consent form, and I agree to participate in this online research study. ○ Yes ○ No

[Introduction] This survey has five sections. While we understand there are many genders, for the purposes of this study, we will ask about specifically women and men.

- Sections 1-3: Rate whether men or women are more likely to do certain things
- Section 4: Elaborate on a few answers you gave in Sections 1-3
- Section 5: Answer general questions about your experiences

The survey will conclude with demographics questions.

Your survey responses are anonymous. We will not ask for identifying information. In this survey, we are interested in your honest thoughts and opinions. There are no right or wrong answers, and your responses have no impact on your compensation.

[General trends] Section [number] contains questions about 3 trends in people's lives. Trend [number] (out of 3):

Based on your personal beliefs and experiences, who is more likely to [[be more interested in learning how to protect their own computer security and privacy; know more about how to protect their own computer security and privacy; be better at protecting their own computer security and privacy]]? ○ Definitely men ○ Probably men ○ Men and women equally ○ Probably women ○ Definitely women ○ Another gender, please specify: _____ ○ Don't know or not sure

[Personal characteristics] Section [number] contains questions about 6 characteristics that people might have. Characteristic [number] (out of 6):

Based on your personal beliefs and experiences, who is more likely to be more [[logical; lazy; overconfident; perceptive; emotional; gullible]] when it comes to computer security and privacy? ○ Definitely men ○ Probably men ○ Men and women equally ○ Probably women ○ Definitely women ○ Another gender, please specify: _____ ○ Don't know or not sure

[Specific tasks] Section [number] contains questions about 10 actions that people could take. Action [number] (out of 10):

Based on your personal beliefs and experiences, who is more likely to [[be a victim of online shopping-related scams; be a victim of dating-related financial scams; verify that a site is using HTTPS when submitting sensitive information online; keep software up-to-date; leave personal devices (e.g., smartphones, computers) unlocked and/or unattended; use anti-virus or anti-malware software on personal computers; use the same password for multiple accounts; use two-factor authentication for personal accounts (by connecting an account to a trusted phone number, backup email address, or phone app); share sensitive information on social media; ask for help if they have questions about protecting their security or privacy]]? ○ Definitely men ○ Probably men ○ Men and women equally ○ Probably women ○ Definitely women ○ Another gender, please specify: _____ ○ Don't know or not sure

[Selected follow-up questions] In Section 4, you will be asked to elaborate on some of the answers you gave in Sections 1-3.

Previous question: [Previous question text]
Your answer: [Previous question answer]

Why do you believe men and women are different in this way? Select all that apply. ○ Biological reasons ○ Non-biological reasons ○ Other reasons, please specify: _____ ○ Don't know or not sure

People believe that men and women are different for many reasons. For each reason you selected above, briefly explain or give an example why you believe that reason was relevant. _____

[Open-ended questions] Section 5 contains questions about your general experiences and beliefs.

Idea: "People of one gender are better than others at doing security- or privacy-related tasks." Prior to taking this survey, had you heard the idea above or something similar? If so, from where? Select all that apply. ○ Heard from friends ○ Heard from family ○ Heard from the news ○ Heard from social media ○ Heard from TV shows or movies ○ Heard from work or job ○ Heard from other, please specify: _____ ○ Never heard of differences among genders when it comes to security or privacy tasks

Have **you** ever been personally affected by a gender stereotype related to computer security or privacy? ○ Yes ○ No ○ Don't know or not sure

[If yes] Please describe (as much as you can) who made the assumption, what the stereotype was, and how you felt or reacted. _____

Do you know **anyone else** who has been personally affected gender stereotype related to computer security or privacy? ○ Yes ○ No ○ Don't know or not sure

[If yes] Please describe (as much as you can) who made the assumption, what the stereotype was, and how they felt or reacted. _____

[Ambivalent Sexism Inventory Questions, see Appendix D.2]

[Demographics] Almost done! This final page contains some demographic questions.

What is your gender? ○ Woman ○ Man ○ Non-binary ○ Prefer to self-describe _____ ○ Prefer not to say

Would you describe yourself as transgender? ○ Yes ○ No ○ Prefer not to say

What is your age? ○ 18-24 ○ 25-34 ○ 35-44 ○ 45-54 ○ 55-64 ○ 65 or older ○ Prefer not to say

How do you identify? Select all that apply, you may select more than one. ○ White ○ Hispanic, Latino, or Spanish origin ○ Black or African American ○ Asian ○ American Indian or Alaska Native ○ Middle Eastern or North African ○ Native Hawaiian or Other Pacific Islander ○ Some other race, ethnicity, or origin _____ ○ Prefer not to say

What is the highest degree or level of school you have completed? ○ High school or less ○ Some college ○ Trade/technical/vocational training ○ Associate's degree ○ Bachelor's degree ○ Master's degree ○ Professional degree or doctorate ○ Prefer not to say

Which of the following best describes your educational background or job field? ○ I have an education in, or work in the field of computer science, computer engineering, or IT ○ I do not have an education in, or work in the field of computer science, computer engineering, or IT ○ Prefer not to say

Which of the following best describes your educational background or job field? ○ I have an education in, or work in the field of **computer security and privacy in particular** ○ I do not have an education in, or work in the field of **computer security and privacy in particular** ○ Prefer not to say

How important is it to you that you be considered good at computer security or privacy tasks? ○ Not at all important ○ Somewhat important ○ Moderately important ○ Very important ○ Extremely important

What is your annual individual income? ○ Less than $20,000 ○ $20,000 to $49,999 ○ $50,000 to $99,999 ○ $100,000 to $250,000 ○ Over $250,000 ○ Prefer not to say

What is your annual household income? ○ Less than $20,000 ○ $20,000 to $49,999 ○ $50,000 to $99,999 ○ $100,000 to $250,000 ○ Over $250,000 ○ Prefer not to say

How comfortable did you feel while answering the questions in this survey? ○ Very comfortable ○ Comfortable ○ Neutral ○ Uncomfortable ○ Very uncomfortable

Thank you so much for your participation in our study! Do you have any final comments or questions? _____

D.2 Ambivalent Sexism Inventory (ASI)

Reproduced here from Glick & Fiske [221].

Below is a series of statements concerning men and women and their relationships in contemporary society. Please indicate the degree to which you agree or disagree with each statement.
Response options: ○ Disagree strongly ○ Disagree somewhat ○ Disagree slightly ○ Agree slightly ○ Agree somewhat ○ Agree strongly

1. No matter how accomplished he is, a man is not truly complete as a person unless he has the love of a woman.
2. Many women are actually seeking special favors, such as hiring policies that favor them over men, under the guise of asking for "equality."
3. In a disaster, women ought not necessarily to be rescued before men.
4. Most women interpret innocent remarks or acts as being sexist.
5. Women are too easily offended.
6. People are often truly happy in life without being romantically involved with a member of the other sex.
7. Feminists are not seeking for women to have more power than men.
8. Many women have a quality of purity that few men possess.
9. Women should be cherished and protected by men.
10. Most women fail to appreciate fully all that men do for them.
11. Women seek to gain power by getting control over men.
12. Every man ought to have a woman whom he adores.
13. Men are complete without women.
14. Women exaggerate problems they have at work.
15. Once a woman gets a man to commit to her, she usually tries to put him on a tight leash.
16. When women lose to men in a fair competition, they typically complain about being discriminated against.
17. A good woman should be set on a pedestal by her man.
18. There are actually very few women who get a kick out of teasing men by seeming sexually available and then refusing male advances.
19. Women, compared to men, tend to have a superior moral sensibility.
20. Men should be willing to sacrifice their own well being in order to provide financially for the women in their lives.
21. Feminists are making entirely reasonable demands of men.
22. Women, as compared to men, tend to have a more refined sense of culture and good taste.

Scoring instructions: Reverse the following items (1 = 6, 2 = 5, 3 = 4, 4 = 3, 5 = 2, 6 = 1): 3, 6, 7, 13, 18, 21. Hostile Sexism Score = average of the following items: 2, 4, 5, 7, 10, 11, 14, 15, 16, 18, 21. Benevolent Sexism Score = average of the following items: 1, 3, 6, 8, 9, 12, 13, 17, 19, 20, 22.

D.3 Main Study Qualitative Codebook

The full codebook, with themes and subthemes, from qualitatively analyzing free-text stereotype rationales in the main study. Codes were not mutually exclusive.

Other stereotypes

- Separate stereotypes: a separate stereotype than the question (e.g., forgetful, lazy, taking shortcuts, protective)
- Stereotypes outside of security and privacy

"Science": justified with science terms, e.g., "proven," "studies," "naturally," or "wired."

Observations

- Self: something they do themselves
- Others: something they have observed others doing; incl. actions, habits, hobbies, traits, e.g., women shop more, men use online dating more; "in my experience," "noticed that," or "women or men I know"

Threat model

- Assets: having (or not having) assets; e.g., "care," "concerned," "nothing to hide"; general valuations of SP, i.e., "want to protect info"
- Threats: recognition (or or not) of threats, e.g., prior experiences that inform what threats they are aware of; more "vulnerable" or "unsafe"
- External threats: external threat, e.g., scammers target this gender more

Assumptions

- About aptitude: assumptions about one gender's knowledge, experience, interest, or usage (or lack thereof), e.g., pay attention more, more capable ("better"); use the internet more, more interested in gaming or social media
- About background: assumptions made about one gender's education or career tendencies

Society: societal or cultural expectations, socialization, or conditioning; including how they want to be perceived

"Just because": no meaningful reason given

Table D.1: Results of multinomial logistic regression models belief in stereotypes by participants' ASIs. DV compares belief towards women and towards men with belief towards neither. (Int.) = Intercept. *Note:* * $p < .05$, ** $p < .01$, *** $p < .001$

Stereotype	DV level	Term	*Estimate*	*Std. Err.*	*Statistic*	*p-value*
Higher sexism scores correlated with beliefs towards women						
Be emotional	men	(Int.)	-2.20	0.80	-2.69	(*n.s.*)
	men	ASI	0.46	0.31	1.47	(*n.s.*)
	women	(Int.)	-1.96	0.56	-3.50	*
	women	ASI	1.00	0.22	4.65	***
Be gullible	men	(Int.)	-4.02	1.01	-3.99	**
	men	ASI	0.76	0.32	2.37	(*n.s.*)
	women	(Int.)	-2.74	0.57	-4.80	***
	women	ASI	0.85	0.19	4.43	***
Be lazy	men	(Int.)	-1.18	0.53	-2.21	(*n.s.*)
	men	ASI	0.27	0.20	1.37	(*n.s.*)
	women	(Int.)	-4.84	0.92	-5.29	***
	women	ASI	1.32	0.27	4.85	***
Fall for shopping scam	men	(Int.)	-6.89	2.14	-3.23	*
	men	ASI	1.45	0.60	2.42	(*n.s.*)
	women	(Int.)	-1.94	0.52	-3.73	**
	women	ASI	0.89	0.19	4.58	***
Ask for help if needed	men	(Int.)	-2.91	0.97	-3.00	(*n.s.*)
	men	ASI	0.44	0.35	1.26	(*n.s.*)
	women	(Int.)	-1.84	0.52	-3.55	*
	women	ASI	0.80	0.19	4.25	***
Reuse passwords	men	(Int.)	-3.56	0.76	-4.70	***
	men	ASI	0.74	0.24	3.15	(*n.s.*)
	women	(Int.)	-4.07	0.85	-4.82	***
	women	ASI	0.86	0.26	3.36	*
Leave device unlocked	men	(Int.)	-1.75	0.60	-2.90	(*n.s.*)
	men	ASI	0.28	0.22	1.29	(*n.s.*)
	women	(Int.)	-3.54	0.70	-5.02	***
	women	ASI	0.95	0.22	4.33	***
Higher sexism scores correlated with beliefs towards men						
Know how to protect	men	(Int.)	-1.48	0.50	-2.96	(*n.s.*)
	men	ASI	0.71	0.18	3.83	**
	women	(Int.)	-8.82	2.67	-3.30	*
	women	ASI	1.99	0.69	2.89	(*n.s.*)
Skilled at protecting	men	(Int.)	-1.63	0.50	-3.29	*
	men	ASI	0.7	0.18	3.95	**
	women	(Int.)	-4.99	2.16	-2.31	(*n.s.*)
	women	ASI	0.57	0.71	0.80	(*n.s.*)
Be perceptive	men	(Int.)	-2.26	0.56	-4.07	**
	men	ASI	0.76	0.19	3.97	**
	women	(Int.)	-2.74	0.71	-3.86	**
	women	ASI	0.64	0.24	2.67	(*n.s.*)

Be logical	men	(Int.)	-3.46	0.62	-5.55	***
	men	ASI	1.26	0.22	5.75	***
	women	(Int.)	-3.56	0.92	-3.86	**
	women	ASI	0.75	0.32	2.30	(*n.s.*)
Verify HTTTPS	men	(Int.)	-2.12	0.52	-4.05	**
	men	ASI	0.70	0.18	3.94	**
	women	(Int.)	-3.81	1.14	-3.36	*
	women	ASI	0.57	0.37	1.53	(*n.s.*)
Install software updates immediately	men	(Int.)	-2.42	0.55	-4.43	***
	men	ASI	0.78	0.18	4.27	**
	women	(Int.)	-2.55	1.01	-2.52	(*n.s.*)
	women	ASI	0.09	0.38	0.24	(*n.s.*)
Use 2FA	men	(Int.)	-2.07	0.54	-3.82	**
	men	ASI	0.62	0.18	3.41	*
	women	(Int.)	-2.24	0.74	-3.01	(*n.s.*)
	women	ASI	0.29	0.26	1.11	(*n.s.*)
Use antivirus software	men	(Int.)	-3.85	0.68	-5.64	***
	men	ASI	1.10	0.22	5.11	***
	women	(Int.)	-3.36	0.93	-3.63	*
	women	ASI	0.50	0.32	1.60	(*n.s.*)
Beliefs not correlated with sexism scores						
Be overconfident	men	(Int.)	0.27	0.54	0.50	(*n.s.*)
	men	ASI	0.37	0.20	1.83	(*n.s.*)
	women	(Int.)	-4.86	1.52	-3.19	(*n.s.*)
	women	ASI	1.16	0.44	2.63	(*n.s.*)
Interested in learning about protecting	men	(Int.)	-1.07	0.50	-2.12	(*n.s.*)
	men	ASI	0.29	0.17	1.72	(*n.s.*)
	women	(Int.)	-0.88	0.73	-1.20	(*n.s.*)
	women	ASI	-0.28	0.28	-1.00	(*n.s.*)
Share sensitive info on social media	men	(Int.)	-3.8	1.35	-2.82	(*n.s.*)
	men	ASI	0.74	0.39	1.91	(*n.s.*)
	women	(Int.)	0.94	0.51	1.82	(*n.s.*)
	women	ASI	0.00	0.18	-0.01	(*n.s.*)
Fall for dating scam	men	(Int.)	-0.45	0.55	-0.81	(*n.s.*)
	men	ASI	0.17	0.19	0.87	(*n.s.*)
	women	(Int.)	-0.80	0.56	-1.43	(*n.s.*)
	women	ASI	0.30	0.19	1.56	(*n.s.*)

Appendix E

SoK (or SoLK?)

E.1 Literature Review

E.1.1 Codebooks

We developed three codebooks for our literature review and detail here the topics we coded for, labels within each codebook, as well as definitions and/or examples of each label.

Type of Security Behavior.: See Table 5.2.

Recruitment / Sample Characteristics.

- **Representative**: successfully matched sample and population characteristics during recruitment
- **Balanced**: similar number of participants in factor groups for analysis, including work that happened to have balanced samples
- **Limited**: sample characteristics were uncontrolled, e.g., snowball / convenience samples, data collected based on non-sociodemographic criteria and descriptively reports sociodemographic data

Study Methods.

- **Self-report**: participants reporting security actions, e.g., interviews conducted in-person or via telephone, surveys
- **Measurement (Observational)**: scraped public data, log data (e.g., from a university, company), or data from software on participants' devices (e.g., mobile app or browser)
- **Measurement (Experimental)**: controlled condition or direct interaction involved, including non-lab study (e.g., MTurk experiment, experiment on social media) or lab study (e.g., in-person, in-lab studies)

E.1.2 Full List of Papers

Tables E.1, E.2, and E.3 show all papers in our literature review.

E.2 Case Study

E.2.1 Sociodemographic Factors

We provide here more details on the sociodemographic factors we study in our case study in Section 5.6. Based on the available log data about Facebook users, we chose six sociodemographic factors to study:

- **Age**, self-reported, bucketed into three groups; 25-34 (46.3%), 35-49 (31.8%), 50+ (21.9%); min: 25, median: 36, mean: 40.14, max: 100
- **Gender**, gender (encoded as binary): women (43.5%), men (56.5%)
- **Educational attainment**, self-reported, scaled per country (e.g., for Brazil, médio incompleto, superior completo, especialização), bucketed into three groups: high school equivalent or less (45.5%), some college (23.6%), BA or higher (30.8%)
- **Geographic location**, one of 16 platform-inferred countries grouped into four regions: *Western* (30.8%): France, Italy, U.S., U.K.; *Latin America* (17.4%): Brazil, Mexico; *Africa and Middle East* (22.4%): Egypt, Kenya, Nigeria, Turkey; *Asia* (29.4%): India, Indonesia, Myanmar, Pakistan, Philippines, Vietnam
- **Internet skill**, measured via Web-Use Skills Index [243, 244], a standardized and validated self-report measure (min: 1.5, median: 3.75, mean: 3.66, max: 5).
- **Technical knowledge**, measured via Pew Research's password knowledge question [422] and three additional questions designed in the same style to assess familiarity with downloads, QR codes, and reacting to posts on the platform of study. Questions ask, "Which of the following best describes" and gives 5 answer options. (Password: 54.1% correct[1]; Download: 61.5% correct; QR: 39.6% correct; Reaction: 46.9% correct)

E.2.2 Regression Results

Table E.4 shows regression results for our case study (see Section 5.6).

[1]Pew Research [422] found that 75% of U.S. survey respondents answered this question correctly; in our dataset 79.2% of U.S. respondents did so.

Table E.1: All 47 papers in our focus dataset, i.e., from our seven selected conferences.

Year	Authors	Title	Conference
2006	Dhamija et al.	Why Phishing Works	CHI
2007	Sheng et al.	Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish	SOUPS
2008	Gilbert et al.	The network in the garden: an empirical analysis of social media in rural life	CHI
2009	Chiasson et al.	Multiple password interference in text passwords and click-based graphical passwords	CCS
	Kumaraguru et al.	School of Phish: A Real-World Evaluation of Anti-Phishing Training	SOUPS
2010	Shay et al.	Encountering stronger password requirements: user attitudes and behaviors	SOUPS
	Sheng et al.	Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions	CHI
	Stutzman et al.	Friends only: examining a privacy-enhancing behavior in Facebook	CHI
2011	Kaye	Self-reported password sharing strategies	CHI
	Sotirakopoulos et al.	On the challenges in usable security lab studies: lessons learned from replicating a study on SSL warnings	SOUPS
	Wimberly et al.	Using Fingerprint Authentication to Reduce System Security: An Empirical Study	IEEE S&P
2012	Bonneau et al.	The science of guessing: analyzing an anonymized corpus of 70 million passwords	IEEE S&P
	Onarlioglu et al.	Insights into User Behavior in Dealing with Internet Attacks	NDSS
2013	Mazurek et al.	Measuring Password Guessability for an Entire University	CCS
2014	Chen et al.	Exploring Internet Security Perceptions and Practices in Urban Ghana	SOUPS
2015	Ion et al.	"...no one can hack my mind": Comparing Expert and Non-Expert Security Practices	SOUPS
	Jia et al.	Risk-taking as a Learning Process for Shaping Teen's Online Information Privacy Behaviors	CSCW
	Wash & Rader	Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users	SOUPS
2016	Cho et al.	Networked Privacy Management in Facebook: A Mixed-Methods and Multinational Study	CSCW
	Harbach et al.	Keep on Lockin' in the Free World: A Multi-National Comparison of Smartphone Locking	CHI
	Redmiles et al.	How I Learned to be Secure: a Census-Representative Survey of Security Advice Sources and Behavior	CCS
2017	Bonné et al.	Exploring decision making with Android's runtime permission dialogs using in-context surveys	SOUPS
	Fiesler et al.	What (or Who) Is Public?: Privacy Settings and Social Media Content Sharing	CSCW
	Hoyle et al.	Viewing the Viewers: Publishers' Desires and Viewers' Privacy Concerns in Social Networks	CSCW
	Pearman et al.	Let's Go in for a Closer Look: Observing Passwords in Their Natural Habitat	CCS
2018	Das et al.	Breaking! A Typology of Security and Privacy News and How It's Shared	CHI
	Machuletz	Webcam Covering as Planned Behavior	CHI
	Redmiles et al.	Examining the Demand for Spam: Who Clicks?	CHI
	Sharif et al.	Predicting Impending Exposure to Malicious Content from User Behavior	CCS
2019	Habib et al.	Impact of Contextual Factors on Snapchat Public Sharing	CHI
2020	Coopamootoo	Usage Patterns of Privacy-Enhancing Technologies	CCS
	Zou et al.	Examining the Adoption and Abandonment of Security, Privacy, and Identity Theft Protection Practices	CHI
2021	Hasan et al.	Your Photo is so Funny that I don't Mind Violating Your Privacy by Sharing it	CHI
	Zhang et al.	Distress Disclosure across Social Media Platforms during the COVID-19 Pandemic	CHI
2023	Bouma-Sims et al.	A US-UK Usability Evaluation of Consent Management Platform Cookie Consent Interface Design on Desktop and Mobile	CHI
	Munyendo et al.	I just stopped using one and started using the other: Motivations Techniques and Challenges When Switching Password Managers	CCS

Table E.2: The remaining papers in our full dataset (1999-2014).

Year	Authors	Title	Venue
2004	Milne & Culnan	Strategies for reducing online privacy risks: Why consumers read (or don'tread) online privacy notices	Jrnl. of Interactive Marketing
	Milne et al.	Consumers' Protection of Online Privacy and Identity.	Jrnl. of Consumer Affairs
2005	Youn	Teenagers' Perceptions of Online Privacy and Coping Behaviors: A Risk–Benefit Appraisal Approach	Jrnl. of Broadcasting & E. Media
2007	Grimes et al.	Email end users and spam: relations of gender and age group to attitudes and actions Computers in	Human Behavior
	Jagatic et al.	Social phishing	CACM
	Kumaraguru et al.	Getting Users to Pay Attention to Anti-Phishing Education: Evaluation of Retention and Transfer	APWG
	Kuo et al.	Assessing Gender Differences in Computer Professionals' Self-Regulatory Efficacy Concerning Info. Privacy Practices	Jrnl. of Business Ethics
2008	Bailey et al.	Analysis of Student Vulnerabilities to Phishing.	AMCIS
	Hazari et al.	An Empirical Investigation of Factors Influencing Information Security Behavior	Jrnl. of Info. Privacy & Security
	Lewis et al.	The Taste for Privacy: An Analysis of College Student Privacy Settings inan Online Social Network	Jrnl. of Comp. Mediated Comm.
	Youn & Hall	Gender and Online Privacy among Teens: Risk Perception, Privacy Concerns, and Protection Behaviors	Cyber Psychology & Behavior
2009	Dinev et al.	User behaviour towards protective information technologies: the role ofnational cultural differences	Information Systems Journal
	Fogel & Nehmad	Internet social network communities: Risk taking, trust, and privacy concerns	Computers in Human Behavior
	Milne et al.	Toward an Understanding of the Online Consumer's Risky Behavior and Protection Practices	Jrnl. of Consumer Affairs
2010	Brandtzæg	Too Many Facebook "Friends"? Content Sharing and Sociability Versus the Need for Privacy in Social Network Sites	Jrnl. of HCI
	Durand	A Comparative Study of Self-Disclosure in Face-to-Face and Email Communication Between Americans and China	N/A (thesis)
	Hoy & Milne	Gender Differences in Privacy-Related Measures for Young Adult Facebook Users	Jrnl. of Interactive Advertising
	Posey et al.	Proposing the online community self-disclosure model	Euro. Jrnl. of Information Systems
	Siripukdee et al.	Empirical Analysis of Human-related Problems on Information Security in Cross-cultural Environment	Japan Society for Info. & and Mgmt.
	Wright et al.	Where Did They Go Right? Understanding the Deception in Phishing Communications	Group Decision and Negotiation
2011	Kruger et al.	An assessment of the role of cultural factors in information security awareness	Information SecuritySouth Africa
	Lomo-David et al.	University Students Computer Security Practices in Two Developing Nations: A Comparative Analysis	SHSU General Business Conference
	Lowry et al.	Privacy Concerns Versus Desire for Interpersonal Awareness in Driving the Use of Self-Disclosure Technologies	Jrnl. of Management Info. Systems
	Maier et al.	An Assessment of Overt Malicious Activity Manifest in Residential Networks	DIMVA
	Special et al.	Self-disclosure and student satisfaction with Facebook Computers in	Human Behavior
2012	Krasnova et al.	Self-disclosure and Privacy Calculus on Social Networking Sites: The Roleof Culture	BISE
	Madden	Privacy management on social media sites	Pew
	Mohebzada et al.	Phishing in a university community: Two large scale phishing experiments	IIT
	Tufekci	Youth and Privacy in Networked Publics: Active and Complex Engagement	ICWSM
2013	Halevi et al.	A pilot study of cyber security and privacy related behavior and personality traits	WWW
	Litt	Understanding social network site users' privacy tool use Computers in	Human Behavior
	Madden et al.	Teens, Social Media, and Privacy	Pew
	Park	Digital Literacy and Privacy Behavior Online	Communication Research
	Rainie et al.	Anonymity, Privacy, and Security Online	Pew
2014	Alseadoon	The Impact of Users' Characteristics on Their Ability to Detect Phishing Emails	N/A (thesis)
	Baek et al.	My privacy is okay, but theirs is endangered: Why comparative optimism matters in online privacy concerns	Computers in Human Behavior
	Blank et al.	A New Privacy Paradox: Young People and Privacy on Social Network Sites	ASA

Table E.3: The remaining papers in our full dataset (2014-2023).

Year	Authors	Title	Venue
2014	Tembe et al.	Phishing in international waters	HotSoS
	Vanderhoven et al.	How Safe Do Teenagers Behave on Facebook? An Observational Study	PLoS One
2015	Anderson et al.	Neural correlates of gender differences and color in distinguishing security warnings and legitimate websites	Jrnl. of Cybersecurity
	Halevi et al.	Spear-Phishing in the Wild	SSRN
	Marshall et al.	Social networking websites in India and the United States: A cross-national comparison of online privacy and communication	Issues in IS
	Park	Do men and women differ in privacy? Gendered privacy and (in)equality in the Internet	Computers in Human Behavior
	Pattinson et al.	Factors that Influence Information Security Behavior: An Australian Web Based Study	HAS
	Posey et al.	The Impact of Organizational Commitment on Insiders' Motivation to Protect Organizational Information Assets	Jrnl. of Management Information Systems
	Whitty et al.	Individual Differences in Cyber Security Behaviors: An Examination of Who Is Sharing Passwords	Cyber Psychology, Behavior, and Social Networking
2016	Aviv et al.	Analyzing the Impact of Collection Methods and Demographics for Android's Pattern Unlock	USEC
	Bertenthal	Attention and Past Behavior, not Security Knowledge, Modulate Users' Decisions to Login to Insecure Websites	ICS
	Chen & Zahedi	Individuals' Internet Security Perceptions and Behaviors: Polycontextual Contrasts Between the United States and China	MIS Quarterly
	Halevi et al.	Cultural and psychological factors in cyber-security II	WAS
	Iuga et al.	Baiting the hook: factors impacting susceptibility to phishing attacks	Human-centric Computing and Info. Sciences
	Kezer et al.	Age differences in privacy attitudes, literacy and privacy management on Facebook	Jrnl. of Psychosocial Research on Cyberspace
	Malik et al.	Privacy and trust in Facebook photo sharing: Age and gender differences	Program
	Petrie et al.	Cultural and Gender Differences in Password Behaviors: Evidence from China, Turkey and the UK	NordiCHI
	Reed et al.	Thumbs up for privacy?: Differences in online self-disclosure behavior across national cultures	Social Science Research
	Sonnenschein et al.	Gender Differences in Mobile Users' IT Security Appraisals and Protective Actions: Findings from a Mixed-Method Study	ISIC
	Tsay-Vogel et al.	Social media cultivating perceptions of privacy	New Media & Society
2017	Anwar et al.	The impact of collectivism and psychological ownership on protection motivation: A cross-cultural examination	Computers in Human Behavior
	Büchi et al.	Caring is not enough: the importance of Internet skills for online privacy protection	ICS
	Butavicius et al.	Understanding susceptibility to phishing emails: Assessing the impact of individual differences and culture	HAISA
	Gavett et al.	Phishing suspiciousness in older and younger adults: The role of executive functioning	PLoS One
	Ifinedo et al.	Effects of Organization Insiders' Self-Control and Relevant Knowledge on Participation in Information Systems Security Deviant Behavior	SIGMIS-CPR
	Sarno et al.	Who are Phishers luring?: A Demographic Analysis of Those Susceptible to Fake Emails	Human Factors and Ergonomics Society
2018	Alohali et al.	Identifying and predicting the factors affecting end-users' risk-taking behavior	Jrnl. of Info. & Comp. Security
	Cain et al.	An exploratory study of cyber hygiene behaviors and knowledge	Jrnl. of Information Security and Applications
	Diaz et al.	Phishing in an Academic Community: A Study of User Susceptibility and Behavior	ArXiV
	Farinosi & Taipale	Who Can See My Stuff? Online Self-Disclosure and Gender Differences on Facebook	OBS
	Griffin	A Demographic Analysis to Determine User Vulnerability among Several Categories of Phishing Attacks	N/A (thesis)
	Lévesque et al.	Technological and Human Factors of Malware Attacks: A Computer Security Clinical Trial Approach	TOPS
	McGill et al.	Gender Differences in Information Security Perceptions and Behaviour	ACIS
	Menard et al.	The impact of collectivism and psychological ownership on protection motivation: A cross-cultural examination	Computers & Security

	Millham et al.	Managing the virtual boundaries: Online social networks, disclosure, and privacy behaviors	New Media & Society
	Redmiles	Net Benefits: Digital Inequities in Social Capital, Privacy Preservation, and Digital Parenting Practices of U.S. Social Media Users	ICWSM
2019	Dev et al.	Personalized WhatsApp Privacy: Demographic and Cultural Influences on Indian and Saudi Users	SSRN
	Lin et al.	Susceptibility to Spear-Phishing Emails: Effects of Internet User Demographics and Email Content	ToCHI
	Ndibwile et al.	A Demographic Perspective of Smartphone Security and Its Redesigned Notifications	Jrnl. of Information Processing
	Shappie et al.	Personality as a Predictor of Cybersecurity Behavior	Psychology of Popular Media Culture
2020	Tifferet et al.	Gender differences in privacy tendencies on social network sites: A meta-analysis	Computers in Human Behavior
	Breitinger et al.	A survey on smartphone user's security choices, awareness and education	Computers & Security
	Epstein et al.	Markers of Online Privacy Marginalization: Empirical Examination of Socioeconomic Disparities in Social Media Privacy Attitudes, Literacy, and Behavior	Social Media + Society
	Herbert et al.	Differences in IT Security Behavior and Knowledge of Private Users in Germany	Wirtschaftsinformatik
	Li et al.	Experimental Investigation of Demographic Factors Related to Phishing Susceptibility	HICCS
	Liu et al.	Effects of Demographic Factors on Phishing Victimization in the Workplace	PACIS
	Oghazi et al.	User self-disclosure on social network sites: A cross-cultural study on Facebook's privacy concepts	Jrnl. of Business Research
	Sombatruang et al.	Attributes affecting user decision to adopt a Virtual Private Network (VPN) app	ICICS
	Thao et al.	Human Factors in Homograph Attack Recognition	ANCS
	Zwilling et al.	Cyber Security Awareness, Knowledge and Behavior: A Comparative Study	Jrnl. of Comp. Info. Systems
2021	Abroshan et al.	COVID-19 and Phishing: Effects of Human Emotions, Behavior, and Demographics on the Success of Phishing Attempts During the Pandemic	IEEE Access
	Abroshan et al.	Phishing Happens Beyond Technology	IEEE Access
	Bhagavatula et al.	What breach? Measuring online awareness of security incidents by studying real-world browsing behavior	EuroUSEC
	Boerman et al.	Exploring Motivations for Online Privacy Protection Behavior: Insights From Panel	Data Comm. Research
	Greitzer et al.	Experimental Investigation of Technical and Human Factors Related to Phishing Susceptibility	ACM Transactions on Social Computing
	Grobler et al.	The importance of social identity on password formulations	Personal and Ubi. Comp.
	Kennison et al.	Who creates strong passwords when nudging fails	Computers in Human Behavior
	Mai et al.	Cyber Security Awareness and Behavior of Youth in Smartphone Usage: A Comparative Study between University Students in Hungary and Vietnam	Acta Polytechnica Hungarica
	Morrison	Understanding U.S. Employees' Personality Traits for Phishing Emails Prevention: A Quantitative Study	N/A (thesis)
	Ouytsel	The prevalence and motivations for password sharing practices and intrusive behaviors among early adolescents' best friendships – A mixed-methods study	Telematics and Informatics
	Roberts	Does Digital Native Status Impact End-User Antivirus Usage?	Jrnl. of Comp. Net. & Comm.
2022	Frank et al.	Contextual drivers of employees' phishing susceptibility: Insights from a field study	Decision Support Systems
2023	Du et al.	Phishing: Gender Differences in Email Security Perceptions and Behaviors	Info. Sys. and Comput. Academic Professionals

Table E.4: Regression results for the relationships between security behaviors (first row, output variables) and sociodemographic factors and platform metrics (first column, input factors). Each column represents the output of one regression model. Numeric cells list the odds ratio (OR) and the 95% confidence interval. Significance of OR: $p < 0.05$ = *, $p < 0.01$ = **, and $p < 0.001$ = ***. LATAM = Latin America, AME = Africa and Middle East, Edu. = Education, SC = some college, BA+ = Bachelor's degree or more, Tech. = Technical, Know. = Knowledge, L30 = Use (Past 30 days)

	Visit Security Settings	**Action Security Settings**	**Stronger Password**	**Use 2FA**
(Intercept)	0.02*** [0, 0.05]	0.02*** [0, 0.06]	192.25*** [29.13, 1268.55]	0*** [0, 0]
Age (35-49)	0.74* [0.59, 0.94]	0.55*** [0.44, 0.70]	1.18 [0.64, 2.17]	0.79* [0.64, 0.96]
Age (50+)	0.63* [0.42, 0.95]	0.38*** [0.23, 0.63]	2.08* [1.07, 4.03]	0.63* [0.43, 0.92]
Gender (woman)	1.20 [0.97, 1.49]	1.44** [1.14, 1.82]	1.55 [0.86, 2.80]	0.88 [0.73, 1.06]
Location (LATAM)	0.90 [0.56, 1.44]	0.90 [0.53, 1.55]	0.64 [0.20, 2.05]	0.85 [0.42, 1.71]
Location (AME)	0.87 [0.52, 1.46]	0.73 [0.40, 1.35]	0.24* [0.08, 0.71]	1.06 [0.55, 2.03]
Location (Asia)	1.94* [1.15, 3.28]	1.61 [0.87, 2.99]	0.16*** [0.06, 0.45]	1.44 [0.68, 3.02]
Edu. (SC)	1.25 [0.41, 3.79]	1.25 [0.32, 4.77]	0.57 [0.04, 8.41]	7.14** [2.14, 23.85]
Edu. (BA+)	1.36 [0.42, 4.39]	1.39 [0.37, 5.16]	0.89 [0.15, 5.30]	5.40** [1.74, 16.72]
Internet Skill	1.41** [1.12, 1.78]	1.44* [1.09, 1.90]	1.10 [0.79, 1.53]	1.84*** [1.42, 2.39]
Tech. Know. (Download)	1.02 [0.81, 1.29]	0.87 [0.68, 1.11]	0.90 [0.49, 1.65]	0.83 [0.66, 1.04]
Tech. Know. (Password)	0.97 [0.77, 1.23]	1.15 [0.89, 1.49]	1.88* [1.09, 3.23]	1.33* [1.01, 1.74]
Tech. Know. (QR)	1.14 [0.92, 1.42]	1.15 [0.90, 1.47]	1.64 [0.80, 3.36]	1.49*** [1.19, 1.87]
Tech. Know. (Reaction)	1.12 [0.90, 1.39]	1.24 [0.97, 1.58]	1.75* [1.02, 3.00]	1.37** [1.11, 1.70]
Platform Tenure (Years)	0.95** [0.91, 0.99]	0.95* [0.91, 0.99]	0.91* [0.84, 1.00]	1.08*** [1.04, 1.13]
Friends	1.00 [0.99, 1.01]	1.00 [0.99, 1.01]	1.01 [0.98, 1.04]	1.02** [1.00, 1.03]
Use (Past 30 days)	1.00 [0.98, 1.01]	0.99 [0.97, 1.00]	0.97 [0.93, 1.01]	1.01 [0.99, 1.03]
Time Spent	1.01 [1.00, 1.18]	1.03 [0.94, 1.12]	0.90 [0.63, 1.29]	1.13* [1.02, 1.24]
Edu. (SC) * Internet Skill	0.91 [0.68, 1.22]	0.89 [0.63, 1.26]	1.30 [0.65, 2.60]	0.63** [0.47, 0.85]
Edu. (BA+) * Internet Skill	0.88 [0.65, 1.18]	0.85 [0.60, 1.19]	0.92 [0.56, 1.50]	0.71* [0.53, 0.95]
L30 * Time Spent	1.00 [1.00, 1.00]	1.00 [1.00, 1.01]	1.01 [0.99, 1.02]	1.00 [0.99, 1.00]
LATAM * Platform Tenure	1.03 [0.97, 1.10]	1.05 [0.98, 1.12]	0.99 [0.87, 1.14]	0.95 [0.87, 1.03]
AME * Platform Tenure	1.05 [0.99, 1.11]	1.09* [1.02, 1.17]	1.08 [0.96, 1.22]	1.00 [0.94, 1.07]
Asia * Platform Tenure	0.97 [0.91, 1.03]	1.00 [0.93, 1.07]	1.12 [0.99, 1.27]	0.95 [0.88, 1.03]