Toward Trauma-Informed Research Practices with Youth in HCI: Caring for Participants and Research Assistants When Studying Sensitive Topics

AFSANEH RAZI, Drexel University, U.S.A
JOHN S. SEBERGER, Drexel University, U.S.A
ASHWAQ ALSOUBAI, Vanderbilt University, USA
NURUN NAHER, University of Central Florida, U.S.A
MUNMUN DE CHOUDHURY, Georgia Institute of Technology, U.S.A
PAMELA J. WISNIEWSKI, Vanderbilt University, USA

Research involving sensitive data often leads to valuable human-centered insights. Yet, the effects of participating in and conducting research about sensitive data with youth are poorly understood. We conducted meta-level research to improve our understanding of these effects. We did the following: (i) asked youth (aged 13-21) to share their private Instagram Direct Messages (DMs) and flag their unsafe DMs; (ii) interviewed 30 participants about the experience of reflecting on this sensitive data; (iii) interviewed research assistants (RAs, n=12) about their experience analyzing youth's data. We found that reflecting about DMs brought discomfort for participants and RAs, although both benefited from increasing their awareness about online risks, their behavior, and privacy and social media practices. Participants had high expectations for safeguarding their private data while their concerns were mitigated by the potential to improve online safety. We provide implications for ethical research practices and the development of reflective practices among participants and RAs through applying trauma-informed principles to HCI research.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **Empirical studies in collaborative and social computing**; • **Social and professional topics → Professional topics**.

Additional Key Words and Phrases: Adolescents, Teens, Datasets, Instagram, Data Collection, Research Ethics, Trauma-Informed Research, Sensitive Research, Online Safety

ACM Reference Format:
Afsaneh Razi, John S. Seberger, Ashwaq Alsoubai, Nurun Naher, Munmun De Choudhury, and Pamela J. Wisniewski. 2024. 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, Article 134 (April 2024), 31 pages. https://doi.org/10.1145/3637411

Authors' addresses: Afsaneh Razi, afsaneh.razi@drexel.edu, Drexel University, 3675 Market St 10th floor, Philadelphia, Pennsylvania, U.S.A, 19104; John S. Seberger, john.s.seberger@drexel.edu, Drexel University, 3675 Market St., Philadelphia, Pennsylvania, U.S.A, 19104; Ashwaq Alsoubai, ashwaq.alsoubai@vanderbilt.edu, atalsoubai@kau.edu.sa, Vanderbilt University, Nashville, Tennessee, USA; Nurun Naher, University of Central Florida, 4000, Orlando, Florida, U.S.A, nurun@Knights.ucf.edu; Munmun De Choudhury, Georgia Institute of Technology, 30318, Atlanta, Georgia, U.S.A, munmund@gatech.edu; Pamela J. Wisniewski, pamela.wisniewski@vanderbilt.edu, Vanderbilt University, Nashville, Tennessee, USA;.

ACM 2573-0142/2024/4-ART134
https://doi.org/10.1145/3637411

1 INTRODUCTION

Engaging technology users in research about their sensitive data and/or adverse life events may cause such users – and the researchers tasked with analysis – unintentional harm (c.f., [5, 29, 80]). Recognition of this problem has yielded productive knowledge about protecting and honoring both participant and researcher welfare (e.g., through post-research interventions [82] concerning bereavement [40], sexual abuse [31], pregnancy loss [6], and mental health challenges [48]). Yet the effects of sensitive research on youth, aged 13 to 21, remain poorly understood. The relative scarcity of work on this topic presents a two-fold problem: first, youth are particularly vulnerable to online risks and harms [84]; second, youth constitute tomorrow's adults – people's whose digitally-oriented behaviors, experiences, and attitudes will shape the social norms of the future regarding today's emergent infrastructures of daily life (e.g., artificial intelligence, social media apps, etc.) [14]. Therefore, there is a present need for meta-level research (i.e., research about research processes) to understand how participants and researchers experience high-risk, high-reward research and how we, as Human-Computer Interaction (HCI) researchers, can proactively mitigate harm by applying trauma-informed practices throughout the research process.

In the social sciences, "trauma-informed" approaches to research are well-established [15]; however, similar approaches have only recently been adopted by the computing fields [18]. Such approaches remain under-utilized by human-computer interaction (HCI) researchers. Trauma-informed approaches can build upon existing ethical and methodological frameworks to inform how ACM Special Interest Group on Computer–Human Interaction (SIGHCI) researchers conduct research with vulnerable participants.

To work toward developing trauma-informed research practices in HCI for youth, we engaged with youth (ages 13-21) who were asked to participate in a study [54] in which they donated their Instagram data for the purpose of studying potentially traumatic online risks they encountered in their private messages. Youth were asked to flag conversations that made them feel uncomfortable or unsafe. As a follow-up, youth participants were asked to participate in a retrospective interview regarding their experience participating in the study. As part of the larger research project, we also had undergraduate research assistants (RAs) annotate youth data for risky interactions. For this paper, we interviewed youth participants *and* RAs regarding their experiences reflecting on the risk-flagged data. In doing so, we conducted a meta-level research study on the potentially traumatic or uncomfortable experiences of participants and the researcher assistants who annotated their data. Our engagement with youth participants and RAs was driven by the following high-level research questions:

- **RQ1:** *When youth are asked to reflect on potentially sensitive or negative experiences from their past social media interactions, how does this affect them?*
- **RQ2:** *When college-aged research assistants are asked to analyze the sensitive and potentially risky social media interactions of youth, how did this affect them?*

To answer these questions, we conducted retrospective semi-structured interviews with youth aged 13 to 21 (n=30) and college-aged research assistants (n=12). Interviews allowed us to probe and understand: (i) *participant* experiences of sharing their private Instagram Direct Messages (DMs) and annotating them for unsafe/uncomfortable interactions [54]; and (ii) *research assistant* experiences of annotating such data. We found that youth who donated their data and reflected upon its contents and significance during the interview process found the experience to be beneficial in terms of contributing to research, increasing their awareness about online risks, their reactions, privacy practices, and social engagements. Although some negative feelings were brought up by the nature of those old uncomfortable interactions, participants expressed that they were glad that they participated. Initial concerns over data privacy were assuaged by the feeling that participation in research would yield better safety for other youth online. Similarly, we found that engagement with sensitive data sometimes made the RAs uncomfortable, but such discomfort was transient. From the annotation process, RAs also learned about online risks, reflected on their own privacy practices, adjusted their online behaviors, and warned the people they know about online risks nowadays. Through analysis and discussion, we make the following contributions to research in HCI and youth online safety:

- We provide insights on the ethical treatment of research participants who engage in more than minimal risk research and research assistants that annotated unsafe sensitive online interactions; and
- We lay foundations for how we can better inform practices around trauma-informed HCI research with youth and their sensitive social media data.

This paper contributes to an emerging body of HCI research that goes beyond data privacy and sharing challenges in research. These challenges include understanding the dynamics at play for what is sensitive or intimate data, the impact of exploring behavioral data on stakeholders, and the open field opportunities for awareness generation and value exchange through these processes.

2 BACKGROUND

In this section, we consider research ethics and practices from two perspectives. First, we consider the need to develop more robust research ethics from the perspective of study participants – particularly those who are asked to engage with potentially traumatic digital stimuli (e.g., historical social media data). Second, we engage in similar consideration from the perspective of researchers, thus recognizing and accounting for the shared condition of *being human* that is distributed among all parties involved in research. Finally, we situate these two perspectives within the trend toward trauma-informed research practices.

2.1 Research Ethics of Working with Minors regarding Online Safety

Many researchers (e.g., [8, 9, 57, 82]) have investigated ethical challenges, especially in working with vulnerable populations, such as adolescents. Because of their developmental stage [69], adolescents are vulnerable to many risk factors that could cause mental health issues [25] such as exploration of sexual identity [32, 55] and excessive use of technology [47]. Walker et al. [82] provided heuristics to consider the needs of vulnerable and marginalized populations involved in research. Such heuristics include considerations for pre-research activities (e.g., needfinding of vulnerable community, relationship of researchers with the community, appropriate compensation, consent), activities during the act of research (e.g., power differentials, data considerations, empowering participants), and post-research activities (e.g., presenting findings, disclosing positionality of researchers, dissemination of the results). Only a few researchers [1, 9, 17] have investigated participants' views about online safety research. Such work primarily focused on participatory design with youth before involving them in research [1, 9, 17] or comparing teens to their parents [1, 9]. For example, Badillo-Urquiola et al. [9] engaged 20 youth in a co-design process to improve adolescent online safety research. They further interviewed 13 parents of the 20 youth participants. On the one hand, they found that adolescents were motivated to share their data to benefit society, while they feared getting in trouble. On the other hand, parents wanted researchers to solve problems facing their teen children. That teens *and* parents sought solutions to certain problems of living life online suggests that there is a delicate balance between participating in research to benefit society and fearing negative repercussions of such participation. The identification of such a delicate balance motivates our present focus on understanding sensitive user experiences both "in the wild" (i.e., when using Instagram) and "in the laboratory" (i.e., when engaging with sensitive Instagram data in a research context).

Further, while prior work [9] emphasized practices for youth protection when designing more than minimal risk research studies, it was conducted as a formative evaluation of participants' needs when designing a high-risk, high-reward research study. In contrast, our study goes a step beyond this in the research process by conducting retrospective interviews with participants about their lived experience of engaging in high-risk, high-reward research to inform better practices for managing the adverse effects of such participation. As such, by recognizing and accounting for the potential trauma of reflecting upon one's own online behaviors, our work contributes to the literature by providing insights into the context of conducting sensitive research with youth private interactions on social media. We make such contributions primarily through the emergent lens of "trauma-informed" research [27, 50].

2.2 Researchers' Welfare in Studies of Sensitive Topics

Institutional Review Boards (IRB) or traditional forms of governance for research, mostly do not consider questions about research ethics and protection of researchers especially for sensitive research. There is a growing body of research acknowledging the sensitivity of conducting research that involves exposure to traumatic information [18]. Specifically, previous research has investigated the potential adverse impact of being involved in such research on the researchers' mental well-being. For example, Mckenzie et al. [41] interviewed eight research assistants who collected and worked on self-harm and suicide attempts clinical notes. They found that research assistants experienced a wide range of challenges when undertaking such research, which includes being emotionally or psychologically unprepared for the level of detail and the sensitivity of the information in the records, being personally drawn into individual stories, and feeling emotional exhaustion from the cumulative effect of processing the data over a long time.

Therefore, researching sensitive topics requires well-documented and well-designed guidelines to protect the researchers' well-being – similar to guidelines for health professionals [20]. For instance, Vidgen et al. [78] shared their developed guidelines for researchers to minimize the harmful mental effects that might be caused by cumulative exposure to viewing and annotating online abusive content [78]. Although these guidelines might be useful to mitigate the impacts of working on such content, these guidelines were researchers' efforts that might not apply to other types of content such as self-harm records or other projects. Therefore, there has been a high demand for universities, funding institutions, and academic leaders to expand the research ethical considerations to protect the researchers and assess the potential risks on their mental well-being [26] – a direction that is emerging and evolving, but with limited evidence to support the design of requisite guidelines. There have not been investigations specifically for the perspectives of research assistants and their mental health when annotating youth's social media conversations for online risks. Our work addresses these gaps.

2.3 Developing Best Practices for Trauma-Informed Research for HCI

Traumatic experiences are difficult to define [19], yet they are very common. (Roughly 70% of people experience trauma at least once in their lives [45].) Broadly, trauma refers to a wide variety of negative experiences, including the experience of abuse and/or violence (physical or emotional) [19]. Chen et al [18] provided the following definition: "physical, emotional, or psychological harm." Yet, taking temporality into account, "trauma" logically refers to: (i) the experience of harm; and (ii) the fallout related to negative experiences (i.e., the ongoing recurrence of negative affect related to "traumatic experiences") [66]. As the computing fields develop frameworks for human-centered research that account for "humans" as more than mere "users" or "data points" [16, 64], it is essential to highlight affect and emotionality not only in the design and deployment of novel technologies, but in the modes of knowledge production that inform such design and deployment.

Trauma-informed approaches have been extensively discussed, developed, and implemented in the fields of mental health and addiction [37], nursing [70], clinical psychology [45], and education and schools [73]. Recently many institutions, organizations, and researchers have explored the implementation of trauma-informed approaches. Such approaches include considerations for participants who have experienced forms of trauma. Trauma-informed best practices are intended to minimize potential harm through heightened sensitivity to traumatic experiences [27, 50]. For example, the Substance Abuse and Mental Health Services Administration (SAMHSA) [34] created a trauma-informed framework which has been used across domains and sectors. Nursing and medical researchers [70] summarized available instruments to examine trauma-informed care (TIC) services and they identified the domains investigated (e.g. policy, mission, impact, etc.) and the populations considered (e.g. providers serving children in welfare system, health and human service organization staff, and survivors of domestic violence). They also have developed an evaluation tool for adolescent health and service users to evaluate the implementation of TIC [70]. There are similar efforts for integrating traumatic stress care into the health care of children and adolescents from pediatric organizations, including the National Child Traumatic Stress Network (NCTIC)[1] and the American Academy of Pediatrics (AAP) [75]. NCTIC defined four R's of a TIC organization as a program which "*realizes* the widespread impact of trauma and understands potential paths for recovery; *recognizes* the signs and symptoms of trauma in stakeholders; and *responds* by fully integrating knowledge about trauma into policies, procedures, and practices and seeking to actively *resist retraumatization*."

In light of this backdrop, trauma-informed research [18, 27, 50] appears as a possibly useful tool in the ethics-oriented maturation of HCI, and the computing fields more broadly. As Chen et al [18] recently wrote, "As an orientation to research and practice, trauma-informed computing has to be an ongoing commitment to improving design processes and artifacts, rather than a one-time checklist or a set of specific techniques" (c.f. [35]). Part of such an "ongoing commitment" must include reflexive consideration of not only trauma facilitated by technology itself, but trauma facilitated through and by means of research practices. Such considerations translate to the adoption of "best practices," similar to the heuristic provided by Walker et al. [82]. Chen et al's notion of "trauma-informed computing" [18] was thus based on main principles in SAMHSA with adaptations to the computing fields defined as a commitment to enhance the life-cycle of building and maintaining technologies by recognizing trauma to avoid technology-related trauma including six principles: safety (i.e., "ensuring that people feel safe when using, designing, or otherwise interacting with technology") trust (i.e., "basis for security, dependability, and confidence in social relationships," peer support (i.e., "connecting with fellow trauma survivors as a vital part of healing and recovery") collaboration (i.e., "ensuring that trauma survivors are actively involved in decisions regarding their care and support") enablement (i.e., "facilitating and developing opportunities for people to fulfill their potential and to develop their own capacity [23]") and intersectionality (i.e., "a lens for understanding how people's lives are mediated by multiple interlocking forms of oppression"). In a recent study, Scott et al. [62] applied the six trauma-informed guiding principles to social media design and moderation, providing examples of how each principle could be enforced both offline and online. The authors argued that to achieve trauma-informed design and moderation, social media companies must implement trauma-informed values and practices (e.g., pre-aware, trauma-aware, trauma-sensitive, trauma-responsive, trauma-informed) overtime. There are promising, albeit limited, examples of HCI research that has incorporated trauma-informed approaches in distinct

[1] https://www.nctsn.org/resources/trauma-informed-integrated-care-children-and-families-healthcare-settings

contexts for adults (e.g. studying Intimate Partner Violence (IPV) [74], participants and organizers of a postpartum health hackathon [21], post-trauma self-regulation apps for people with intellectual and developmental disabilities [77], and the theorization of speculative vulnerability [66].) Further, while trauma-informed guiding principles are valuable, they only explored some examples of good practices (Table 2 in [18]) in application areas of UX Research and Design, Security and Privacy, Artificial Intelligence and Machine Learning, and Organizational Culture. They do not go as far as to suggest how to operationalize these principles in practice, especially in novel contexts, such as working with youth and their sensitive data. Moreover, due to a lack of a systems-based perspective, some types of HCI researchers need to draw upon trauma-informed approaches *across* multiple fields while designing, developing, and conducting research for and with varied stakeholders. Therefore, we investigated the effect of potentially traumatic experiences of youth on social media when conducting research regarding their online risks.

On a complementary note, the HCI research community has called for connecting sensitive research to trauma-informed approaches [33, 71]. For instance, recently HCI researchers [39] have started to look at integrating design research activities with the therapeutic process for children of trauma backgrounds. They designed reflective storytelling and co-design activities with therapists to create a framework for guiding the design of technologies that support safety, connection, and reflection in scaffolding social-emotional learning for children [39]. As far as social media and youth online interaction research and trauma-informed research practice are concerned, there is still much to be explored. Our study is the first to use a trauma-informed lens to inform best practices in HCI research with youth regarding their online risk experiences. More broadly, as the computing fields – particularly those which assert the centrality of the human in computing – devote increased attention to concepts and practices traditionally left to philosophy (e.g., practical ethics), we identify an opportunity to enfold more sensitive methods and practices into the culture of computing research. We contend that the development and deployment of more robust trauma-informed research practices in the computing fields is necessary in order to normalize a more holistic, and therefore genuinely "human-centered," approach to understanding the role that our designs and the research that supports them play in human experience.

3 METHODS

We performed a retrospective interview study, a technique where participants are asked to reflect on their past experiences [43]. We engaged with youth (13-21 years old) who recently participated in a research study [54] where they were asked to take a web-based survey regarding their social media usage and mental health, upload their Instagram data to our system, and then review and flag their direct message conversations for private interactions that made them or someone else feel uncomfortable or unsafe.

3.1 Background on Original Instagram Data Donation Study

The goal of the original study [54] was to improve adolescent online safety by creating an ecologically valid dataset for training machine learning models for youth online risk detection [2, 3, 53]. For data collection, we invited participants who met the inclusion criteria: 1) speaks English and lives in the United States; 2) is a current active Instagram user and has used Instagram for at least 3 months when they were 13-17 years old; 3) has at least 15 direct message conversations; 4) at least two of the direct conversations made them or someone else feel uncomfortable or unsafe; and 5) the participants are willing to share their Instagram data with us for research purpose. After enrollment in the study, and filling out a survey about their online risk experiences, they were asked to download their Instagram data file to upload it to our secure online system. Since more than half (72%) of adolescents use Instagram, making it one of the most popular social media sites for young people, we chose Instagram as the platform of preference [7].

Then, participants were instructed to review their Direct Messages (DM) and select any messages that made them feel "uncomfortable" or "unsafe" while marking all other conversations as "safe." For every unsafe conversation flagged, the participants then annotated their data at the message level for risk level and risk type. Our risk categories were based on Instagram's reporting functionality.[2] Although in prior work [54], we presented participants with pre-defined risk types, we also stressed that risky interactions were not restricted to these categories and that they should self-assess the circumstances that made them feel uncomfortable or unsafe. We designed the first study based on the annotation of youth themselves on their conversations because including the viewpoints of individuals who have experienced online risks is essential for allowing the machine learning models to identify implicit indications of those risks [36, 56]. Participants were compensated for their time and data with a $50 Amazon gift card.

3.2 Interview Study Design

For the present study, we invited participants who started or completed the first part of the study. After the completion of the first part (initial study [54]), participants could opt-in to the second part of the study (interview study) by filling out an online form. We also reached out to the participants who started the first part but did not complete the study after two months to see if they were interested in participating in the interview study. Our aim for interviewing them was to find out the reason why they did not complete the study. The first author was the lead in designing, deploying, and data collection for the initial study, and most of the co-authors were involved in the process. In designing our study, we used a trauma-informed perspective, drawing broadly from what trauma-informed means. Specifically, we drew from the CDC [3] and SAMHSA's [34] six principles of trauma-informed care (1. Safety, 2. Trustworthiness and Transparency, 3. Peer Support, 4. Collaboration and Mutuality, 5. Empowerment Voice and Choice, 6. Cultural, historical and gender issues). Examples of this approach are taking all precautions to preserve their privacy, providing help resources during the study, and asking participants retrospectively how the study made them feel. More specifically, we designed a semi-structured interview script based on how participants felt participating in the original study, reviewing unsafe/risky conversations, and how that affected them. Interview questions are listed in Appendix B. We asked follow-up questions to clarify interesting discussion points during the conversations.

Then we asked participants who completed the study if they would be comfortable with us sharing our screen to show them the conversations they flagged for risks to remind them of those interactions. For those who discontinued involvement in our prior research, we asked why they did not continue the study and what we could have done better. We also interviewed RAs who annotated participant's data as third-parties (please refer to Appendix B for interview questions). We conducted interviews over a 30-minute-long scheduled Zoom session. The interviews averaged 25 minutes with a minimum of 20 minutes and a maximum of 40 minutes. We had a risk mitigation plan in place to ensure the safety of the participants. We incentivized participation with a $20 Amazon gift card distributed to the participant upon completion of the interview.

3.3 Participant Information

We conducted a total of 30 interviews with youth participants, of which 22 participants completed the first part of the study and passed the eligibility requirement, but 8 of them did not continue to

[2] https://www.facebook.com/help/instagram/192435014247952

[3] https://www.cdc.gov/orr/infographics/6_principles_trauma_info.htm

upload and flag their Instagram data. Table 3 in the appendix presents the demographics of the participants and their risk flagging information. We conducted a χ^2 test of independence (i.e.,a between-group analyses [68]) to examine any differences between the participants who passed the eligibility requirement versus those who discontinued the study. We found that there were no significant differences yielded between these two groups based on sex ($p - value = 0.17$), age ($p - value = 0.06$), and race ($p - value = 0.50$). In addition to interviews with data-donating youth participants, we conducted 12 interviews with research assistants (RAs) who helped us annotate the youth data and agreed to participate in our interview study. All of the RAs were undergraduate students in their 20's. Such RAs were mentored by three faculty and four Ph.D. students specializing in the field of online safety. Please refer to Table 4 for RAs' demographics. We held training workshops for RAs on how to handle such sensitive data based on our data management plan, security audit from our institutions, and child-mandated abuse reporting instructions. In addition to providing guidelines for annotations and definitions of risk types and levels, we encouraged open discussions regarding any concerns related to the annotations and resolved disagreements through consensus during weekly group discussions. We explained to RAs that the goal of their annotations is to have ground truth for creating training datasets for building machine learning models to detect online risks automatically. We anticipated that undergrads were uniquely qualified to review the data, given their proximal age range to youth and the training they received. RAs had a wide range of backgrounds, from computer science to criminology, sociology, and psychology.

3.3.1 Ethical Considerations. It was crucial to protect the participants' anonymity, privacy, and security because the dataset from the first part comprised extremely sensitive data. Our institution's IRB approved both parts of the study. The participants over 18 years old were required to fill out an adult consent form while participants under 18 years old required parental consent and teens' assent before participating. In the consent and assent forms, we included information about the research, research process, potential benefits, and risks of participating in this research. Additionally, participants were informed about what information would be collected, how it would be stored and protected, and how their data would be used in research.

We informed the participants and RAs of the researchers' mandated reporter status for Child Sexual Abuse Material (CSAM) and our duty to report imminent dangers. We provided clear warnings to participants in the first part of the study to avoid uploading any digital images that contained a minor's nudity as well as detailed instructions on how to delete such data before uploading data to our system. But in case RAs found such material, they had instructions to report them as soon as possible to the officials with assistance from the PIs. In order to further protect participant privacy, we additionally secured a National Institute of Health Certificate of Confidentiality. We instructed the RAs on safety precautions when completing data annotations such as avoiding using any cloud-based services or copying data on personal computers, and limiting data storage to devices that have been approved by and are secure by the institution. Specifically, we developed an annotation tool that the RAs were required to use. Moreover, we offered practical advice to support the well-being of the RAs, such as taking enough breaks because some of the information could be upsetting or graphic. We also provided "Help Resources" (such as helplines for suicide prevention and sexual victim assistance) on a webpage that was presented to our participants and RAs. We encouraged both groups to reach out to us with questions regarding their concerns and online safety. The RAs were given access to past research from the lab regarding online safety and we communicated our availability for consultation.

3.4 Qualitative Data Analysis Approach

Interviews were recorded and transcribed verbatim. We used thematic analysis [13] to develop systematic understanding of the interview data. After preliminary data coding, the first author discussed the initial coding with the rest of the research team to refine and finalize the codebook. Then, the first author coded all the interview transcripts with frequent check-ins by the last author to validate the codes. The coauthors reviewed the consistency of the codes iteratively throughout the data analysis phase. We conducted a thematic analysis of emergent themes. Some of the codes were not mutually exclusive and were double-coded (including codes in Table 1 Mixed_Feelings and Negative_Emotions; Contribute_Research, Personal_Interest, Incentive, Only_for_Research_Purpose, Keep_private, Do_not_Sell_Data, Aggregate_Data; and codes in Table 2), which led to percentages for each theme that would add to more than 100%.

3.4.1 Participant Interviews Qualitative Coding (RQ1). The codebook for participants (displayed in Table 1) included their motivation to participate in the study (receiving incentive, their interest in the topic, and contributing to the research), how they envisioned their data to be used/not be used (data be kept private/confidential, only be used for purpose of research, and not be sold to make a profit), and the emotions they experienced while flagging the data as we categorized into general categories based on their description of their feelings (positive, negative, mixed, neutral). It also included what they learned based on the study and reviewing their conversations (changed the way they communicate, changed their social media use and habits, reflected more on people's intentions and their response), and/or the reason participants discontinued (technical issues, privacy issues).

3.4.2 Research Assistants' Interviews Qualitative Coding (RQ2). Next, the code book for RA participants (displayed in Table 2) included the reasons they got surprised (frequency of online risks, personal conversations and debates, how risks escalated/evolved by sending more media), their concerns (reporting abuse/illegal material, were distressed, felt uncomfortable), learned (ground-truth importance for machine learning, advised people they know, become more privacy aware, made personal reflections, gained more knowledge about online risks, had positive thoughts), and how to provide support they needed (know mentors are there to help and guide, be strategic and tell them not to take conversations personally/take breaks occasionally, motivate them by emphasizing the importance of their work, have workshops and assign smaller sets of conversation, improve annotation tool).

4 RESULTS

In this section, we answer each research question by presenting major themes identified our during analyses. The major themes that emerged with the frequency of codes are presented in Table 1 for participants' and Table 2 for research assistants' interviews.

Table 1. Participants' Codebook. Total participants: $n = 30$ ($n = 22$ completed and $n = 8$ discontinued the initial study).
Note: Percentages calculated out of $n = 22$ completed are marked with ** and out of $n = 8$ discontinued marked with *

Themes	**Codes (Count, Percentage)**	**Example**
Surprise over accessibility and durability of historical data.	Download_Ability ($n = 15$, 50%)	*"Nothing really surprised me except the part where we actually took the data from Instagram. That's the only part that was pretty shocking because I didn't know you can do that"* P13
Discomfort from engaging with memories brought up by Instagram data.	Mixed_Feelings ($n = 12$, 55%)**	*"It definitely brought up some of the emotions that I had felt during the conversation, but it also made me it helped me process it again, I thought to myself oh Maybe this wasn't quite as bad as I thought"* P17
	Negative_Emotions ($n = 11$, 50%)** (Discomfort ($n = 13$, 61%)), Weird ($n = 6$, 27%), Upsetting($n = 5$, 23%), Angry ($n = 5$, 24%))**	*"I guess I realized that like there were times when I felt uncomfortable but then I sort of ignored that feeling and just kept texting person"*P5
Increased awareness from reflection on historical data	Communication_Improvement ($n = 11$, 37%)	*"If someone you know it's like threatening, they're harassing you just like block them instead of like replying"* P29
	Social_Media_Use ($n = 11$, 37%)	*"Make me a bit more cautious about how I should use social media"* P13
	Reflections_Dynamics ($n = 8$, 26%)	*"I'll make sure to understand who they are, why are they messaging me, what do they expect from me, that's what I'll do more"* P27
Willingness to help online safety with high expectations for safeguarding their privacy	Contribute_Research ($n = 18$, 60%)	*"I enjoy in helping out universities with their studies"*P19
	Personal_Interest ($n = 13$, 43%)	*"I wanted to help out and I know a lot of people that have struggled with their mental health due to social media online"* P27
	Incentive ($n = 8$, 27%)	*"I did not have a job, and needed some extra money. Happy to help research and to receive compensation"* P16
	Only_for_Research_Purpose ($n = 21$, 70%)	*"I assume my data is being used for researchers to see how teens are and impacted by social media"* P29
	Keep_Private ($n = 14$, 47%)	*"I do not want my data being shared outside the group of researchers"* P5
	Do_not_Sell_Data ($n = 4$, 13%)	*"I do not want my data to be used for advertising"* P26
	Aggregate_Data ($n = 7$, 23%)	*"I would hope the data conversations are being summarized instead of individualized"* P9
Discontinued Participants had technical difficulties with some privacy concerns	Tech_Difficulties ($n = 5$, 63%)*	*"I did not complete the Instagram data upload; the upload would never finish"* P24
	Privacy_Concerns ($n = 2$, 25%)*	*"Did not complete the study, because I felt the Instagram data upload was mildly intrusive"* P7
	Insufficient_Data ($n = 1$, 12%)*	*"I haven't had much uncomfortable conversations"* P21

Table 2. Research Assistants' Codebook

Themes	Code (Subcode)	Example
Surprised to see the types of risks teens are exposed to, but did not cause emotional distress	Surprised_Frequency_Risks (n = 7, 58%)	*"Really sad that it was happening so frequently with so many people."* RA6
	Uncomfortable (n = 4, 33%)	*"uncomfortable .... for the occasional meme that was suggestive that would be sent in group chats between friends."* RA1
	Sad (n = 2, 17%)	*"People getting so many bad messages from strangers that affect them negatively. It breaks my heart. It was crazy."* RA6
	Concerned_Reporting (n = 2, 17%)	*"Only concern was finding something that I would have to report, not sure what the emotional consequences would be like. "* RA2
Learned more about online risks made them reflect on their own past experience, privacy, and online safety	Reflected (n = 7, 58%)	*"Yeah, I thought some of the ways that I might have hurt someone either on Social Media or in person"* RA10
	Privacy_Awareness (n = 6, 50%)	*"I deleted social media. I saw all the things that happened, and how much time it takes "* RA10
	Experienced_Risks (n = 6, 50%)	*"It gave me a different perspective because given context something can be risky or not. Younger guys use SM as an escape from real life"* RA12
	Gave_Advice (n = 4, 33%)	*"I'll tell my friends if you get a message from a stranger that doesn't have a pic, just block them and don't answer."* RA6
	Positive_Impacts (n = 3, 8%))	*"Overall, a positive mental health impact, you can see how similar people are, and how they think the same things as you."* RA1
Research team support was crucial	Support (n = 6, 50%)	*" Reiterate that it's okay if you need a break from annotation tool... Important if one person's messages is too much, it can be reassigned to someone else ... "* RA2

4.1 Youths' Experience in Participating in Sensitive Research (RQ1)

Youths' experiences of reflecting on past interactions on Instagram that made them feel uncomfortable or unsafe included several themes (presented in Table 1), which we unpack in detail in the sections that follow.

4.1.1 Surprise over accessibility and durability of historical data. Half of the youths we interviewed (n=15, 50%) were surprised that they could even download and share their Instagram data for the purpose of research. Our participants were intrigued by the idea that their Instagram was portable and could be reviewed in the form of XML files downloaded to their computers. Their surprise was further rooted in their ability to see the chat histories that they had long forgotten as our web-based system presented these conversations back to them for risk-flagging. P30 provides a representative quote, highlighting the opacity of Instagram's data archiving practices:

> *"I didn't know that you could download your data for Instagram. After downloading actually I spent a long time looking at my data, because I'm like wow I thought these things weren't there anymore, it was pretty cool. I think I learned a lot from it, because I*

> *didn't even know the comments were documented. Because, on the app itself there's no way for you to look at your past comments, or it's not something that they were saved, but then I was like here all the comments are."* P30

Surprise over the durability of their Instagram data was also coupled with concern. Participants realized that such data could fall into the hands of third parties in the future: *"I think if someone hacked into your account, they could download all that data and have all that data"* (P2). Participants further indicated surprise over the frequency of their historical engagement with Instagram, as well as surprise over the presence of chats they did not consider as unsafe at the time such chats occurred. Collectively, the results described above indicate youths' ambiguous understanding of Instagram data durability, how the use of the platform may have future repercussions, and where responsibility for protecting against undesirable data-driven outcomes should reside.

4.1.2 Discomfort from engaging with memories brought up by Instagram data. For the youths in our study, their Instagram data were like windows into their pasts, which often brought up fond memories that allowed them to reminisce about events that they had otherwise forgotten:

> "*It [tagging historical Instagram data] was kind of like going to the past, like looking at old pictures. So it did bring back some old times. It also did bring back some good feelings, as well as like the good memories of that time, so it kind of just like to come back in time to that place or that point in my life."* P2

Yet, for most participants, engaging with the past through data was also uncomfortable, as P2 continued: *"But it definitely did make me a little bit uncomfortable coming back to those messages."*

Youths described a variety of feelings arising from the process of flagging chats in their Instagram data. They frequently described mixed feelings (n = 12, or 55%) and negative emotions (n = 11, or 50%) as results of tagging and reflecting on old chats in their Instagram histories. More than half of the participants (n = 13, 61%) specifically expressed "discomfort" related to engaging with old memories and the reflective process of actively considering them. Some of this discomfort stemmed from developing a new perspective on past interactions. P5, for example, regretted continued engagement in chats that they now realize they should have ignored or terminated:

> *"Reflecting on messages made me feel uncomfortable because while I was rereading the messages, there were instances that I felt uncomfortable at the time, but I kept engaging in the conversation instead of ignoring them. I have negative feelings looking back at my messages, nothing could be done on your side. "* P5

Some of the discomforts were less about reflecting on the data but instead about their past actions from when they were younger and more naive. For instance, P10's experience of tagging her historical Instagram data was marked by mild regret, indicating that she would like to have acted differently in past chats:

> *"I felt a little regret reviewing these past messages, wish I could have taken back the things I said. It wasn't too negative an experience, just realizing what I should do in these types of situations in the future. My perspective of a few of the conversations changed after reviewing them. I realized I should be better about reporting and blocking risky messages immediately instead of responding."* P10

Yet, some participants (n = 5, or 23%) indicated that negative emotions resulting from engagement with their historical data exceeded mere discomfort. Participants indicated being "upset" (P22,P28), "worried" (P29,P15), and "nervous" (P29).[4] The root of such negative affective experiences was

[4] One participant (P29) even requested that the interviewer not share their screen with them so as to avoid being reminded of unsafe chats.

complex. For example, P29 described a worry that her old Instagram data could negatively impact her in the future, as well as her present mental health:

> *"It made me feel like very nervous and maybe worried because even though those messages were a long time ago, they could still have an impact on you somehow maybe in the future. Just reading your own messages and reflecting on yourself could have maybe an impact on your mental health. You read messages and you're like 'Oh, I said this in the past, what was I thinking?"'* P29

Such reflections often made participants feel worse about themselves and the adverse situations that they experienced when they were younger. A few participants felt "angry" ($n = 5$, 24%), because reflecting on unpleasant interactions made them revisit situations in which they felt powerless:

> *"Made me feel uncomfortable, angry, and powerless, not super pleasant to go through, it happened a long time ago and it brought up stuff and didn't make me feel that great."* P11

Some participants ($n = 6$, 27%) mentioned they felt "weird" while reviewing their historical DMs. Such weirdness was tied to the content of specific chats and compounded participants' sense of being powerless against the emotional impacts of revisiting unsafe chats. Yet, those feelings were mostly because of the nature of the interactions that happened in the past, and participants said that there was nothing that we could do better to alleviate those feelings. In line with the six principles of the Trauma-Informed Framework [34], this theme was associated most closely with the *Safety* principle. The negative emotions, including the discomfort relate to the psychological safety of using Instagram and participating in the study itself.

4.1.3 Increased awareness from reflection on historical data. On the positive side, many participants ($n = 11$, or 37%) said that reflection had helped them improve the way they communicated with others. In reviewing and flagging their past conversations, participants assessed and reflected on how they dealt with different situations and people. Such improvement included being more attentive to phrasing and subtext. Participants further indicated that reflection helped them understand their friendships and how they deemed people to be trustworthy. P1 provided an example:

> *"I think it was probably a beneficial experience just in terms of scrolling and evaluating past relationships in terms of looking at why did I hang out with this person, how did I react fast, but also, it was kind of a weird feeling. But I think it's largely beneficial to see things that I had sent a couple of years ago, not necessarily because they were explicitly bad, but just because of the way that I communicate. I think it has changed, it was just interesting to see the difference."* P1

In an unexpected finding, eleven participants (37%) indicated that active reflection about their historical Instagram data encouraged them to decrease their current social media use. Participants mentioned how they used to respond to every message, but after reviewing those past interactions they learned that they could simply stop responding and ignore more messages:

> *"After looking at all of the unsafe messages, I realize probably have been best if I just didn't respond at all. Because they can't really do anything if I don't respond, but if I do then it's just giving them the attention that they wanted in the first place."* P22

Further, participants stated that engagement in this research project made them more cautious about giving information to people on social media. It taught them to be more mindful about who they add and which large groups are worth joining. Notably, some participants indicated that participation in this research encouraged them to avoid joining large groups altogether. Yet, participants acknowledged that such caution on Instagram comes with trade-offs. Consider the following quote from P17:

> *"I guess to evaluate, is it worth putting myself in that uncomfortable potentially threatening situations for the sake of my freedom of speech, and I decided the answer was yes."* P17

Here P17 describes how exercising caution in her Instagram interactions impacts her right to express her opinions freely. Some participants, on the other hand, identified features of Instagram as means to find the balance between caution and expression. For instance, P10 indicated that reporting posts and blocking users are valuable alternatives to self-censorship or withdrawal from expression on the platform:

> *"It wasn't too negative an experience, just realizing what I should do in these types of situations in the future. My perspective of a few of the conversations changed after reviewing them. I realized I should be better about reporting and blocking risky messages immediately instead of responding."* P10

Similarly, participants expressed their appreciation of the features that Instagram provides, such as showing DMs of people who you do not follow in a separate tab. In some ways, they learned to use the features of Instagram in a way to safeguard themselves from uncomfortable or unsafe situations. Such features mediated expression and communication on Instagram in a beneficial way:

> *"Well, I'm much older now, but I think you learn as you age how to use implemented features. If someone sends you a DM, it doesn't show it to you unless you accept the message. So you can just block that person right away, and block any new accounts they make. I think that's very smart of them, but I do like the idea of if someone makes you uncomfortable literally just block them that's what I do now."* P18

Youths also acknowledged the value of making their social media accounts private as a way of mediating their online social interactions. They mentioned recent changes they made in their social media behaviors, such as deleting more posts and being more concerned about their privacy. They discussed how they use privacy features on Instagram now compared to before for instance only sharing Instagram stories with only close friends:

> *"I'm careful with posting a picture of myself now, or just people I tag and stuff like that. I believe I have my Instagram set to private as well. I used to post a lot of pictures of myself until that one girl started saying stuff about me. I barely post pictures of myself anymore on social media. And if I do it either has a filter on it now, or I'll post it as an Instagram story that only all my close friends in real life can see."* P19

After flagging their conversations for risks, youths also reflected more on the dynamics of interactions online and people's intentions and how they respond to them ($n = 8$, 27%). For instance, participants said they would spend more effort understanding who the people they engaged with were, why they were messaging them, and discerning how to respond. Participants also thought about how to be clearer about what they want from other people and how to articulate that more effectively:

> *"I definitely think about what I send on the Internet more. I make sure that people are going to interpret. I send the way that I want them to interpret it and if I don't think they will, then I'll clarify differently."* P14

They reflected on their feelings toward conversations and mentioned how important it was to validate their own feelings toward a conversation, rather than ignoring them. By being more self-aware, they could have responded differently:

> *"Probably just like know that my feelings towards the conversation are valid and not like probably shouldn't ignore them and just continue the conversation. Like I probably should*

> *have been like "Hmm why do I feel uncomfortable?" instead of just continuing to talk to them."* P5

In sum, P6 expressed how the study helped her recall situations *"that [were] kinda messed up"* online, which made her more self-aware and able to learn from her past mistakes. This sentiment was shared among most of our participants who were a bit older and wiser than when they first started engaging with others via Instagram. The *Empowerment* principle of the Trauma-Informed Framework [34] aligned with the perceptions of participants' increased awareness and resolve toward taking different actions in the future resulting from their participation in the original study.

4.1.4 Willingness to help online safety with high expectations for safeguarding their privacy. The youths in our study recounted experiences of growing up alongside social media. Having seen first-hand the potential toxicity of social media (SM), youths discussed their willingness to participate in our research study in hopes of improving SM conditions for other youths. The majority of youths (n = 18, 60%) in our study were motivated by contributing to research on adolescent online safety or they were interested in the subject (n = 13, 43%). Additionally, some participants were motivated to receive the incentive (n = 8, 27%). Some participants (n = 4, 13%) explicitly mentioned that they do not want their data to be sold to third parties, be used for advertisement, or be used for "profit purposes" (P12). They expressed interest in topics related to SM and understanding social patterns such as how important it is to solve online safety issues for adolescents and how they would be glad to contribute to research in this area. For instance, P2 stated that although it took substantial amount of time to complete the study, contributing to the research was worth the time:

> *"I'm glad to participate but took longer than expected, I believe it was worth it. It was worth it because it made it possible for my data to be used for research."* P2

Based on consent/assent forms statements, most participants (n = 21, 70%) envisioned their data being used solely for research and the provision of insights for solving problems. Participants had different viewpoints on how their data could be used for research purposes. They discussed how their data could be used to provide statistics and trends about interactions in SM and be used as evidence of unsafe interactions happening. Then the youth SM data would be used to improve their experience and make online safe spaces for teens. Moreover, research could demonstrate how youth use online spaces and how harassment and other risk types happen. It could also be used to tell companies that they need to do a better job at keeping youth safe on their platforms:

> *"It'd be something like telling different companies hey you know you guys just aren't doing this right and that so we need some things to stop. Overall just make social media very much better experience for people my age."* P24

Almost half of the participants (n = 14, 47%) expressed how important it is to keep their data private and confidential (e.g., *"as long as my name isn't plastered everywhere."* (P6)) by de-identification and anonymization techniques to protect their personal information from being disseminated outside of the research team or shared publicly. Participants expressed that they trusted researchers from accredited universities, as long as their de-identified data was solely used for research purposes. Further, all participants mentioned they do not regret participating in the study and expressed their expectations on how their data could be used to create online safety solutions to prevent risky interactions. Youth (n = 7, 23%) did not want their data to be investigated individually (*"do not track what I do"*(P25)) and wanted their data to be used to make aggregated insights such as presenting *"Mass data analysis"*(P5) or ' '*summarized conversations"*(P9).

Some noted other concerns such as how they imagined their data being kept secure (physically), how they would like the results of the study being shared with them, or how they are worried about other people's information involved in the conversations, more than their own privacy:

> *"To be completely honest I don't know how ethical it was to be sharing my own data when it was also other people's conversations. I obviously didn't have a reliable way to contact all 500 of them, but I trusted you're going to use the data in a safe way, but I wasn't totally sure about that and that's the point of consideration that we had at first, we started the study. I would not want any of the information to be identifiable to the account that sent them to me I don't mind if I am identified, but I don't want that specifically to be published, or the specific words."* P17

Finally, for privacy and legal purposes, we provided instructions on how to delete content from their data file before uploading. Participants cited other reasons, as to why they deleted some high-risk data, for instance, the data their parents would "kill"(P16) them for having, but they mentioned how they "wanted to put those in"(P19). The *Trust* principle of the Trauma-Informed Framework was most relevant when understanding participants' willingness to contribute to sensitive research, but only with the high expectation that the researchers would protect their privacy.

4.1.5 Discontinued Participants had technical difficulties with some privacy concerns. Some (5 out of 8) of the participants in our interviews who discontinued the study did so because of upload or technical issues. They could complete the first part of the study on any device (e.g., mobile device), but in order to upload their Instagram data they needed a desktop computer. The reason is the limited space on mobile devices and difficulty to store the file, browse, and upload it:

> *"I did not complete study Phone space was limited Have Chrome book from school; the site was blocked on my computer. I do not have the technological availability to complete the study."* P4

Some participants who discontinued the study ($n = 2$) expressed privacy concerns related to sharing their private messages. For instance, P7 expressed how much he wanted to contribute to this important research, but it is privacy intrusive to him to upload all his Instagram data and someone potentially read all his conversations:

> *"I just personally found it to be like mildly intrusive. However, everything else, it's for a good cause and that's important. Privacy is a very broad term but also, I think, it's just the ability to keep things confidential. I'm sure you don't really like it when people look through your texts and stuff like that, and I know that's the point of this study; however, to relinquish so much privacy in a quick file send, that concept is frankly intimidating. I prioritize not anonymity, but my right to self whenever I'm online and interacting with others, and for that reason I found it a bit uncomfortable to bargain with that topic."* P7

Although P7 suggested that if we had only asked him to upload his uncomfortable conversations not all of his Instagram data, he would complete the study. He also suggested that reviewing unsafe interactions with a conductor's assistance would make the research more controlled and it would make it easier.

No participants who discontinued participation in our study attributed discontinuation to negative affective experiences. For instance, P8 mentioned that they would have deleted the most uncomfortable content, such that reviewing conversations would not be triggering for them:

> *"I would have deleted anything any conversation that made me feel really uncomfortable. So I probably wouldn't have been dealing with really triggering material because I wouldn't have it. So I don't think it would have made me feel too negatively. Definitely, I don't think I would have that many remaining because if they upset me that much I would have blocked the person and deleted it.* P8

Overall, most of the participants who did not complete the study had technical difficulties that should be minimized and it was not because of privacy issues. For participants with privacy concerns, other options such as uploading partial messages should be provided. Similar to the previous subsection, this theme also mapped most closely to the *Trust* principle of the Trauma-Informed Framework [34]. In order to ensure trustworthy computing, it's essential to ensure that technology artifacts, processes, and organizations function transparently, predictably, and reliably, while also giving users the ability to make mistakes and correct them, if necessary [38].

4.2 Researchers' Experience Analyzing Sensitive Youth Data (RQ2)

Below, we provide insights on the experiences of research assistants (RAs) who annotated the data (themes and codes presented in Table 2).

4.2.1 Surprised to see the types of risks teens are exposed to, but did not cause emotional distress. Most (n = 7, 58%) of the RAs in our project were surprised to see frequently youths encounter online risks. Such surprise, however, was not emotionally distressing to RAs. RAs mentioned surprise at teens communicating about mature topics inappropriate to their age. Some RAs (n = 4, 33%) mentioned that some messages or media made them uncomfortable. RAs explained that most negative feelings associated with reviewing conversations were because they "could not do anything to help" (RA7) the situations. However, such feelings were transient:

> *"I guess a lot of the creeps were guys messaging girls. How frequent it was and the things they would say surprised me. That stuff that was like really grossed me out in the moment. Nothing like that hurt me mentally or physically that I carried on."* RA10

Some topics were more disturbing to RAs than others. Such topics included as sexually suggestive media, discussions about sexual orientation, or adults (e.g., grown males) trying to seduce teenage girls.RAs used some strategies for those uncomfortable feelings such as moving through those faster as RA8 mentioned:

> *"Sometimes when people send images that were kind of graphic memes that would make me uncomfortable! I just tried to I guess get through those conversations faster, so I can move on to another conversation."* RA8

In a few cases (n = 2, 17%), RAs were worried if they find something that they have to report (e.g., child pornography). Less frequently (n = 2, 17%), RAs mentioned that they were sad to see how frequent the unsafe interactions were that youth deal with, though, the sad feelings were only negative feelings that disappeared after a short amount of time:

> *"The only messages that were triggering only lingered in my mind for about 10 minutes; mostly related to sexuality. Bothered by messages that insulted or judged someone based on their sexuality. Sad to see that kids are growing up believing these things."* RA11

Overall, RAs were surprised by the frequency with which teens are exposed to online risks. At times it was unpleasant for RAs to review those interactions, but it only affected them at the times they were reviewing those messages and they did not carry the negative feelings with them. To some extent, this relates to the *Safety* principle of the trauma-informed Framework, as RA's were concerned about reporting imminent dangers to participants or child abuse materials, while also trying to maintain their own psychological safety.

4.2.2 Learning more about online risks made them reflect on their own past experience, privacy, and online safety. Most of the RAs (n = 7, 58%) reflected about how the annotation task affected their perspectives. They reflected on how social media is used for friendships, online social dynamics, and social connections, and how those features inherently bring more vulnerabilities. Consider the following quote from RA7:

> *"I thought more about how risky it is for other people. Sometimes there is a sense that you have more followers the better, then you accept more people, but you don't really know them and they are looking at your content and start making your friends their friends,.. it is dangerous sometimes."* RA7

The annotation task not only made RAs think about the situations where they (or other people) were vulnerable, but also made them reflect if they ever hurt someone. Annotating social media conversations of youth for online safety made half (n = 6, 50%) of the research assistants more privacy-aware. Some limited the use of social media or removed their accounts from their lives. They became more cognizant of who their friends and connections are and removed the people that they do not know:

> *"I became more aware of who is on my social media. I removed anybody that I didn't know."* RA11

The task made RAs more cautious of online safety, especially toward younger friends and family. Some RAs (n=4, 33%) reported giving advice to their younger siblings, relatives, or friends:

> *"I was among the first people on messaging apps and experienced it for the first time. More innocent at the time. Dialogue online was very useful and safe, but now it's become dangerous. It worried me about my younger sister to stay away!... I am using social media less now, this task confirmed to use it less... I recommend other people to stay away too like my sister."* RA9

Similarly, other RAs also compared their own experiences to what youth experience online and mentioned they had less risky interactions compared to what is happening online.

A few RAs (n = 3, 8%) mentioned that the annotations had positive impacts on them when they reflected on their past experiences. They also felt they have positive impacts by helping to detect online risks for youth:

> *"Made me reflect on how I felt the same way that these teen participants did struggle with their mental health, and now life is much better. Somewhat a positive experience."* RA2

This theme mapped to the *Empowerment* principle of the Trauma-Informed Framework by enabling RAs to gain more insights into online risks which made them reflect more on their own experiences and made them give advice to people they care about.

4.2.3 Research team support was crucial. RAs felt that most of their feedback was consistently addressed or incorporated. They (n = 6, 50%) noted the importance of the support, motivational conversations, and general encouragement from their mentors and the team. They mentioned that some conversations were tricky and unclear to annotate, but the team members were always there to consult on confusing cases. They provided invaluable feedback to improve the process such as taking frequent breaks, reminding them of the "bigger picture" of the project to motivate them, or as RA6 suggested providing strategies to RAs to not take messages personally, and facilitate opportunities for them to work together:

> *"Make sure annotators are not taking those messages personally or close to their heart. Give a chance to the annotators for their task to be assigned to another annotator, have small workshops once a month and tell them what you are doing is good and encourage people. How much value it has for."* RA6

Broadly, this theme mapped to the Peer Support principle of the Trauma-Informed Framework. The support and tools that the mentors provide for RAs are of crucial importance which can alleviate some of the negative experiences resulted from the nature of reviewing unpleasant online interactions.

5 DISCUSSION

5.1 Applying a Trauma-Informed Lens to Youth Experiences

A key emergent finding from our study is that social media research has the potential to re-traumatize participants [62]; and in our case, youth who experienced negative interactions online. As human-subjects research grows to include digitally mediated experiences, protocols need to be updated to reflect new contexts. Such updating should include the normalization of trauma-informed research practices. While Chen et al.'s work provides valuable insights for trauma-informed evaluation of *computing systems* based on the SAMHSA's [34] six dimensions of trauma-informed care, our work focuses on developing best practices in HCI for trauma-informed *computing research* with youth. Therefore, we reflect on the strengths of and lessons learned from our study as it relates to research practices aligning with the six dimensions of trauma-informed care:

- *Safety:* In terms of digital safety, we ensured that the system for the study was secure by provisioning additional security audits from a special security team. Also, we provided procedures in place in terms of child abuse reporting, confidentiality of their data, and being able to delete their data they do not want to share. However, what we learned in terms of safety is that we needed to take adequate care of participants' mental well-being while participating in our study.
- *Trust:* We learned from our interviews that participants trusted us as representatives of a large university to donate their data. Researchers should work to build trust with participants by taking additional steps, such as acquiring the National Institutes of Health (NIH)'s Certificate of Confidentiality, to protect identifiable research information from forced disclosure such as parents asking about their teen's social media activities.
- *Peer Support:* Sharing lived experiences can help trauma survivors in their healing journey [27]. Research assistants also mentioned the importance of this by having group workshops to discuss and share when annotating potential traumatic experiences.
- *Collaboration:* Our studies ensured that potential trauma survivors are actively collaborating in the development of new technologies and their voices are being heard in decision making. Youth in our study experienced online risks, however, we did not specifically have people who experienced trauma help inform the design of the study. In the future researchers should involve youth who encountered serious risks such as suicidal ideation or sexual solicitation to inform the research. Taking into account the perspectives of youth for designing research and technologies is a way to ensure their opinions are valued and integrated.
- *Enablement:* In studies of sensitive topics, it is essential to give more control to youth to enable them to make decisions that are best for them. For instance, giving more options to partially donate their data or remove the data they do not want to share with researchers is a necessity. In the future, researchers should also involve youth in more generative processes toward creating solutions such as co-design. As youth are sensitive and may not be willing to provide all their data, participants should have "alternative paths" [10] to participate in a study and only share data they are most comfortable with.
- *Intersectionality:* In our study, we had a representative sample of youth, including a large sample of LGBTQ+ youth participants. Yet, we also need to consider diversity in other terms such as neurodiversity, physical ability, etc. As youth's trauma experiences are greatly intertwined with their identity, researchers need to consider identity from different perspectives and at various social *power relations* [52].

Although we confirmed that our initial study aligned with the recommendations for trauma-informed computing, we also uncovered several gaps where these guidelines were general and therefore insufficient in understanding and anticipating some of the trauma-informed guidance that we should have provided to our youth participants in research practices. Therefore, we make the following novel recommendations for trauma-informed research in HCI with youth, especially when engaging them in reflective exercises involving their sensitive social media data.

5.1.1 Using a Trauma-Informed Lens to Anticipate Challenges for HCI Research with Youth. With specific focus on youth as a vulnerable population, we amend the six dimensions of trauma-informed computing Chen et al. [18]. An important difference between working with youth versus adults is that there is a potential for more trauma as participating in this type of research could potentially be related to mandated child abuse. According to some states' laws, a teen could be removed from their household and be placed in child welfare if they are found to be sexting. As researchers, we need to minimize the chance of youth being traumatized by creating trauma-sensitive procedures. For instance, there should be procedures about what to do if there is an imminent risk versus a risk that has happened in the past.

We make several recommendations for ethical HCI research practices in studies involving youth's re-engagement with potentially traumatic online experiences based on our results:

- *Anticipate knowledge gaps.* Youth were surprised by the accessibility and durability of their historical data. Hence an implication for trauma-informed computing is that there needs to be more education and awareness around our digitized lives and our digital footprints to help reduce vulnerability to speculative forms of digitized trauma [42, 66]. Therefore, researchers need to provide training/education to reduce participants' level of surprise or clearly describe the data prior to asking them to download and/or share it.
- *Anticipate emotional vulnerabilities.* We found that participants experienced discomfort from engaging with memories brought up, which might be different than what is traumatic for adults due to the unique developmental stage of adolescence. Therefore, as researchers, we need not to assume that what would be traumatic to us would be the same for our participants. Also, participants should have access to contact information for crisis support and help resources available at all times. Pursuant to the level of risk faced by participants, we recommend that debriefing sessions could be facilitated between participants and trained mental health professionals (involvement of mental health professionals has been adopted in HCI researchers [85]), such that any negative emotional outcomes of research can be addressed productively. At the very least, risky studies should include opportunities for participants to take a break if they are uncomfortable. Second, we recommend continued contact with study participants in the form of transparent communication about how their participation in research benefited others. For example, setting up a website to catalog useful resources and including the results from the study. Notably, such continued contact would need to receive ethical approval.
- *Anticipate potential ways of working through trauma as part of research.* We found that youth awareness increased from reflection on historical data which shows promising impacts on them. For example, many LGBTQ+ youth self-selected to participate in our study, implying that they want to share their experiences. The research community should not underestimate the value of giving youth a platform to talk about their negative online experiences. As we found that participants' reflections increased their privacy awareness and changed their social media habits, we could streamline the study into a learning opportunity for them by providing them with more training, such as training about managing one's privacy or interpersonal boundaries on social media as a way of supporting youth from feeling powerless. In the mental health space, researchers [44, 61] have created interventions based on self-reflection and investigated ways that self-reflections may help people change their behaviors. This

research shows that aside from supporting mental health, self-reflections could be helpful for youth online risk behavior.

- *Anticipate youth's needs for empowerment.* Since youth are willing to help with online safety with high expectations for safeguarding their privacy, researchers should anticipate youth's unique developmental needs for empowerment when contributing to research. Youth empowerment to have impacts on research programs should be based on justice-centered design (JCD) principles to overcome societal inequities [17]. The HCI research community has been giving the survivors of traumatic events (e.g. studying sex workers [72], domestic abuse survivors [28, 74]) an empowering platform to tell their stories. We cannot refrain from conducting research about sensitive topics and should not treat studying traumatizing subjects as taboo simply because there might be a risk of re-traumatization. Such an approach might unintentionally take away survivors' voices in trying to heal from these experiences. But most trauma-informed approaches have been applied in HCI research in different contexts for adults. For instance, as mentioned in Section 2, a trauma-informed approach has been applied in the context of IPV [5], which tends to be a topic more focused on adults than youth.
- *Anticipate providing hands-on technical assistance.* We often refer to youth as "digital natives," [11] which implies that they have been immersed in technology, but we need to question whether they have the necessary technical skills. Most participants who discontinued did so because of technical difficulties. Researchers should not assume that young people possess the technical skills required to overcome technical issues on their own. It is therefore important to consider these issues when designing online studies for youth. Where possible, then, we recommend the inclusion of live help sessions for participants in online studies, particularly when such studies are not completed under the aegis of platforms like Amazon Mechanical Turk, Qualtrics, etc. We further recommend that despite the ease of conducting online studies, researchers – junior and senior alike – maintain vigilance about basic, epistemologically-grounded practices. Data are always already intermediary (see Drucker [24] on the concept of "capta"). Yet, data are obligatory passage points between daily life and participation in large sociotechnical structures, even when interactions with data-driven infrastructures are sources of discomfort [67] and deceptively limited empowerment [65]. To study data in and of itself is increasingly a proxy for studying *people* in a digital world. But studying data is not the same as studying people. As data-driven research moves researchers away from direct engagement with people and toward conducting big data analyses, interacting with people directly gets edged out. Yet, people must remain central in research design and practice.

Future researchers should study how digital artifacts and researching those artifacts can produce trauma and the appropriate ways to educate youth on how to maintain and manage (e.g. young adults naturally practice retrospective impression management on their Facebook posts and connected users such as alternating past content [60]) this potential trauma. Our work is an effort toward taking a trauma-informed approach to research in HCI for youth. It is imperative that the HCI community continue this line of research to examine methods for interacting with youth trauma survivors ethically and practically and to create trauma-informed frameworks for research with youth. In addition to informing trauma-based research, this leads to a new critical HCI community research agenda on helping youth deal with digital forms of trauma.

[5] https://www.ipvtechresearch.org/research

5.2 Applying a Trauma-Informed Lens with College-Age Research Assistants

Overall, we did not find evidence that RAs were traumatized by participating in the analysis described in this paper. The reason may be because they were reflecting on the risky experiences of youth, rather than their own. RAs were given the chance to opt-in to data analysis, so this probably lessened the risks, since they knew what they were getting into.

Moreover, there might be some stigma related to expressing emotions, especially in highly technical fields such as Computer Science. Taboo lingers in such academic arenas to maintain professionalism in research and exposing emotions is considered going against professionalism – an untenable position given the role that personal experience plays in such knowledge production techniques as thematic analysis [13]. Thus it is even more important to provide a safe environment for researchers to express their personal experiences. We encourage HCI researchers to utilize strategies from social science researchers and other fields that work more with human subjects. These strategies for researcher self-care include personal self-assessment of emotional risk factors, emotional proximity, and distance, physical health and wellbeing, mental time-out, social support, and enabling environment [79]. It is worth noting that such strategies may challenge the temporalities of publishing in Computer Science and the computing fields – higher levels of care toward RAs and participants may stretch the periods of time required to produce meaningful *and ethically tenable* research artifacts. Nevertheless, trauma-informed frameworks [30, 58, 81] should be integrated for researchers when studying topics of a sensitive nature. Such implementation may consist of teaching self-care strategies to help overcome emotional fatigue; such strategies should be embodied by lab leaders and senior researchers. We provide the following implications from our results for taking a trauma-informed approach with research assistants

- *Anticipate the need for training.* We found that RAs were surprised to see the frequency and the types of risks teens are exposed to which made some emotional discomfort but that did not cause emotional distress for them. However, if the sensitive nature of the data were more triggering with regard to RAs' lived experiences (e.g., rape, pregnancy loss, etc.), then we would have likely seen more negative responses. This should also be disclosed in the write-up of the research. Therefore, research teams need to have well-thought-out training materials. Also, it needs to be considered whether the population annotating the data is well-suited for the task. In our case, college-aged undergrads were a good fit, rather than adults, because they are generationally closer to the participants' age.
- *Anticipate personal benefits to the researcher.* We found that RAs learned more about online risks which made them reflect on their own past experience, privacy, and online experience. Engagement with participant data allowed RAs opportunities for meaningful self-reflection. Such opportunities for self-reflection empower RAs to feel they are "Scholar activists," [86] who can make a positive impact by helping online safety for youth. Each sensitive research is unique, and challenges and emotional distress related to it might be different in nature [79]. So before starting a research process, a well-being care plan should be prepared to allow researchers to express any fatigue or trauma that may be experienced and ways to overcome it. When conducting sensitive research this question could be asked:"From the human-subjects perspective, could we consider having a consent form for researchers?" One of the useful protecting measures we could have performed to protect researchers before they accept to work on a project is to have a consent form for researchers, explaining the potential benefits and harms in annotating data and conducting the research. Moreover, the RAs who participated in our study were given the instructions that the annotations they complete will be used for training machine learning models to detect youth online risks. This may have influenced how they are handling the potentially upsetting content of such sensitive data.

- *Anticipate inclusion.* Aligned with our finding that the support from the research team was crucial, we need to make sure that the RAs know how much their work is important and appreciated. It is imperative to provide an environment for RAs to mitigate any negative impact on them. We found that providing help and feedback from research mentors is crucial to support the RAs.

Recently, researchers have used sensitive or intimate data in their research processes and actively involved participants (data donors, crowd workers, or citizen scientists) in the process as a way to get a more contextual understanding of people and their behaviors. Researchers and participants in these activities engage in highly personal exploration, unlike those in data science. This paper contributes to an emerging body of HCI research that extends beyond challenges related to data privacy and sharing. In addition, these processes allow for the generation of awareness and the exchange of value through an understanding of the dynamics that play out when it comes to what is sensitive or intimate data.

5.3 Implications from the Retrospection of Youth on Their Social Media Data

The implications of our research go beyond trauma-informed research to suggest possible ways to reduce trauma through social media platforms. We found that youth were surprised that user data were stored in a long-term way. Such surprise was cause for discomfort. For the youth who participated in our study, the permanence and opacity of Instagram data implied a non-obvious form of trauma: the discomfort of confronting and reconciling oneself with data traces that represent immature actions taken in earlier developmental stages of life.

When engaging with data representing forgotten interactions, the youth in our study described various levels of discomfort about how they acted (or failed to act) online in the "old times" (P2). Immature actions appropriate to teenage development (e.g., bickering), are rendered potentially harmful in the future precisely because of their permanence. Such findings speak to the need to explore the future-facing, preventative functions of trauma-informed computing.

That data about forgotten interactions could be downloaded long after such interactions had passed from user memory highlights the need to understand online risks at multiple timescales, not only from multiple perspectives. The functional permanence of Instagram data means that old data can be made new again at any time. That is, such data – like trauma – may resurface. Combined with anxiety about how data traces might be viewed, interpreted, or used in youth's futures, the disparity between what is clearly visible to the user (i.e., recent interactions) and what may be available to interested parties (e.g., researchers, hackers, message recipients, data-brokers, and their customers) implicates platforms like Instagram in bringing about "speculative vulnerabilities" [66]. Such speculative vulnerabilities constitute sites for future data-oriented trauma. By identifying speculative vulnerabilities that arise because of the time-based characteristics of data (i.e., the fact that such data may be recalled long after youth have outgrown the behaviors such data represent), researchers and practitioners may be able to help users avoid the trauma of encountering personally deprecated data traces (i.e., records of behaviors that one has long since outgrown). Thus, from our findings we provide our first implication. In creating what are essentially archives of interactions [63] stored for future analysis, should the "need" for such analysis arise [12], researchers have an emergent commitment to protect participants from past tense interactions to minimize potential retraumatization.

Many youth participants demonstrated personal growth through their reflections on past chats. They used mechanisms provided by platforms (e.g., blocking) to manage their social interactions. For instance, youth in our study frequently discussed cleaning their data. By "cleaning," they meant removing data that they found to be embarrassing. The same youth appeared to want Instagram to take more responsibility for protecting against harmful interactions on the platform. As such, it may be inferred that: (1) engagement with social media platforms normalizes digitally-oriented social behaviors in non-digital social spaces; and (2) individual reactions to data-represented pasts are functions of the individual at the time of reaction. We recommend that Instagram (and similar social media platforms) accept greater responsibility for users by developing and deploying policies limiting the durability of user data in social media – particularly data derived from users yet to achieve legal majority (i.e., youth). Such limitation would align with European standards already in place (e.g., the right to be forgotten [59]), but would also constitute a colorable variation on canonical privacy literature in the United States: Warren and Brandeis's "right to be left alone" [83]. To "be left alone" requires redefinition as people's subjective experience of their worlds extends further into the arcane and powerful machinery of platforms and digital institutions. While the discourse of privacy took a turn towards "control" in the mid-twentieth century [4] and subsequently the maintenance of contextual integrity [46], the current characteristics of app culture [67] create a problematic synergy that hints at speculative forms of use-based trauma [66]. Such synergy emerges among "control," "context," and what is known as "responsibilization." Responsibilization refers to the placing of responsibilities previously held by institutions (e.g., companies, universities, etc.) onto individuals [22, 49, 51]. In the present context, we use "responsibilization" to account for participant feelings of being responsible for "cleaning" embarrassing data post hoc. The responsibility for solving the problems of such synergy should not fall to users, let alone youth; nor should problems created by the control-context-responsibility triad go unaddressed as they create futures in which today's youth are inherently vulnerable *because* of their historical data. By accounting for the synergy among control, context, and responsibilization, trauma-informed computing may do more than mitigate the negative impacts of prior trauma: it may prevent future data- and use-related trauma by letting one's data disappear from memory in the same way trivial daily interactions may disappear from memory.

5.4 Limitations and Future Research

The findings we present here come with several limitations that can be addressed in future work. Given the dynamics between researchers and participants, it is likely that our data contains effects of social desirability bias and recall bias [76]. Most of our participants were motivated to participate in research and were interested in online safety topics, so we may not have captured the opinions of everyone (e.g., an effect of self-selection bias). We also encountered a surprising amount of participant attrition, often due to technical issues. Participants who were susceptible to leaving the study because of technical difficulties constitute a meaningful population to study in future work. Youth relationships with their data traces evolve over time. As such, longitudinal work is required to better situate our findings in time. Generally, the first study was constrained by its context based on social media risks, so more data from trauma-informed studies in other settings would be beneficial. Finally, our study results have potential implications in a broader range of contexts where the population might not seem at risk initially. Future HCI research dealing with personal information need to take trauma-informed approaches, even if the research topic does not appear to be sensitive initially (e.g. co-monitoring using smart-home IoT; such technology might be beneficial to most but harmful to partners of intimate violence. Frequently critical reports on research stakeholders' experiences (e.g. participants and researchers) are not documented through publication. Future research should go beyond just conducting the research to complete a retrospective interrogation of the research. Also, our research occurred in parallel with other trauma-informed works, such as Chen et al. and Scott et al. [18, 34, 62] , Therefore, while these works did not directly inform our study design, for future research, it would be helpful to leverage these frameworks or others when applicable, and more generally, trauma-informed principles to inform future study designs when engaging with vulnerable users and dealing with sensitive topics. A primary goal of HCI research is to study and design technology in ways that benefit people; therefore, we believe that incorporating evidence-based principles of trauma-informed care into our research practices is a logical step in the right direction.

6 CONCLUSION

At times, the nature of our HCI research may have potential risks of re-traumatization to the human subjects in which we study. With that realization comes the responsibility to continually develop and integrate trauma-informed best practices into the critical research that we conduct, especially when it involves vulnerable populations, such as youth. We believe that the HCI community is up for this challenge.

ACKNOWLEDGMENTS

This research is supported in part by the U.S. National Science Foundation under grants #IIP-2329976, #IIS-2333207, #CNS-1942610 and by the William T. Grant Foundation grant #187941. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsors. We would also like to thank all the participants who donated their data and contributed to the research.

REFERENCES

[1] Zainab Agha, Neeraj Chatlani, Afsaneh Razi, and Pamela Wisniewski. 2020. Towards Conducting Responsible Research with Teens and Parents Regarding Online Risks. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3383073

[2] Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Joshua Gracie, Munmun De Choudhury, Pamela J. Wisniewski, and Gianluca Stringhini. 2022. 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* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 148, 14 pages. https://doi.org/10.1145/3491102.3501969

[3] Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Chen Ling, Munmun De Choudhury, Pamela J. Wisniewski, and Gianluca Stringhini. 2023. Getting Meta: A Multimodal Approach for Detecting Unsafe Conversations within Instagram Direct Messages of Youth. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 132 (apr 2023), 30 pages. https://doi.org/10.1145/3579608

[4] Irwin Altman. 1976. A conceptual analysis. *Environment and behavior* 8, 1 (1976), 7–29.

[5] Nazanin Andalibi and Andrea Forte. 2015. Social computing researchers as vulnerable populations. In *ACM Conference on Computer Supported Cooperative Work & Social Computing Workshop on Ethics for Studying Sociotechnical Systems in a Big Data World.*

[6] Nazanin Andalibi and Andrea Forte. 2018. Announcing Pregnancy Loss on Facebook: A Decision-Making Framework for Stigmatized Disclosures on Identified Social Network Sites. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173732

[7] Monica Anderson and Jingjing Jiang. 2018. Teens, Social Media & Technology 2018 | Pew Research Center. http://www.pewinternet.org/2018/05/31/teens-social-media-technology-2018/

[8] Alissa N. Antle. 2017. The ethics of doing research with vulnerable populations. *interactions* 24, 6 (Oct. 2017), 74–77. https://doi.org/10.1145/3137107

[9] Karla Badillo-Urquiola, Zachary Shea, Zainab Agha, Irina Lediaeva, and Pamela Wisniewski. 2021. Conducting Risky Research with Teens: Co-designing for the Ethical Treatment and Protection of Adolescents. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–46.

[10] Steve Benford, Chris Greenhalgh, Gabriella Giannachi, Brendan Walker, Joe Marshall, and Tom Rodden. 2012. Uncomfortable Interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*

(Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2005–2014. https://doi.org/10.1145/2207676.2208347

[11] Sue Bennett, Karl Maton, and Lisa Kervin. 2008. The 'digital natives' debate: A critical review of the evidence. *British journal of educational technology* 39, 5 (2008), 775–786.

[12] Geoffrey C. Bowker. 2008. *Memory Practices in the Sciences.* MIT Press, Cambridge, MA.

[13] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* American Psychological Association, Washington, DC, US, 57–71. https://doi.org/10.1037/13620-004

[14] Jennie Carroll, Steve Howard, Frank Vetere, Jane Peck, and John Murphy. 2002. Just what do the youth of today want? Technology appropriation by young people. In *Proceedings of the 35th annual Hawaii international conference on system sciences.* IEEE, 1777–1785.

[15] Robey B. Champine, Jason M. Lang, Ashley M. Nelson, Rochelle F. Hanson, and Jacob K. Tebes. 2019. Systems Measures of a Trauma-Informed Approach: A Systematic Review. *American Journal of Community Psychology* 64, 3-4 (2019), 418–437. https://doi.org/10.1002/ajcp.12388 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajcp.12388

[16] Stevie Chancellor, Eric P. S. Baumer, and Munmun De Choudhury. 2019. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 147 (Nov. 2019), 32 pages. https://doi.org/10.1145/3359249

[17] Neeraj Chatlani, Arianna Davis, Karla Badillo-Urquiola, Elizabeth Bonsignore, and Pamela Wisniewski. 2022. Teen as research-apprentice: A restorative justice approach for centering adolescents as the authority of their own online safety. *International Journal of Child-Computer Interaction* (2022), 100549.

[18] Janet X. Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. 2022. Trauma-Informed Computing: Towards Safer Technology Experiences for All. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 544, 20 pages. https://doi.org/10.1145/3491102.3517475

[19] Constance J Dalenberg, Elizabeth Straus, and Eve B Carlson. 2017. Defining trauma. In *APA handbook of trauma psychology: Foundations in knowledge, Vol. 1.* American Psychological Association, 15–33.

[20] Virginia Dickson-Swift, Erica L James, Sandra Kippen, and Pranee Liamputtong. 2006. Blurring boundaries in qualitative health research on sensitive topics. *Qualitative health research* 16, 6 (2006), 853–871.

[21] Catherine D'Ignazio, Rebecca Michelson, Alexis Hope, Josephine Hoy, Jennifer Roberts, and Kate Krontiris. 2020. "The Personal is Political": Hackathons as Feminist Consciousness Raising. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 150 (oct 2020), 23 pages. https://doi.org/10.1145/3415221

[22] Niall Docherty and Asia J. Biega. 2022. (Re)Politicizing Digital Well-Being: Beyond User Engagements. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 573, 13 pages. https://doi.org/10.1145/3491102.3501857

[23] Lynn Dombrowski, Ellie Harmon, and Sarah Fox. 2016. Social justice-oriented interaction design: Outlining key design strategies and commitments. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems.* 656–671.

[24] Johanna Drucker. 2014. *Graphesis: Visual forms of Knowledge Production.* Harvard University Press, Cambridge, MA.

[25] Holly R Farley. 2020. Assessing mental health in vulnerable adolescents. *Nursing2020* 50, 10 (2020), 48–53.

[26] Jessica L. Feuston, Arpita Bhattacharya, Nazanin Andalibi, Elizabeth A. Ankrah, Sheena Erete, Mark Handel, Wendy Moncur, Sarah Vieweg, and Jed R. Brubaker. 2022. Researcher Wellbeing and Best Practices in Emotionally Demanding Research. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 72, 6 pages. https://doi.org/10.1145/3491101.3503742

[27] Center for Substance Abuse Treatment et al. 2014. Trauma-informed care in behavioral health services. https://www.ncbi.nlm.nih.gov/books/NBK207201/

[28] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. 2018. "A Stalker's Paradise" How Intimate Partner Abusers Exploit Technology. In *Proceedings of the 2018 CHI conference on human factors in computing systems.* 1–13.

[29] Nicolas E Gold, Raul Masu, Cecile Chevalier, and Fabio Morreale. 2022. Share Your Values! Community-Driven Embedding of Ethics in Research. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 12, 7 pages. https://doi.org/10.1145/3491101.3516389

[30] Jordan Goodwin and Emmy Tiderington. 2020. Building trauma-informed research competencies in social work education. *Social Work Education* 41, 2 (2020), 1–14.

[31] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. Safe Sexting: The Advice and Support Adolescents Receive from Peers Regarding Online Sexual Risks. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 42 (apr 2021),

31 pages. https://doi.org/10.1145/3449116

[32] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. 'If You Care About Me, You'll Send Me a Pic'-Examining the Role of Peer Pressure in Adolescent Sexting. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 67–71.

[33] Juan Pablo Hourcade, Alissa N Antle, Michail Giannakos, Jerry Alan Fails, Janet C Read, Panos Markopoulos, Franca Garzotto, and Andrea Palumbos. 2019. Child-computer interaction sig: Designing for refugee children. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–4.

[34] Larke N. Huang, Rebecca Flatow, Tenly Biggs, Sara Afayee, Kelley Smith, Thomas Clark, and Mary Blake. 2014. SAMHSA's Concept of Truama and Guidance for a Trauma-Informed Approach. http://hdl.handle.net/10713/18559 This manual introduces a concept of trauma and offers a framework for becoming a trauma-informed organization, system, or service sector. The manual provides a definition of trauma and a trauma-informed approach, and offers 6 key principles and 10 implementation domains..

[35] Shannon Kelly, Benjamin Lauren, and Kaitlyn Nguyen. 2021. Trauma-Informed Web Heuristics for Communication Designers. In *Proceedings of the 39th ACM International Conference on Design of Communication* (Virtual Event, USA) *(SIGDOC '21)*. Association for Computing Machinery, New York, NY, USA, 172–176. https://doi.org/10.1145/3472714.3473638

[36] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela Wisniewski, and Munmun De Choudhury. 2021. You Don't Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[37] Maritt Kirst, Anjana Aery, Flora I Matheson, and Vicky Stergiopoulos. 2017. Provider and consumer perceptions of trauma informed practices and services for substance use and mental health problems. *International Journal of Mental Health and Addiction* 15, 3 (2017), 514–528.

[38] Bran Knowles, Mark Rouncefield, Mike Harding, Nigel Davies, Lynne Blair, James Hannon, John Walden, and Ding Wang. 2015. Models and patterns of trust. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 328–338.

[39] Lian Loke, Aaron Blishen, Carl Gray, and Naseem Ahmadpour. 2021. Safety, Connection and Reflection: Designing with Therapists for Children with Serious Emotional Behaviour Issues. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 162, 17 pages. https://doi.org/10.1145/3411764.3445178

[40] Michael Massimi, Wendy Moncur, William Odom, Richard Banks, and David Kirk. 2012. Memento Mori: Technology Design for the End of Life. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI EA '12)*. Association for Computing Machinery, New York, NY, USA, 2759–2762. https://doi.org/10.1145/2212776.2212714

[41] Sarah K Mckenzie, Cissy Li, Gabrielle Jenkin, and Sunny Collings. 2017. Ethical considerations in sensitive suicide research reliant on non-clinical researchers. *Research ethics* 13, 3-4 (2017), 173–183.

[42] Anna Menyhért. 2020. Trauma studies in the digital age. In *The Routledge Companion to Literature and Trauma*. Routledge, 241–256.

[43] Sandra Metts, Susan Sprecher, and William R Cupach. 1991. Retrospective self-reports. In *Studying interpersonal interaction*, B. M. Montgomery & S. Duck (Ed.). Guilford Press, New York, NY, 162–178.

[44] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2018. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* 52, 6 (2018), 446–462.

[45] Yuval Neria, Evelyn J Bromet, Sylvia Sievers, Janet Lavelle, and Laura J Fochtmann. 2002. Trauma exposure and posttraumatic stress disorder in psychosis: findings from a first-admission cohort. *Journal of consulting and Clinical Psychology* 70, 1 (2002), 246.

[46] H. Nissenbaum. 2009. *Privacy in Context: Technology, Policy, and the Integrity of Social Life.* Stanford University Press. https://books.google.com/books?id=_NN1uGn1Jd8C

[47] World Health Organization. 2021. *Adolescence mental health.* Retrieved 2022-12-11 from https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health/

[48] Jessica Pater and Elizabeth Mynatt. 2017. Defining Digital Self-Harm. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1501–1513. https://doi.org/10.1145/2998181.2998224

[49] Lucy Pei, Benedict Salazar Olgado, and Roderic Crooks. 2022. Narrativity, Audience, Legitimacy: Data Practices of Community Organizers. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 328, 6 pages. https://doi.org/10.1145/3491101.3519673

[50] Kate Portman-Thompson. 2020. Implementing trauma-informed care in mental health services. *Mental Health Practice* 23, 3 (2020).

[51] Jarkko Pyysiäinen, Darren Halpin, and Andrew Guilfoyle. 2017. Neoliberal governance and 'responsibilization'of agents: reassessing the mechanisms of responsibility-shift in neoliberal discursive environments. *Distinktion: Journal of Social Theory* 18, 2 (2017), 215–235.

[52] Yolanda A Rankin and Jakita O Thomas. 2019. Straighten up and fly right: Rethinking intersectionality in HCI research. *Interactions* 26, 6 (2019), 64–68.

[53] Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2023. 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, Article 89 (apr 2023), 29 pages. https://doi.org/10.1145/3579522

[54] Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Nurun Naher, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2022. Instagram Data Donation: A Case Study on Collecting Ecologically Valid Social Media Data for the Purpose of Adolescent Online Risk Detection. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 39, 9 pages. https://doi.org/10.1145/3491101.3503569

[55] Afsaneh Razi, Karla Badillo-Urquiola, and Pamela J. Wisniewski. 2020. Let's Talk about Sext: How Adolescents Seek Support and Advice about Their Online Sexual Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376400

[56] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J. Wisniewski. 2021. A Human-Centered Systematic Literature Review of the Computational Approaches for Online Sexual Risk Detection. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 465 (oct 2021), 38 pages. https://doi.org/10.1145/3479609

[57] Afsaneh Razi, Seunghyun Kim, Munmun De Choudhury, and Pamela Wisniewski. 2019. Ethical considerations for adolescent online risk detection AI systems. In *Good Systems: Ethical AI for CSCW (The 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing)*.

[58] Elizabeth Reeves. 2015. A synthesis of the literature on trauma-informed care. *Issues in mental health nursing* 36, 9 (2015), 698–709.

[59] Jeffrey Rosen. 2011. The right to be forgotten. *Stan. L. Rev. Online* 64 (2011), 88.

[60] Sarita Schoenebeck, Nicole B. Ellison, Lindsay Blackwell, Joseph B. Bayer, and Emily B. Falk. 2016. Playful Backstalking and Serious Impression Management: How Young Adults Reflect on Their Past Identities on Facebook. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1475–1487. https://doi.org/10.1145/2818048.2819923

[61] Stephen M Schueller, Martha Neary, Jocelyn Lai, and Daniel A Epstein. 2021. Understanding People's Use of and Perspectives on Mood-Tracking Apps: Interview Study. *JMIR mental health* 8, 8 (2021), e29368.

[62] Carol F Scott, Gabriela Marcu, Riana Elyse Anderson, Mark W Newman, and Sarita Schoenebeck. 2023. 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 341, 20 pages. https://doi.org/10.1145/3544548.3581512

[63] John S Seberger. 2021. Into the Archive of Ubiquitous Computing: The Data Perfect Tense and the Historicization of the Present. *Journal of Documentation* 71, 1 (2021), 18–37. https://doi.org/10.1108/JD-11-2020-0195

[64] John S Seberger. 2021. Reconsidering the user in IoT: the subjectivity of things. *Personal and Ubiquitous Computing* 25, 3 (2021), 525–533.

[65] John S. Seberger, Marissel Llavore, Nicholas Nye Wyant, Irina Shklovski, and Sameer Patil. 2021. Empowering Resignation: There's an App for That. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 552, 18 pages. https://doi.org/10.1145/3411764.3445293

[66] John S. Seberger, Ike Obi, Mariem Loukil, William Liao, David Wild, and Patil Sameer. 2022. Speculative Vulnerability: Uncovering the Temporalities of Vulnerability in People's Experiences of the Pandemic. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW, Article 17 (dec 2022), 20 pages. https://doi.org/InPress

[67] John S. Seberger, Irina Shklovski, Emily Swiatek, and Sameer Patil. 2022. Still Creepy After All These Years:The Normalization of Affective Discomfort in App Use. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 159, 19 pages. https://doi.org/10.1145/3491102.3502112

[68] Donald Sharpe. 2015. Chi-square test is statistically significant: Now what? *Practical Assessment, Research, and Evaluation* 20, 1 (2015), 8.

[69] Mary Cipriano Silva. 1995. Ethical guidelines in the conduct, dissemination, and implementation of nursing research. (1995).

[70] Laura Sinko, Dana Beck, and Julia Seng. 2022. Developing the TIC grade: a youth self-report measure of perceptions of trauma-informed care. *Journal of the American Psychiatric Nurses Association* 28, 6 (2022), 455–463.

[71] S Revi Sterling. 2013. Designing for trauma: the roles of ICTD in combating violence against women (VAW). In *Proceedings of the Sixth International Conference on Information and Communications Technologies and Development: Notes-Volume 2*. 159–162.

[72] Angelika Strohmayer, Jenn Clamen, and Mary Laing. 2019. Technologies for social justice: Lessons from sex workers on the front lines. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.

[73] M Shelley Thomas, Shantel Crosby, and Judi Vanderhaar. 2019. Trauma-informed practices in schools across two decades: An interdisciplinary review of research. *Review of Research in Education* 43, 1 (2019), 422–452.

[74] Emily Tseng, Diana Freed, Kristen Engel, Thomas Ristenpart, and Nicola Dell. 2021. A digital safety dilemma: Analysis of computer-mediated computer security interventions for intimate partner violence during COVID-19. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[75] Andre Tylee, Dagmar M Haller, Tanya Graham, Rachel Churchill, and Lena A Sanci. 2007. Youth-friendly primary-care services: how are we doing and what more needs to be done? *The Lancet* 369, 9572 (2007), 1565–1573.

[76] Thea F Van de Mortel. 2008. Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, The* 25, 4 (2008), 40–48.

[77] Krishna Venkatasubramanian and Tina-Marie Ranalli. 2022. Designing Post-Trauma Self-Regulation Apps for People with Intellectual and Developmental Disabilities. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–14.

[78] Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy, 80 – 93.

[79] Joanne Vincett. 2018. Researcher self-care in organizational ethnography: Lessons from overcoming compassion fatigue. *Journal of Organizational Ethnography* 7, 1 (2018), 44–58.

[80] Jessica Vitak, Katie Shilton, and Zahra Ashktorab. 2016. Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work an Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 941–953. https://doi.org/10.1145/2818048.2820078

[81] Laura A Voith, Tyrone Hamler, Meredith W Francis, Hyunjune Lee, and Amy Korsch-Williams. 2020. Using a trauma-informed, socially just research framework with marginalized populations: practices and barriers to implementation. *Social Work Research* 44, 3 (2020), 169–181.

[82] Ashley Marie Walker, Yaxing Yao, Christine Geeng, Roberto Hoyle, and Pamela Wisniewski. 2019. Moving beyond 'one size fits all': research considerations for working with vulnerable populations. *Interactions* 26, 6 (Oct. 2019), 34–39. https://doi.org/10.1145/3358904

[83] Samuel D Warren and Louis D Brandeis. 1890. Right to privacy. *Harv. L. Rev.* 4 (1890), 193.

[84] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, Daniel F. Perkins, and John M. Carroll. 2016. Dear Diary: Teens Reflect on Their Weekly Online Risk Experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3919–3930. https://doi.org/10.1145/2858036.2858317 event-place: San Jose, California, USA.

[85] Daisy Yoo, Odeth Kantengwa, Nick Logler, Reverien Interayamahanga, Joseph Nkurunziza, and Batya Friedman. 2018. Collaborative reflection: a practice for enriching research partnerships spanning culture, discipline, and time. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–11.

[86] Meg Young, PM Krafft, and Michael A Katell. 2021. A Call for Scholar Activism. *AI Activism* 28 (2021), 43.

A PARTICIPANT DEMOGRAPHICS TABLE

Table 3. Participants' Demographics

ID	Study Status	Age	Gender	Race	Unsafe Conversations
P1	Passed	19	Female	White/Caucasian,Prefer to Self-Identify	3
P2	Passed	18	Male	White/Caucasian	18
P3	Passed	15	Gender-fluid	Black/African-American	9
P4	Discontinued	18	Non-Binary	White/Caucasian	0
P5	Passed	18	Non-Binary	Asian or Pacific Islander	4
P6	Passed	18	Female	White/Caucasian,Black/African-American,Hispanic/Latino	2
P7	Discontinued	18	Male	Not Disclosed	0
P8	Discontinued	21	Non-Binary	White/Caucasian,Asian or Pacific Islander	0
P9	Discontinued	21	Female	Asian or Pacific Islander	0
P10	Passed	14	Female	Asian or Pacific Islander	14
P11	Passed	17	Male	White/Caucasian	32
P12	Passed	16	Female	Black/African-American	6
P13	Discontinued	19	Female	Asian or Pacific Islander	10
P14	Passed	15	Female	White/Caucasian	11
P15	Passed	18	Male	Black/African-American	4
P16	Passed	17	Female	White/Caucasian,Black/African-American	9
P17	Passed	18	Female	White/Caucasian,Asian or Pacific Islander	15
P18	Passed	21	Female	White/Caucasian	3
P19	Passed	16	Female	White/Caucasian	4
P20	Passed	14	Female	White/Caucasian	22
P21	Discontinued	20	Female	Black/African-American	2
P22	Passed	14	Female	Black/African-American,American Indian/Alaska Native	12
P23	Discontinued	17	Female	White/Caucasian,Asian or Pacific Islander	0
P24	Discontinued	16	Male	White/Caucasian,Hispanic/Latino	0
P25	Passed	15	Female	Black/African-American	26
P26	Passed	17	Male	White/Caucasian,Hispanic/Latino	16
P27	Passed	14	Female	Black/African-American,American Indian/Alaska Native	4
P28	Passed	16	Male	White/Caucasian	5
P29	Passed	17	Female	Hispanic/Latino	15
P30	Passed	20	Female	Asian or Pacific Islander	4

Table 4. RAs' Demographics

ID	Gender	Degree	Major	Race-Age
R1	Not-self identify	Undergraduate	Computer Science	Hispanic/Latino-NA
R2	Female	Undergraduate	Computer Science	Hispanic/ Latino-27
R3	Male	Undergraduate	Social Sciences	Hispanic/Latino-22
R4	Female	Undergraduate	Computer Science	Some other race-NA
R5	Male	Undergraduate	Computer Science	White-NA
R6	Female	Undergraduate	Information Technology	Asian-22
R7	Male	Undergraduate	Computer Science	Hispanic/ Latino-NA
R8	Female	MS Student	Computer Science	Asian-23
R9	Male	MS Student	Computer Science	Two or more races-25
R10	Male	Undergraduate	Computer Science	White-NA
R11	Female	Undergraduate	Psychology	White/Hispanic/Latino-22
R12	Male	Undergraduate	Computer Science	White-NA

B INTERVIEW QUESTIONS

B.0.1 Participants' Interview Questions.

- What motivated you to participate in the study?
- Did anything about the study surprise you? Please explain.
- Did anything about the study make you feel uncomfortable? Please explain.
- Was there any part of the study that was confusing or difficult for you to complete? Please explain.
- Was there any information that we should have asked for during the study that we did not ask for?
- I'm going to share my screen and show you some of the content you flagged for risk to ask you further questions about them. Is that okay?
- When you reviewed this conversation for the study, how did bringing up that memory make you feel? Please explain.
- (For participants that did not complete this part ask:) What was the reason that you did not complete flagging your unsafe conversations?
- Based on our explanation in the consent form, how do you envision your data being used? Are there any ways you would not want researchers to use your data?
- Overall, are you glad that you participated in this study, or do you regret having participated? Please explain.
- Based on your participation in this study, what have you learned or what might you do differently regarding your interactions with others on social media?
- For the ones who did not complete the study: At what point in the study did you decide not to continue? Why this part of the study made you hesitant to participate? Anything we could have done better? Why did you initially decide to participate in the study? (When they explained something they did not like ask:) Was it explained before in informed consent?

B.0.2 RAs' Interview Questions.

- As an annotator, can you please explain what your job responsibilities were?
- What were some of the insights you gained from annotating the youth Instagram data?
- Did anything about the annotation surprise you? Please explain.
- Did anything about the annotation make you feel uncomfortable? Please explain.
- Did you have any problems or concerns during the annotation process? Please explain. Is there anything that we could have done for you to address those issues?
- Have you personally experienced any of the uncomfortable or unsafe experiences the youth in our study encountered? Please explain to the extent that you are comfortable. Did this influence your data annotations in any way?
- Did the annotation task make you reflect and think about your past experiences?
- After the annotation, did anything change in the way that you use social media or Instagram?
- How could we have better supported you in your annotation role?

Received January 2023; revised July 2023; accepted November 2023