“Strike at the Root”: Co-designing Real-Time Social Media Interventions for Adolescent Online Risk Prevention

ZAINAB AGHA, Vanderbilt University, USA
KARLA BADILLO-URQUIOLA, University of Notre Dame, USA
PAMELA J. WISNIEWSKI, Vanderbilt University, USA

Adolescent online safety researchers have emphasized the importance of moving beyond restrictive and privacy invasive approaches to online safety, towards resilience-based approaches for empowering teens to deal with online risks independently. Unfortunately, many of the existing online safety interventions are focused on parental mediation and not contextualized to teens’ personal experiences online; thus, they do not effectively cater to the unique needs of teens. To better understand how we might design online safety interventions that help teens deal with online risks, as well as when and how to intervene, we must include teens as partners in the design process and equip them with the skills needed to contribute equally to the design process. As such, we conducted User Experience (UX) bootcamps with 21 teens (ages 13-17) to first teach them important UX design skills using industry standard tools, so they could create storyboards for unsafe online interactions commonly experienced by teens and high-fidelity, interactive prototypes for dealing with these situations. Based on their storyboards, teens often encountered information breaches and sexual risks with strangers, as well as cyberbullying from acquaintances or friends. While teens often blocked or reported strangers, they struggled with responding to risks from friends or acquaintances, seeking advice from others on the best action to take. Importantly, teens did not find any of the existing ways for responding to these risks to be effective in keeping them safe. When asked to create their own design-based interventions, teens frequently envisioned “nudges” that occurred in real-time. Interestingly, teens more often designed for risk prevention (rather than risk coping) by focusing on nudging the risk perpetrator (rather than the victim) to rethink their actions, block harmful actions from occurring, or penalizing perpetrators for inappropriate behavior to prevent it from happening again in the future. Teens also designed personalized sensitivity filters to provide teens the ability to manage content they wanted to see online. Some teens also designed personalized nudges, so that teens could receive intelligent, guided advice from the platform that would help them know how to handle online risks themselves without intervention from their parents. Our findings highlight how teens want to address online risks “at the root” by putting the onus of risk prevention on those who perpetrate them – rather than on the victim. Our work is the first to leverage co-design with teens to develop novel online safety interventions that advocate for a paradigm shift from youth risk protection to promoting good digital citizenship.

CCS Concepts: •**Human-centered computing~Human computer interaction (HCI)~Empirical studies in HCI**

Additional Key Words and Phrases: Adolescent Online Safety, User Experience, Co-Design, Participatory Design, Social Media, Design-Based Interventions, Nudges

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

2573-0142/2023/04 – Art 149 $15.00

© Copyright is held by the owner/author(s). Publication rights licensed to ACM.

https://doi.org/10.1145/3579625

ACM Reference format:
Zainab Agha, Karla Badillo-Urquiola, and Pamela J. Wisniewski. 2023. "Strike at the Root:" Co-designing Real-Time Social media Interventions for Adolescent Online Risk Prevention. *Proc. ACM Hum.-Comput. Interact,* 7, CSCW1, Article 149 (April 2023), 32 pages, https://doi.org/10.1145/3579625

1 INTRODUCTION

In today's digital age of constant connectivity, adolescent online safety has become an increasingly important topic. There is growing awareness on the types of risks posed to teens by social media, such as cyberbullying [36], sexual risks [40], exposure to explicit content [9], or information breaches [43]. As such, social media companies are under scrutiny for the online risks posed to teens on their platforms, leading to increased efforts for online safety, such as screen time limits [55], age verification restrictions [31], and parental controls [54,57]. At the same time, new regulations have been put into effect globally for protection of youth online, such as the KIDS Act [32] in the U.S. for ensuring that teens are protected from harmful content and manipulation online, and the Safety by Design initiative in the EU [56] for ensuring that teen-centric online safety features are at the core of online platforms.

Given the importance and urgency of promoting adolescent online safety, there have also been extensive efforts within the Human Computer Interaction (HCI) and SIGCHI communities dedicated towards understanding teens online risk experiences and the best ways to ensure their safety in developmentally appropriate ways. Research on this topic has confirmed the prevalence of online risks [14,30,50] and studied risk-aversive approaches to dealing with unsafe interactions, such as technology restrictions [29], mediation [24], or monitoring [23] of teens social media using parental controls [4]. However, teens are at a unique developmental stage, where they often find parental involvement to be privacy invasive and desire more autonomy in their online interactions [19,47]. Therefore, the narrative has shifted towards strength-based approaches to online safety that can empower teens to be resilient in the face of risks [49,58], with the help of resources and tools that can guide them towards safer online interactions [8,26,46]. While there is a significant body of research on understanding the importance of risk-coping and resilience on teens' online risk experiences [48,49,58], there is a lack of knowledge on *how* to design and implement online safety interventions to help teens build resilience and manage online risks. Most research focuses on the prevalence and negative consequences of online risk experiences on youth. In a comprehensive review of the adolescent online safety literature, Pinter et al. [38] found limited research on teen-centric solutions for mitigating online risks. We address this gap by moving beyond an understanding of resilience based approaches, towards designing actionable teen-centric solutions that empower youth to manage their own online safety.

It is challenging to design effective interventions that cater to teens' online safety needs without their participation as primary stakeholders. To design effective online safety interventions, we need a better understanding of teens lived online risk experiences and their ideas for how to solve this critical problem. Several researchers have emphasized on the importance of involving teens as partners in the process of co-designing for online safety solutions (e.g., [6,8,34]). However, previous co-design efforts for adolescent online safety have either focused on a single type of online risk (such as cyberbullying [5,17]), younger children (not teens) [6,34], or focused on redesigning parental controls [6,34]. While these works highlight the usefulness of co-design in this space, there is a need to better understand how teens themselves would design online safety interventions for the myriad of online risks that are personally relevant to them that go beyond parental controls. We address this gap by co-designing online safety interventions with teens that move beyond parental controls and that are contextualized to teens' diverse online risk experiences, including but not limited to cyberbullying. Therefore, we pose the following high-level research questions:

- **RQ1**: *a) What types of situations make teens feel uncomfortable or unsafe online? b) How do teens currently deal with these situations and are these approaches effective?*
- **RQ2**: *What design-based interventions do teens recommend for dealing with these online risk scenarios?*
- **RQ3**: *How do these solutions reflect and/or depart from the status quo of existing online safety practices and research?*

To answer these research questions, we designed and conducted nine User Experience (UX) Bootcamp sessions with 21 adolescents (ages 13-17) based in the United States. First, we taught them important UX design skills, so that they could create storyboards and high-fidelity, interactive prototypes for online safety interventions. In the process, we asked teens to share and design for their personal experiences when using social media, especially those that made them feel uncomfortable or unsafe online. Using the skills learned in the trainings, teens were guided to participate in three research activities: a) Creating storyboards for unsafe online interactions, b) Whiteboarding ideas for online safety solutions, and c) Developing high-fidelity prototypes for online safety interventions. Due to COVID-19, this study was conducted virtually via Zoom, using online UX tools (i.e., Canva, FigJam, and Figma) for each of the research activities, respectively.

Overall, teens most often described risk scenarios that included information breaches from strangers, sexually inappropriate messages from strangers, and cyberbullying and harassment from people they knew (RQ1a). Most of these risks happened in private conversations (e.g., Direct Messages) on Instagram. In responding to risks with strangers, teens used a combination of blocking, reporting, and deleting. In contrast, teens struggled with responding to risks from acquaintances and friends, often attempting to refuse their unwanted advances or seeking help from others on what actions to take. Unfortunately, teens did not find their existing approaches to be effective in keeping them safe and often reported that risks persisted (RQ1b). To help deal with these online risks effectively, teens designed real-time online safety interventions which often resembled "nudges". A nudge is designed to alter people's behavior through positive reinforcement without compromising their decision-making autonomy [44]. Teens designed interventions for tackling the situation at multiple stages (e.g., before, during, after) and from the perspective of both the victim and the perpetrator. Interestingly, teens designed more often for risk prevention by focusing on the perpetrator by encouraging them to rethink their actions, by blocking harmful actions from occurring, or penalizing perpetrators for inappropriate behavior. In case the risk could not be prevented, teens designed informative alerts, personalized sensitivity filters, and guided actions for assisting the risk victim (RQ2). Overall, teens challenged the status quo by shifting the responsibility of online safety to those perpetuating online harm by dealing with the online risks at the root cause, rather than focusing on victim protection or parental control (RQ3). Prior online safety solutions primarily focused on reactive approaches that protected teen victims after experiencing a risk. In contrast, our paper provides novel recommendations that prevent online risks *before* they occur by targeting designs toward the risk perpetrator. Additionally, this work is the first to employ co-design with teens to develop a broad range of online safety features through storyboards and high fideltity, interactive protoypes. To summarize, our work makes the following novel contributions:

- We used a new approach of running UX Bootcamps to equip teens with the skills required to contribute as equal partners in the co-design process. Our work is the first to co-design resilience-based interventions with teens to address a broad range of online risks derived from teens.

- Our findings move beyond understanding online risks in isolation. Instead, we provide a holistic understanding of teens' unsafe online experiences in full context, from *what* types of online risks they faced, *where*, *with whom*, and *how* they responded, and the *outcomes* of their unsafe online interactions.
- We provide teen-centered design-based recommendations based on co-design with teens that break from the status quo of victim protection, emphasizing on *risk prevention* (warning and penalizing the perpetrator) and *teen empowerment* (providing control and guidance for managing risks, rather than treating teens like victims).

2 RELATED WORK

We situate our work within adolescent online safety and co-design to illustrate how we build upon and extend prior work towards designing online safety interventions with teens.

2.1 Adolescent Online Safety and Risks

The current paradigm for adolescent online safety focuses heavily on "abstinence-only" approaches that attempt to shield teens from experiencing any and all online risks [4,29] by restricting access and parental controls. Recently, the narrative has shifted beyond the restrictive approaches towards strength-based approaches to online safety that empower teens to be resilient in the face of risks and can guide them towards safety [49,58]. This strength-based perspective was initially promoted in the context of adolescent risk behavior, where Zimmerman et al. [52] conceptualized resilience theory to inform intervention design for improved adolescent health. In the context of adolescent online safety, Wisniewski et al. [50] were one of the first to investigate resilience-based approaches online through a web-based diary study with adolescents in the US to understand how resilience plays a key role in protecting teens from experiencing online risks. Their results indicated that resilience (e.g., handling negative feelings effectively) is helpful in mitigating negative effects of online risks and helping teens develop interpersonal skills to deal with online risks. Since then, there has been a significant focus on promoting resilience-based approaches for adolescent online safety using different methods ranging from the traditional surveys [24], interviews [1,8,19], and diary studies [2,50], to novel methods such as app-based feature analyses [48], qualitative analyses of app reviews [25], and social media data analyses [39]. Yet, much of the prior work in this space focused on understanding online risks, the need for resilience-based approaches, or associated challenges, resulting in a lack of design-based solutions. In our work, we draw from and expand upon this prior work by moving beyond resilience as a concept, towards resilience-based actionable solutions co-designed with teens. In the next section, we cover previous research on the types of interventions designed for online safety and highlight how our research expands upon this work.

2.2 Design-based Interventions for Improved Adolescent Online Safety

While there is a significant body of research on understanding teens' online experiences, there is a lack of resilience-based interventions designed and implemented for helping teens with online risks [38]. The few efforts dedicated towards designing solutions are focused on parental controls, rather than teen empowerment [38]. In a parallel research area, researchers have explored design-based interventions, such as nudges, to help adolescents make better privacy decisions online. For instance, Alemany et al. [3] conducted an experiment in which they nudged teens about the privacy risks of publishing content on social media and found that nudges were effective in preventing teens from making potentially negative privacy disclosures online. In 2020, Masaki et al. [33] conducted an online survey with adolescents to compare how different nudge-based designs (e.g., negative/affirmative framing) influenced teens' decisions in scenarios featuring privacy and safety threats. They concluded that adolescents found nudges with negative framing (e.g., "90% of the users wouldn't do this") to be more effective in reducing privacy risk behaviors. Our research builds upon this work to understand whether designed-based interventions, like nudges, can also reduce risk behaviors and exposures related to adolescent online safety.

The investment in design-based interventions for adolescents' privacy, security, and wellbeing goes beyond academic endeavors, as there have been several industrial efforts for implementing and evaluating interventions as well. For example, in December 2021, Meta announced a new feature for encouraging teens to take a break, when they have been dwelling on one topic for too long, to improve their wellbeing [55]. OpenWeb [59] experimented to see whether users revised their content before posting when made aware of its potential harmful consequences. They found that 34% of the users edited their harmful posts when prompted. Similarly, we aimed to design online safety interventions that can help teens safely navigate unsafe experiences online. However, a limitation of the aforementioned research is that they examined the effects of the interventions with teens but did not involve teens in the design and develop of these solutions. Therefore, we address this gap and extend beyond the current online safety literature and industry efforts by designing interventions for online safety *with* teens through co-design. The next section summarizes research that employed co-design methods with youth and how this informed our study design.

2.3 Effectively Co-designing for Adolescent Online Safety Interventions

Co-design refers to the involvement and partnership of stakeholders in the design process [18]. Equal representation in the design process is particularly important for teens as a vulnerable population [45], with unique perspectives. In the context of adolescent online safety, co-design has been used previously to understand teens' unsafe experiences and design solutions for safety [6,7,34]. In 2014, Bowler et al. led one of the first co-design efforts for online safety, in which they conducted focus groups with teens and undergraduate students to design storyboards depicting cyberbullying interactions and interventions that could help alleviate those risks [53]. They found that teens designed for empathy and empowerment, that allowed transparency about the emotional effects of cyberbullying, which could impact the cyberbullies actions. Garaigorbil and Martinez-Valderray investigated the effects of Cyberbully 2.0, an educational program aimed at reducing face-to-face and cyberbullying, with teens in Spain. They found that their program significantly reduced cyberbullying and increased empathy [22]. Ashktorab et al. conducted participatory design sessions with high-school students to design solutions for mitigating cyberbullying [5], such as filtering certain personalized or explicit words, getting help from therapists, and for apps to report back to the cyberbullied users on the actions taken for safety.

While the aforementioned works focused solely on cyberbullying, McNally et al. extended these works to co-design parental monitoring software with children (ages 7-11). They found that children preferred parental controls that enabled risk coping, encouraged parent-child communication, and automated monitoring which notified parents at the time of risk [34]. Additionally, Badillo-Urquiola et al. conducted participatory design sessions with children to understand their needs for protection against stranger danger and found that younger children envisioned features that provided them alerts to ask for help, parental support, and automated assistance [6]. Both these studies were focused on solutions for younger children. In subsequent work, Badillo-Urquiola et al. [7] designed youth-centric online safety features with college students using a value sensitive design approach. Overall, a majority of the existing co-design efforts focused on more restrictive online safety approaches, such as redesigning parental controls [34], and employed co-designed with populations other than teens (i.e., younger children [6] or college students [7]). To our knowledge, the only existing co-design effort for online safety that engaged teens focused on cyberbullying risk prevention [5], exemplifying the need for our research. As such, we extend beyond this prior work as the first study to co-design online safety interventions with teens (ages 13-17) that address a broad range of online safety concerns raised by teens.

2.4 Addressing Knowledge Gaps in the Current Literature

In addition to the limitations of the prior work that we addressed in the sections above, there are over-arching knowledge gaps in the literature this work serves to close. Prior work has established the need for resilience-based solutions that promote adolescent online safety [49,58]; yet, our research is the first to use a grounded approach to generate design-based ideas that are actionable ways that address how we can actually start move towards accomplishing this goal. In contrast to Alemany and Masaki et als. [3,33] works that proposed nudge-based interventions that were evaluated by teens, we saw the need to work directly with teens to conceptualize online safety solutions that they felt would meet their needs. This approach allowed us to take a more holistic approach of first understanding the types of online risks modern-day youth encounter online and then working with teens to conceptualize ways to overcome and/or prevent these uncomfortable situations. As such, a core strength of our research is the ecological validity of the types of risks teens experience online and the types of interventions they felt would be most appropriate for addressing these situations. Therefore, our research is wholly teen-centered in that we let teens create the scenarios for which they designed online safety solutions without constraining them to any partcular design space (e.g., parental controls or nudges). We describe our methods for this approach in more detail below.

3 METHODS

In this section, we summarize our study methodology, recruitment efforts and participants' demographics, along with an overview of our data analysis approach.

3.1 Study Overview

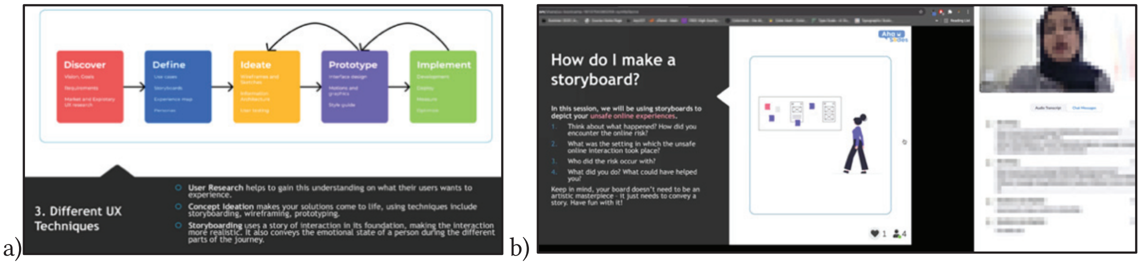
We worked directly with teens to co-design online safety interventions contextualized to online risk experiences relevant to their lives. To do this, we conducted nine UX Bootcamps virtually via Zoom with 21 adolescents (ages 13-17). The UX Bootcamp consisted of trainings and research activities conducted over a span of two days with the end goal of designing high-fidelity, interactive prototypes for online safety. The trainings covered several topics, including adolescent online safety, UX design, storyboarding, and prototyping. Using the skills learned in the trainings, teens were guided to participate in three research activities: a) creating storyboards for unsafe and uncomfortable online interactions, b) whiteboarding ideas for online safety solutions, and c) developing prototypes for design-based interventions for online safety. Due to COVID-19, this study was conducted virtually via Zoom, using online tools (i.e., Canva, FigJam, and Figma) for each of the activities, respectively. Each activity was a combination of individual work, co-designing with researchers, and group discussions. At the conclusion of each Bootcamp, teens were asked to complete an exit survey to provide feedback for the Bootcamp.

3.2 UX Bootcamp Training Activities

Our aim was to involve teens in the design of online safety interventions in a way that made them well-equipped to be equal partners during co-design, while also benefitting them in the process. Prior work highlights how teens often lack the skills and training to act as equal partners in the design process [16,51] and that impactful research with teens needs to be mutually beneficial; teens require motivation and incentives for their participation in research [8]. Therefore, as part of the UX Bootcamp training, we first introduced teens to the topics of online safety and user experience. We also provided UX training for creating storyboards and low/high-fidelity prototypes. We summarize each of the training activities below.

3.2.1 Introduction to Adolescent Online Safety and Risks. We began the session with introductions, followed by guidelines for the session and an icebreaker activity. We began the trainings with an introduction to adolescent online safety with recent news headlines regarding teens' online safety concerns and introduced them to relevant adolescent online safety research addressing these concerns and discussing methods that involve youth in research (e.g., participatory design). After establishing context on adolescent online safety, we asked teens a warm-up question about what they considered to be an unsafe or uncomfortable interaction online. For this part of the Bootcamp, we also provided the option for teens to enter their responses anonymously through interactive Aha slides [60]. We ended this section of the training by defining the goals of the Bootcamp, which included, a) learning about UX skills and tools, and b) applying these skills to co-design safety features for unsafe online interactions.

3.2.2 Teaching User Experience (UX) and Storyboarding. We started the UX training with a warm-up activity asking teens to compare two user interfaces for the login page of an app and select and justify their favorite interface design to help teens start thinking from a design perspective. Next, we introduced teens to user experience concepts (**Fig. 1a**), such as the importance of user experience [21], the five stages in the design thinking process [15], and concept ideation techniques (e.g., storyboarding, wireframing, prototyping) [61]. Building upon the introductory material, we narrowed the focus of the UX training to storyboarding. We familiarized teens with the process of creating storyboards, and how they may be used in the UX design cycle. Then, we provided guidelines (**Fig. 1b**) and an example for creating a storyboard to depict an unsafe online interaction faced by the teens, someone they know, or a hypothetical scenario based on real-life. We ended this portion of the training by demonstrating the features of Canva [55], the online tool we used for the storyboarding activity.



3. Different UX Techniques

- User Research helps to gain this understanding on what their users wants to experience.
- Concept Ideation makes your solutions come to life, using techniques include storyboarding, wireframing, prototyping.
- Storyboarding uses a story of interaction in its foundation, making the interaction more realistic. It also conveys the emotional state of a person during the different parts of the journey.

How do I make a storyboard?

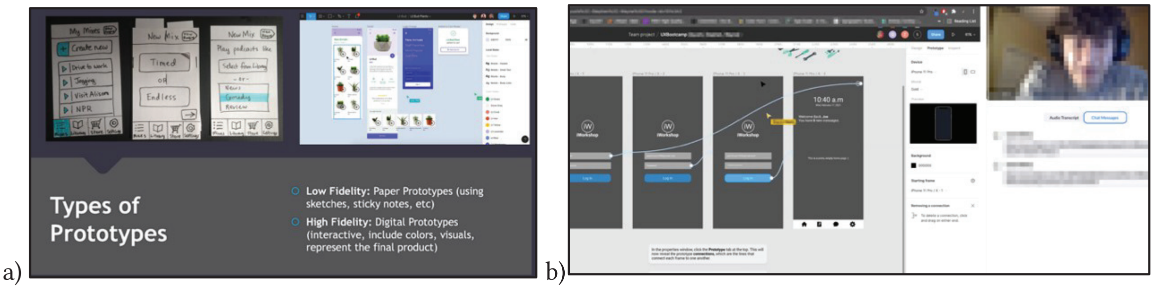
In this session, we will be using storyboards to depict your unsafe online experiences.

1. Think about what happened? How did you encounter the online risk?
2. What was the setting in which the unsafe online interaction took place?
3. Who did the risk occur with?
4. What did you do? What could have helped you?

Keep in mind, your board doesn't need to be an artistic masterpiece - it just needs to convey a story. Have fun with it!

Fig. 1. Introduction to UX, b) Storyboard Training.

3.2.3 Teaching Low and High-Fidelity Prototyping. Our prototyping training started with an introduction to low and high-fidelity prototypes (**Fig. 2a**). The low-fidelity prototyping activity (described in section 3.3.2) served as a brainstorming and preparatory exercise for their final prototypes. We demonstrated the purpose of low-fidelity prototypes and examples on how to effectively create them using FigJam, [62] which a virtual whiteboarding tool. For the final prototyping training, we designed an interactive workshop embedded within Figma, which was the primary application used for the prototyping activity. Figma [63] is a web-based prototyping tool widely used in the industry to design and brainstorm product ideas. The workshop was divided up into sections covering a different learning principle or tool in Figma (e.g., frames, connections, navigation) and was followed by a practice activity (**Fig. 2b**). Each teen would pair up with one of the researchers and follow along to learn about Figma. By the end of this training, teens were familiar with Figma and prototyping, and demonstrated their abilities by prototyping a chat-based interface.



Types of Prototypes

- Low Fidelity: Paper Prototypes (using sketches, sticky notes, etc)
- High Fidelity: Digital Prototypes (interactive, include colors, visuals, represent the final product)

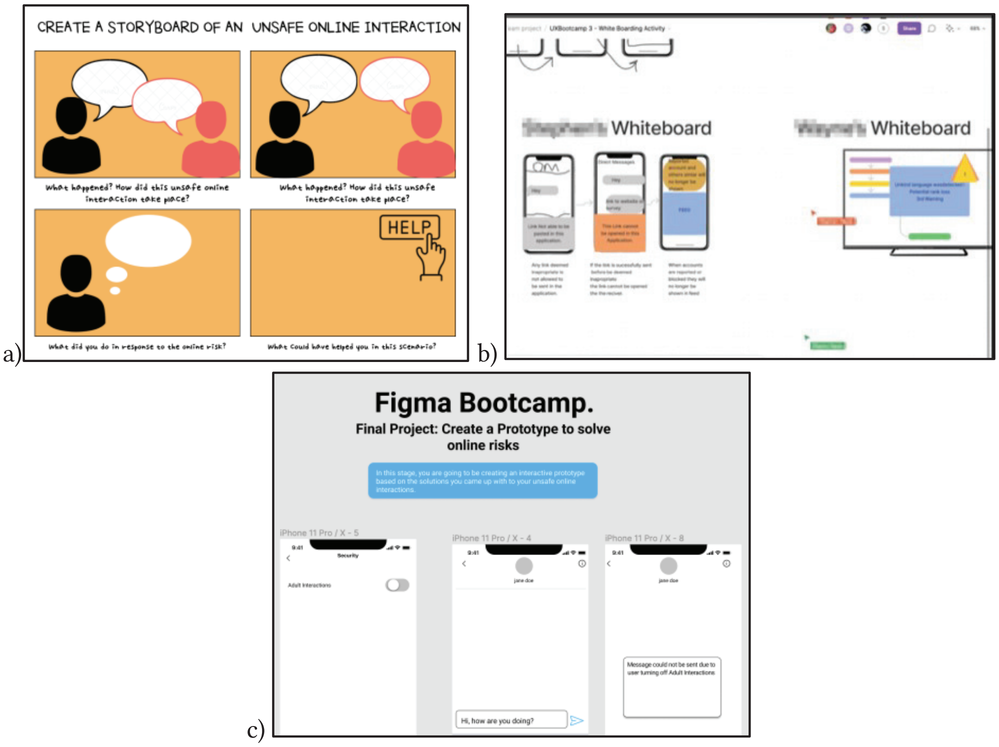
Fig. 2. a) Introduction to Low/High-fidelity Prototyping, b) Figma Training.

3.3 Research Activities

Each of the training activities was accompanied with a research task involving co-design activities with teens for informing and designing interventions for online safety. In this section, we will summarize these design activities.

3.3.1 Designing Storyboards for Unsafe Online Interactions. After the training for UX and storyboarding, teens were asked to create their own storyboards based on an unsafe and uncomfortable online interaction scenario. To ensure teens' comfort in a group setting, we allowed the teens to share a scenario either based on their personal experiences, anonymized experience of friends, or a hypothetical scenario of a common online risk. Teens were provided with a storyboard template (**Fig. 3a**) along with instructions for recreating the scene, visualizing responses and reactions towards the risk, and demonstrating possible solutions. Throughout the process, teens were asked follow-up questions by researchers to understand their scenario and proposed solutions. For example, “Who were you interacting with? How did you react to this situation? What could have helped you in this situation?”. Teens spent about 45 minutes to complete their storyboards. After they finished, each teen was asked to present their storyboard to the group. The other teens and researchers provided feedback on each storyboard.

3.3.2 Low and High-Fidelity Prototyping for Online Safety Features. Building upon the storyboards, teens were first asked to use the low-fidelity prototyping method to brainstorm details of their proposed online safety solution for dealing with the unsafe scenario described in the storyboard (**Fig. 3b**). Researchers asked probing questions to help teens brainstorm their ideas for online safety features, and helped teens organize and structure the different elements and flow of their safety design. Teens had about 60 minutes for the whiteboarding activity. After they finished, each teen was asked to present their whiteboard low-fidelity prototype to the group which helped identify the limitations of their ideas and get suggestions for improvement before the final implementation of their ideas in prototypes.


a)
CREATE A STORYBOARD OF AN UNSAFE ONLINE INTERACTION
What happened? How did this unsafe online interaction take place?
What happened? How did this unsafe interaction take place?
HELP
What did you do in response to the online risk?
What could have helped you in this scenario?
b)
[redacted] Whiteboard
[redacted] Whiteboard
c)
Figma Bootcamp.
Final Project: Create a Prototype to solve online risks
iPhone 11 Pro / X - 5
iPhone 11 Pro / X - 4
iPhone 11 Pro / X - 8
9:41
Security
Adult Interactions
jane doe

In this stage, you are going to be creating an interactive prototype based on the solutions you came up with to your unsafe online interactions.

Hi, how are you doing?

Message could not be sent due to user turning off Adult Interactions

Fig. 3. a) Storyboarding template provided to teens to depict their unsafe online interactions, b) Teens’ low-fidelity prototypes for brainstorming ideas for safety interventions, c) Teens’ high-fidelity prototyping project for designing online safety interventions.

The final activity was the high-fidelity prototyping for online safety features specific to their unsafe scenario presented in the storyboard (**Fig. 3c)**. This final project was done in breakout rooms where each teen was paired with one or more researchers for assistance. Teens had approximately 1 hour and 30 minutes to complete their final prototype. Throughout the design process, researchers asked follow-up questions about the teens idea to help refine the prototype or identify missing components of the idea. Once the individual prototypes for online safety were completed, the group reconvened in the main room in Zoom where teens presented their prototypes in an interactive demo. While presenting, teens were prompted to explain their unsafe online interaction, their prototyped solution to that online risk, and how it would impact adolescent online safety. After each teen presented their prototype, researchers opened the floor to the group for asking questions or sharing comments regarding that prototype. We ended the Bootcamp by summarizing the designs presented and appreciating teens for their contribution to online safety. After the bootcamp, we sent an exit survey to teens for getting their feedback and suggestions on the training and design activities of the bootcamp.

3.4 Data Analysis Approach

The data collected included audio and video recordings, participant responses through Aha slides, design artifacts (storyboards, whiteboard wireframes, prototypes), and feedback survey data. The recordings were transcribed using Otter AI, after which they were manually revised by the researchers for any errors. To address each of our research questions, we performed qualitative thematic analysis [12], as it's suitable for generating new themes and insights from the data. We began by reviewing the transcripts along with the designs artifacts to generate initial codes. The first author coded the first few sessions to generate an initial codebook. This codebook was then used as a guide during the coding process by the research assistants. We used an iterative "follow-the-leader" approach during data coding process, where the research assistants consulted the first author after coding each session and any time they had questions or when new codes emerged. After coding each session, the first author met with the research assistants to form a consensus, resolve conflicts, and update the codebook accordingly. By the last session, we reached theoretical saturation, where no new codes emerged from our analyses [28]. Therefore, we concluded data collection. Our codes were then refined and grouped conceptually into themes to create our final codebooks.

To answer RQ1a and RQ1b, we coded the storyboards and teens responses to follow-up questions for the risk types, public vs. private risks, social media platforms (**Table 2**), and coping responses (**Table 3**). To answer RQ2 and RQ3, we analyzed the prototyped features co-designed by teens for online safety interventions and conceptualized their novel themes (**Table 4**). Many of the teens' storyboards and prototypes were double coded as they represented multiple risk types, coping responses, or multiple ideas for interventions, making the total percentages greater than 100%. The data coding was done by three researchers, who met consistently to discuss and merge their codes, as well as resolve any conflicts. The thematic analysis was completed by the first author with feedback from all co-authors.

3.5 Participant Recruitment and Demographics

We obtained Institutional Review Board (IRB) approval before recruiting participants. Interested teens had to complete an eligibility survey to confirm that they are from the United States, between the ages of 13-17 years old, and have access to reliable internet, as well as a webcam and microphone. After confirming eligibility, teens were informed that parental consent is required to participate in the study. After parental consent was obtained electronically via Qualtrics, teens were asked to provide their own assent for participation. Most teens were between 15 and 17 years of age (62%, N=13), with a mean age of 15.2 and a standard deviation of 1.4 years. We had a diverse sample of teens with participants identifying as Hispanic/Latino (5%), White/Caucasian (15%), Black/African

American (33%), and Asian (47%). We had a good gender representation with 9 female (43%) and 12 male (57%) participants (**Table 1**). We recruited teens by distributing flyers and Bootcamp information to schools and STEM organizations via emails, phone calls, and social media. We also held meetings with schoolteachers, principals, coaches, and youth program coordinators to spread the word about the Bootcamp. While we recruited a diverse sample, all participants had access to education and technology, and may not be representative of those who are more vulnerable to online risks, such as those from different cultural or socio-economic backgrounds. Teens received certificates of completion for participating in the Bootcamp. The activities for each Bootcamp were conducted over two days (on the weekend or after-school), with each session lasting 3.5 hours per day. A total of seven researchers helped conduct the bootcamps with 3-5 researchers per session. Recruitment began in March 2021 and concluded in November 2021.

Table 1. Participants’ Demographic Information

Group	ID	Age	Sex	Ethnicity	Risk Type
Group 1	P1	17	F	Asian	Information Breach, Sexual Risks
	P2	17	F	Asian	Sexual Risks
	P3	17	F	Asian	Information Breach
Group 2	P4	17	F	Asian	Information Breach, Sexual Risks
Group 3	P5	16	M	Asian	Information Breach, Sexual Risks
	P6	15	F	White/Caucasian	Online Harassment
	P7	14	M	Asian	Information Breach
Group 4	P8	15	F	Asian	Information Breach
	P9	15	M	Black/African American	Information Breach, Sexual Risks
	P10	13	M	Black/African American	Online Harassment
Group 5	P11	14	M	White/Caucasian	Information Breach, Online Harassment
	P12	17	M	Asian	Information Breach
	P13	16	F	Asian	Information Breach, Sexual Solicitation
Group 6	P14	14	M	Hispanic/Latino	Information Breach, Online Harassment
	P15	13	F	Black/African American	Information Breach, Sexual Solicitation
Group 7	P16	15	M	Black/African American	Online Harassment
	P17	15	M	Black/African American	Online Harassment
Group 8	P18	14	M	Black/African American	Information Breach
	P19	14	M	Black/African American	Information Breach
Group 9	P20	14	F	Asian	Online Harassment
	P21	16	F	White/Caucasian	Information Breach, Sexual Solicitation

4 FINDINGS

In this section, we summarize the key findings answering each of our research questions. We use illustrative quotes and design artifacts from the bootcamp sessions as examples.

4.1 Storyboarding What Makes Teens Feel Uncomfortable or Unsafe Online (RQ1)

We were first interested in understanding what types of situations most often made teens feel uncomfortable or unsafe online (RQ1a), how do teens deal with these situations, and if teens believe their existing approaches to handle these situations are effective (RQ1b). The following sections detail our findings.

4.1.1 Unsafe and Uncomfortable Online Experiences Storyboarded by Teens (RQ1a). To ensure their comfort, teens were given the option to either share a personal unsafe experience, an anonymized experience of someone they knew, or a hypothetical scenario based on realistic online risks. Despite the option to share a hypothetical scenario, most teens shared their personal real-life online risk experiences (81%, N=17). Only a few teens shared anonymized and fictionalized versions of their own or their friends' unsafe or uncomfortable encounters online (19%, N=4). **Table 2** summarize the types of online risks teens commonly reported in their storyboards and the characteristics of their unsafe online interactions. Percentages within each column are calculated based on the total for that specific risk type.

Although teens were not prompted to share an unsafe or uncomfortable experience from a specific platform, many teens designed their storyboard for risks encountered on Instagram (47%, N=10). Most of these experiences were private conversations through direct messages (67%, N=14) with strangers (57%, N=12), often adult predators. However, the risks varied; teens reported experiences of information breaches (71%, N=15), sexual risks (38%, N=8), and cyberbullying (33.3%, N=7). While **most teens reported information breaching risks** (71%, N=15), the most common type of encounter was personal information requested, doxed, or posted online without the teen's consent (52%, N=11). Teens were often asked for information that could reveal their identity (e.g., name, age, phone number) or location (e.g., home address, school) and could potentially lead them to offline threats. For example, P4, a 15-year-old Female, received messages from an older stranger asking for her age and location, making her feel uncomfortable (**Fig. 4**). In a few cases, teens were even offered incentives (e.g., money) in exchange for their information.

Table 2. Summary of teens' unsafe experiences online (RQ1a)

Dimension	Codes	Total* (N=21)	Info Breaches (71%, N=15)	Sexual Risks (38%, N=8)	Cyberbullying (33%, N=7)
Platform	Instagram	**48%, N=10**	**53%, N=8**	**63%, N=5**	**43%, N=3**
	Discord/Gaming	19%, N=4	13%, N=2	13%, N=1	**29%, N=2**
	Other	14%, N=3	**20%, N=3**	13%, N=1	14%, N=1
	Twitter	10%, N=2	7%, N=1	13%, N=1	14%, N=1
	Snapchat	10%, N=2	7%, N=1	0%	0%
Public vs. Private	Direct Message	**67%, N=14**	**73%, N=11**	**88%, N=7**	**43%, N=3**
	Public Post	19%, N=4	13%, N=2	0%	**43%, N=3**
	Group Chat	14%, N=3	13%, N=2	13%, N=1	14%, N=1
Relationship	Stranger	**57%, N=12**	**60%, N=9**	**88%, N=7**	**43%, N=3**
	Acquaintance	24%, N=5	27%, N=4	0%	**29%, N=2**
	Friend	19%, N=4	13%, N=2	13%, N=1	**29%, N=2**

Total equals the total number of risk scenarios. Risk types are N>21 due to multiple risks in the same scenario.

Yet, most of **the information breaching encounters occurred in parallel with other risks**, such as sexual risks, cyberbullying, and harassment. This concurrence of information breaches with other risks explains why it was the most prominently reported risk type. For example, many requests for information were often accompanied with inappropriate sexual comments or requests to meet in person. For instance, P5 got inappropriate messages from a middle-aged adult with a request to meet up, that could lead to offline risks.

> "*She said I looked cute and wanted to hang out with me somewhere.*" –P5 (16 yr. old, Male)

Sexual risks were another common type of unsafe online interactions (38%, N=8) reported in the teens’ storyboards. Unlike information breaches and cyberbullying, no sexual risk experiences were reported in public posts and only one happened in a semi-public group chat. While one teen reported an incident with a friend, the **majority of the sexual experiences were perpetuated by strangers** (87.5%, N=7), specifically adult predators (62.5%, N=5). Most of the risk experiences involved the teen receiving sexually suggestive comments about their appearance or inappropriate compliments (75%, N=6). These sexual messages were sometimes accompanied with requests to meet in-person or asking about the teen’s location, which escalated the risk. For instance, P1 storyboarded her interaction with multiple strangers sending her predatory messages along with asking for personal information and offering money (**Fig. 5**).

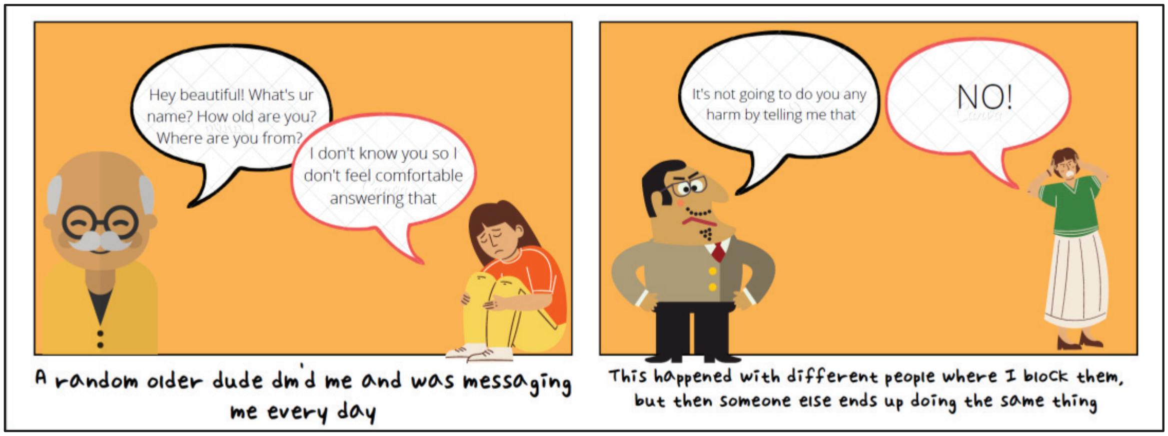

Hey beautiful! What's ur name? How old are you? Where are you from?
I don't know you so I don't feel comfortable answering that
A random older dude dm'd me and was messaging me every day
It's not going to do you any harm by telling me that
NO!
This happened with different people where I block them, but then someone else ends up doing the same thing

Fig. 4. P4’s storyboard on facing an information breaching risk with an older stranger.

> “*You have no idea who they are. And it starts out as an innocent conversation and everything. But then they start asking pretty personal questions...” –P1 (17 yr. old, Female*)

Some sexual requests asked for a nude photo or to sext. These requests often disturbed teens, especially when they were from people they knew. For example, P2, a 17-year-old Female, received explicit photos from a friend, who insisted her to exchange nudges. P2 explained feeling unsafe and disappointed when her friend pressured her to send photos, when she was looking for a platonic friendship.

> “*Throughout the conversation, I was a little sad that she kept asking for my pics instead of getting to know me.”* –P2 (17 yr. old, Female)

Other types of sexual encounters included receiving explicit content in the form of photos and videos that the teen did not wish to see. These messages were often in the form of a spam link which concealed the content but exposed the teen to explicit imagery once clicked. P15 explained:

> “*This stuff [content] they sent to me isn't appropriate. So like, I didn't even click on it, but the way it appeared, I was like, no, this can't be good” –P15 (13 yr. old, Female*)

Several teens also storyboarded experiences of cyberbullying and harassment (33.3%, N=7). Unlike information breaches and sexual risks, cyberbullying happened relatively more within public facing environments (i.e., public posts and semi-public group chats) (57%; N=4). However, the types of cyberbullying situations varied in severity. Some teens reported being bullied online through hurtful messages (24%, N=5), such as mocking the teens' photos, appearance, or family. While other teens reported on more severe risks, such as targeted threats (9.5%, N=2) with the goal to harm the teen, their social reputation, or their family. Teens discussed how sometimes the hurtful comments received had consequences on a teen's mental wellbeing. For example, P6 explained how her friend was consistently bullied on Instagram which made her friend depressed:

> "*They [classmates] were sending her these like very mean messages, criticizing her, like her looks her body and just saying she's not good at this stuff... she was like falling into depression*" –P6 (15 yr. old, Female)

Beyond social media, there were also instances of unsafe interactions in gaming environments. For example, while playing Fortnite, a teen (P10, 13-year-old, Male) was attacked and intimidated by a competitor in the game with serious threats (**Fig. 6**):

> "*He would call me inappropriate names, and then say that he would do disgusting things to my mother, or and he would say he would kill people in my family.*" –P10 (13 yr. old, Male)

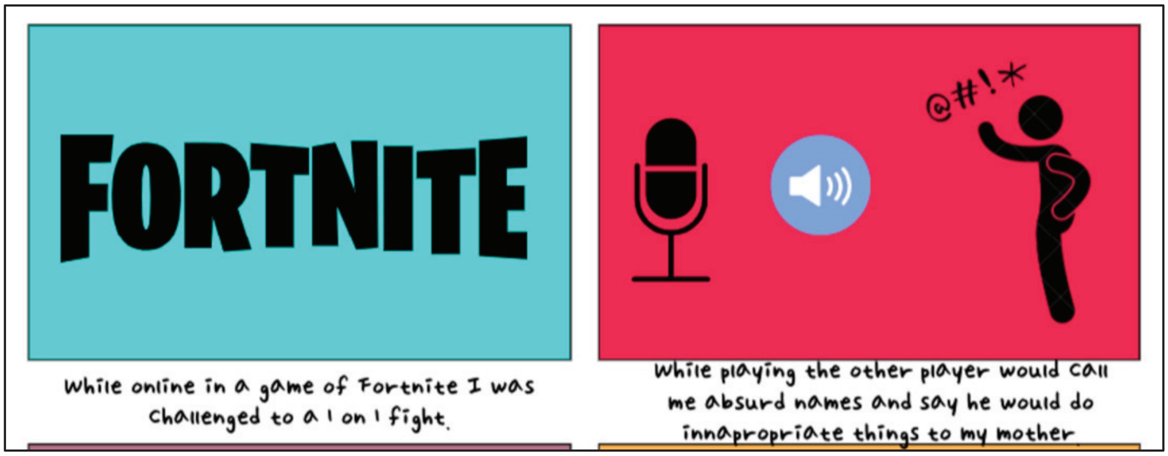

FORTNITE
@#!*
While online in a game of Fortnite I was Challenged to a 1 on 1 fight.
While playing the other player would call me absurd names and say he would do innapropriate things to my mother

Fig. 6. P10's storyboard about facing targeted threats from a player in Fortnite.

Another difference of cyberbullying risks is that teens often faced these risks with people they knew, such as acquaintances (28.6%, N=2) and friends (28.6%, N=2), rather than strangers (42.8%, N=3). In the next section, we summarize the approaches teens took to manage the types of unsafe and uncomfortable interactions they experienced.

4.1.2 Teens' Responses to Online Risks and their Outcomes (RQ1b).

Teens' responses included blocking, reporting ignoring or deleting the risk, refusing to comply with the uncomfortable request, or confirming safe actions with others (**Table 3**).

The outcomes of their unsafe and uncomfortable experiences have been grouped into three categories; persisted (i.e., the risk continued, either in that instance or a similar risk happened later), ineffective (i.e., the teen still felt unsafe or the response did not change their safety) and undetermined (i.e., the final outcome was not specified by the teen) (**Fig. 7**). As shown in the Sankey diagram, regardless of the type of response to the unsafe experience, teens felt that the risks persisted, and the responses ineffective in keeping them safe (**Fig. 7**). In this section, we elaborate on teens responses to online risks in relation to risk types and their outcomes.

Table 3. Summary of teens’ responses to online risks and their outcomes (RQ1b)

Dimension	**Codes**	**Response Type**					
		Block	**Report**	**Ignore**	**Confirm**	**Refuse**	**Delete**
		62%, N=13	**33%, N=7**	**19%, N=4**	**19%, N=4**	**14% N=3**	**14%, N=3**
Risk Type	Info Breaches	**77%**	**57%**	**75%**	50%	67%	100%
	Sexual Risks	39%	29%	25%	**75%**	**67%**	**67%**
	Cyberbullying	15%	**57%**	50%	25%	0%	0%
Relation	Stranger	**69%**	**71%**	**50%**	**50%**	33%	**100%**
	Acquaintance	31%	29%	25%	0%	0%	0%
	Friend	0%	0%	25%	**50%**	**67%**	0%
Outcome	Persisted	**46%**	29%	50%	**50%**	33%	**100%**
	Ineffective	15%	29%	**50%**	25%	**67%**	0%
	Undetermined	39%	**43%**	0%	25%	0%	0%

Despite blocking strangers or acquaintances in response to information breaches or sexual risks, teens reported risks persist. Blocking was the most common coping mechanism that teens (62%, N=13) used in response to the risk. Teens often used blocking in combination with other ways of coping such as reporting, deleting, and ignoring. Blocking was most commonly used on Instagram (46%, N=6) in response to information breaching risks (77%, N=10), and for some sexual risks (38.5%, N=5) Additionally, teens most commonly blocked strangers (69%, N=9) and never their friends. Despite blocking the other person after an unsafe interaction, many teens not only felt unsafe but reported that the risk persisted (46%, N=6), either with the same person who made a fake account, or similar risks occurred with other accounts (**Fig. 8a**). For example, P20, a 14-year-old Female explained her experience with people online who kept making inappropriate comments on her personal photos. Even after blocking many of them, new accounts appeared, and the unsafe comments continued.

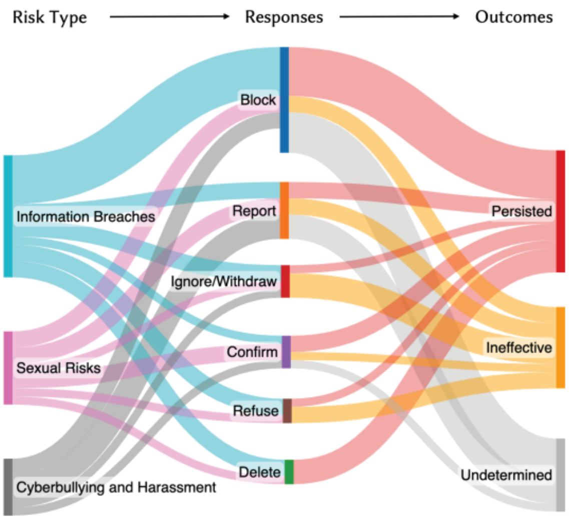


Fig. 7. Sankey Visualization of Teens’ Unsafe Experiences, Responses to Risks, and Outcomes.

> “*They keep stalking you and like, keep commenting on all your posts. And when you block them once, another account gets created to keep doing the same thing?*” –P20 (14 yr. old, Female)

A few teens (15.4%, N=2) did not report the risk reoccurring after blocking, but still felt that blocking was ineffective as they felt that damage had already been done or they had to take further action to feel safe. At the same time, the outcome of a few teens' safety after blocking remained undetermined (33.3%, N=5) as they did not indicate whether blocking led to safety.

Teens usually reported strangers in response to information breaches or cyberbullying, which was often unreliable or ineffective. Reporting was the second most frequently used action by the teens in response to online risks (33%, N=7), who mostly used reporting and blocking together to deal with unsafe interactions. Teens used reporting more frequently in response to information breaches (57%, N=4) and cyberbullying (57%, N=4). Similar to blocking, teens used reporting with risks from strangers the most (71%, N=5) on Instagram. However, some teens often found the reporting feature to be ineffective and wanted a more reliable reporting system that would carefully fact-check risks before reporting an account, to avoid false accusations.

> "*Instagram or whatever social media should see both sides of the thing [after reporting], and not just one side. Like they would be able to look up history on that and be able to check it more thoroughly." –*P14 (14 yr. old, Male)

A few teens expressed that the risk continued despite reporting (28.6%, N=2). This was either due to similar instances of the risks repeating, or the continued negative impact of the risk on their mental health despite reporting. For example, P20 reported a risk of cyberbullying with an acquaintance, and explained how the mental health impact of the risk cannot be undone with reporting, indicating the need for additional resources that can help teens with mental health.

> "*So, when you get bullied, it's not like, it just goes away after you report it…you're mentally impacted. It's just stuck in your mind"* –P20 (14 yr. old, Female)

Lastly, the outcome of reporting the risk was undetermined for many teens (42.8%, N=3), where the teens indicated neither a positive nor negative consequence of reporting. Therefore, reporting may be a response that makes some teens feel that the situation was neutralized.

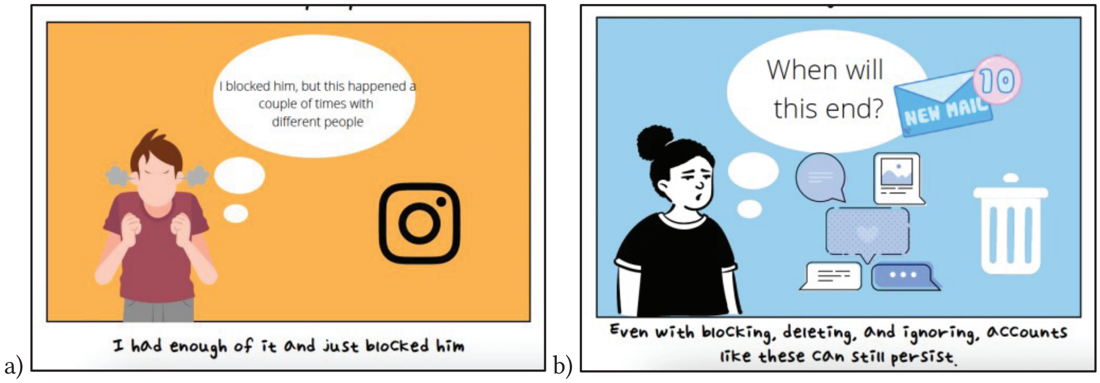

I blocked him, but this happened a couple of times with different people
I had enough of it and just blocked him
a)
When will this end?
NEW MAIL
10
Even with blocking, deleting, and ignoring, accounts like these can still persist.
b)

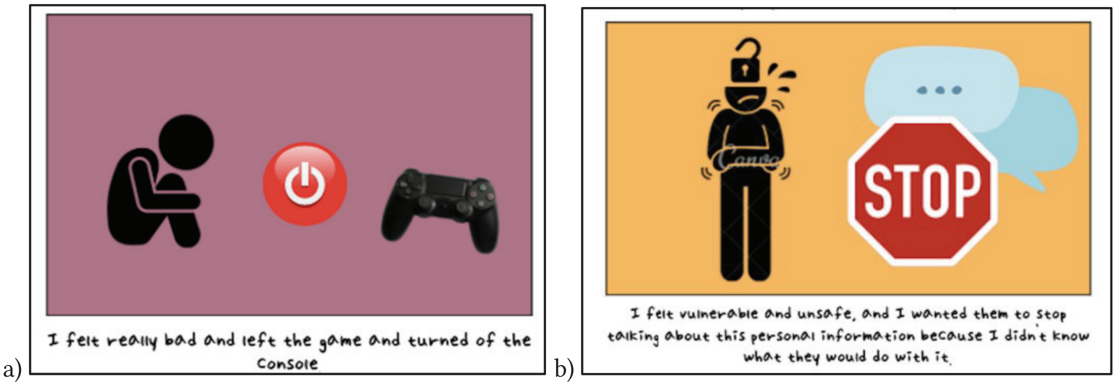
Fig. 8. a) P4's storyboard about risks persisting after blocking, b) P1's storyboard describing how multiple responses still lead to unsafe outcomes.

Other teens ignored or withdrew from the risk (19%, N=4), or confirmed safe actions with others (19%, N=4). Teens ignored information breaching risks the most (75%, N=3), followed by cyberbullying (50%, N=2). Interestingly, teens ignored strangers (50%, N=2) as much as their acquaintances and friends combined (50%, N=2). Similar to other responses, teens found ignoring to be an ineffective strategy. For example, P18 ignored another player in a game asking for his personal information but was still followed by the other player and pursued for information, and later blocked him, indicating that teens often had to use multiple ways of coping and yet did not feel fully safe. Having similar experiences with predators and bot accounts, P1 summarized how these risks can persist despite multiple responses (**Fig. 8b**).

> “*Predators, bots, etc. can continuously send DMs. Even with blocking, deleting, and ignoring, accounts like these can still persist.*” –P1 (17 yr. old, Female)

Some teens (19%, N=4) relied on advice from people they trusted to share the encounter and get confirmation on safe actions to take. Teens reported taking advice from their parents or friends on the best ways to respond to the risk. Unlike other responses, confirmation was most frequently used by teens after receiving sexual risks (75%, N=3). and often used with risks faced with friends (50%, N=2). Yet, risk persisted for half of these teens despite getting help from others (50%, N=2), as they often faced similar encounters later in time. For instance, P15 was added in a semi-public group chat with strangers where explicit content was being shared. She recalled that she turned to her parents who helped her block the group, but a few months later she was added back in a similar group with inappropriate content being shared.

> “*So I went to tell my parents [about the risk]. I don't know how to block, but my dad showed me. But, three months after, they put me in the [unsafe] group again.*” –P15 (13 yr. old, Female)



I felt really bad and left the game and turned of the console

I felt vulnerable and unsafe, and I wanted them to stop talking about this personal information because I didn't know what they would do with it.

Fig. 9. a) P10’s storyboard about withdrawing from the game when he faced cyberbullying, b) P8 explained how she struggled with saying no to her friends from sharing her information.

Relatively fewer teens (14%, N=3) refused an unsafe or uncomfortable request or attempted to stop the other person from their unsafe action. Specifically, teens used refusal such as saying “No” or “Stop” for requests for personal information or sexual requests. Moreover, refusal was the only response teens used the most with friends (66.7%, N=2), but still found it challenging to resist the peer pressure or stand up for themselves (**Fig. 9b**).

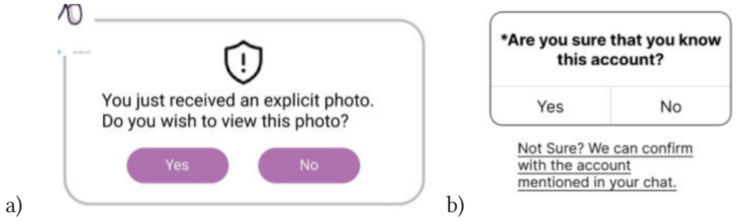
> “*I said no. Because she keeps asking [for photos]. And it was very hard to like, switch the topic because she was so persistent. And so I had trouble like responding to her.*” –P2 (17 yr. old, Female)

Lastly, a few teens deleted the unsafe or uncomfortable messages or content after receiving it. Deleting was always coupled with blocking or other responses, and no teens used delete on its own. Teens usually deleted instances of information breaching risks or sexual risks, with strangers and it always resulted in negative outcomes. Overall, teens often expected their response to a risk to not only keep them safe in that moment, but to provide long-term safety for similar risks. Although blocking and reporting sometimes neutralized the situation, most of the teens not only felt unsafe but reported the same risks happening again, in an unrelenting cycle. Next, we summarize teens designs for new and improved safety features.

4.3 Teens' Design-Based Recommendations for Improved Online Safety (RQ2)

In this section, we describe the features and real-time interventions designed by teens for helping them deal with unsafe interactions online. Most of the teens developed features that were built upon the idea that the social media platforms provide accurate and automated risk detection in real-time. Therefore, accurate risk assessment served as a prerequisite for their design-based recommendations. Moreover, their features relied on risk detection from different dimensions. For example, the features required accurate detection of the different risk types faced by teens such as requests for personal information, personal photos, or harmful words sent to the teen. Some teens also designed for unsafe or uncomfortable scenarios that relied on accurate detection of potentially explicit content, harmful links that could lead to information breaches, or voice data risk detection for cyberbullying during gaming. Other types of detection required accurately detecting accounts, such as strangers, adults, or people over a certain age. **Table 4** summarizes the features co-designed by teens to improve online safety in real-time.

Risk Alerts for Teens. Many teens (90.5%, N=19) designed alerts for teens to warn them before they get exposed to an unsafe situation. Some of these alerts were aimed towards **warning about unsafe content** received such as request for personal information, explicit photos, harmful language. These alerts described the risk of the message received and often provided recommendations for safe actions to take. These actions included reminders to not view the unsafe content, to avoid engaging in unsafe interactions (such as sharing of sensitive information), or to report or block the other person (**Fig. 10a**).



You just received an explicit photo. Do you wish to view this photo?

*Are you sure that you know this account?

Not Sure? We can confirm with the account mentioned in your chat.

Fig. 10. a) A risk alert warning teen about the explicit content received. b) A contact risk warning alerting about a stranger message.

Teens also created alerts that informed them when someone else posted content about them such as their information or photos. This alert allowed teens to consent to information being posted about them, and the other person would not be able to continue with sharing this information without the teen's approval. P4 explained:

> "*They would have to ask permission to like, post or comment on that thing, instead of directly posting it, you know?*" –P4, (17 yr. old Female)

Teens also designed **contact risk alerts for teens** that would be triggered on contact from certain types of users, that may pose harm, such as strangers or predators (**Fig. 10b**). Similar to content warnings, these alerts suggested ways to be safer. For example, teens designed stranger danger warnings and the ability to disable messages from specific individuals. P4 extended this idea

Table 4. RQ2: Themes within Teens Co-designed Online Safety Interventions

Themes	Features	Exemplary Quotes
Risk Alerts for Teens, (90.5%, N=19)	Content Risk Alerts (N=13)	"*The thing I made was to basically* ***alert or warn before they open the message*** *that - oh there's something bad coming up*" – P15 (13 yr. old, Female)
	Contact Risk Alerts (N=5)	“*So, it would get triggered within new messages. It will* ***send a notification on whether or not you know this account****, and you're sure that you know, you can trust it*” –P13 (16 yr. old, Female)
	Personal Content Consent (N=2)	“*They would* ***have to ask permission to like, post or comment*** *on that thing, instead of directly posting it kind of, you know?*” –P4, (17 yr. old, Female)
Proactive Warnings and Restriction for the Perpetrator (52%, N=11)	Block Unsafe Content or Contact (N=11)	“*Give them the option to go to their settings and* ***disable accessibility to their direct messages from strangers***” –P1 (17 yr. old Female)
	Trusted Circle (N=2)	“*In the settings, it [the app] has the trusted circle ... and I put* ***it should be reserved for extremely close friends****, close cousins, siblings, parents…*” –P11 (14 yr. old, Male)
Sensitivity Filters (47.6%, N=10)	Personalized Filters and Censoring (N=10)	“*Something that would* ***prevent harmful stuff from getting to people****... You get to choose... to hide it completely or you just want to blur it.*” –P6 (15 yr. old, Female)
Robust Risk Reporting (43%, N=9)	Risk Proof Upload (N=6)	“*I think maybe the questions [with reporting] could be like, you know, if you were the one who experienced the risk to* ***share some like, evidence.***” –P4, (17 yr. old Female)
	User Reports and Rating (N=2)	“*Once you click on his profile...* ***you see he has a lot of reports, you can click on user reports*** *and then it'll show you like the number of reports he has*” –P4, (17 yr. old Female)
	Auto-Report Similar Users (N=1)	“*You're like 'I don't like this' so you go report it. So then all the reported accounts and* ***all similar accounts will no longer be shown*** *on your feed.*" –P9 (15-year-old, Male)
Educational Warnings and Penalty (38%, N=8)	Preventive Educational Warning (N=3)	“*The system sends them a message telling [them] that they've asked for personal info and* ***it reminds them not to give or ask for personal info****...*" –P18 (14 yr. old, Male)
	Educational Guidelines (N=4)	“***Establishing guidelines*** *on what can or should or shouldn't be talked about in these social media settings.*” –P8 (15 yr. old, Female)
	Risk Penalty (N=4)	“*We give you a* ***5-minute penalty****... If you're doing the same thing again, we'll give you a second one [penalty]... For the third warning, they no longer wait for you to improve and ban you for three days*” –P10 (13 yr. old, Male)
Guided Actions (33%, N=7)	Safety Action Recommendations (N=4)	“*When they click on that [unsafe message]* ***it tells you to block*** *or more information [pop-up]...*" –P15 (13 yr. old, Female)
	Response Recommendations (N=3)	“*The* ***virtual assistant will help you [the user] respond*** *to these [unsafe situations] type situations [unsafe situations]...*" –P2 (17 yr. old Female)

and designed a warning which allowed you to warn your friends or receive warnings from them on contact with an unsafe user. This would allow teens to be cautious of any further interaction with the reported user. P13 also designed a contact warning based on her interaction with a stranger, claiming to be a mutual friend.

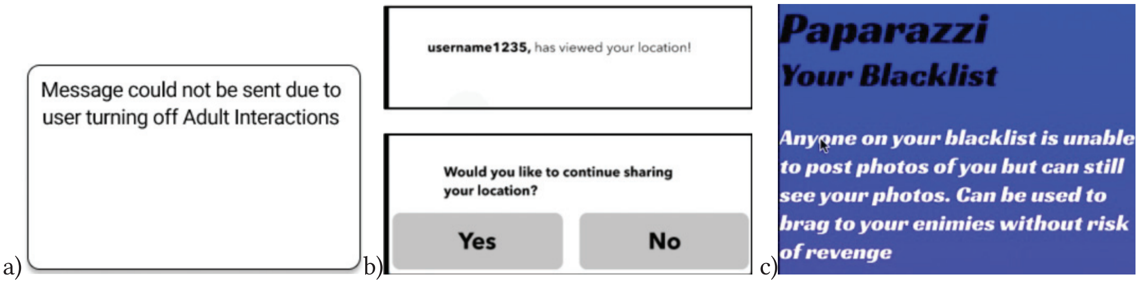
> "*So, it would get triggered within new messages. It will send a notification on whether or not you know this account, and you're sure that you know, you can trust it" –P13 (16 yr. old, Female*)

Along with the warning, P13 designed a feature that allowed teens to directly reach out to other friends or family to confirm if they knew the stranger claiming to have some mutual connection with the teen. In summary, a most of the teens designed alerts for the teen to be warned about unsafe content and potentially unsafe or uncomfortable interactions with strangers.

Proactive Warnings and Restriction for the Perpetrator. More than half of the teens (52%, N=11) designed warnings for the risk perpetrator which prevented the risk from being sent. Some of these warnings were aimed towards unsafe content, i.e., the person perpetuating the risk would receive a warning to reconsider their actions when unsafe language or content is detected. Other warnings prevented the teen from being contacted by certain specified individuals, for example strangers, adults, or known predators. In such a scenario, the teen would have the option to turn off messages or contact from the specified group in their privacy setting. For example, P5 designed a feature in the settings which allowed teens to turn off interactions with adults over a certain age, to avoid predatory messages. When someone from the limited age group tried to contact the teen, their message would be blocked with a warning (**Fig. 11a**). Similarly, P1 designed such a feature for preventing access to strangers. She explained:

> "*Give them the option to go to their settings and disable accessibility to their direct messages from strangers" –P1 (17 yr. old Female*)

Some other teens designed features for removing access to certain information or preventing certain actions from users. For example, P12 designed a warning that popped up when their location was viewed by someone (through Snapchat maps), providing them with the option to remove location access from that user (**Fig. 11b**). Similarly, P13 designed a blacklist feature for a Poparazzi app that disabled blacklisted users from the ability to post pictures of the teen (**Fig. 11c**). Another design for limiting access was a trusted circle, designed to limit sharing personal information and activity to close friends and family only. In summary, teens designed a variety of features for preventing the risk perpetrator from initiating unsafe actions or restricting the perpetrator's ability to interact with the teen.



Message could not be sent due to user turning off Adult Interactions

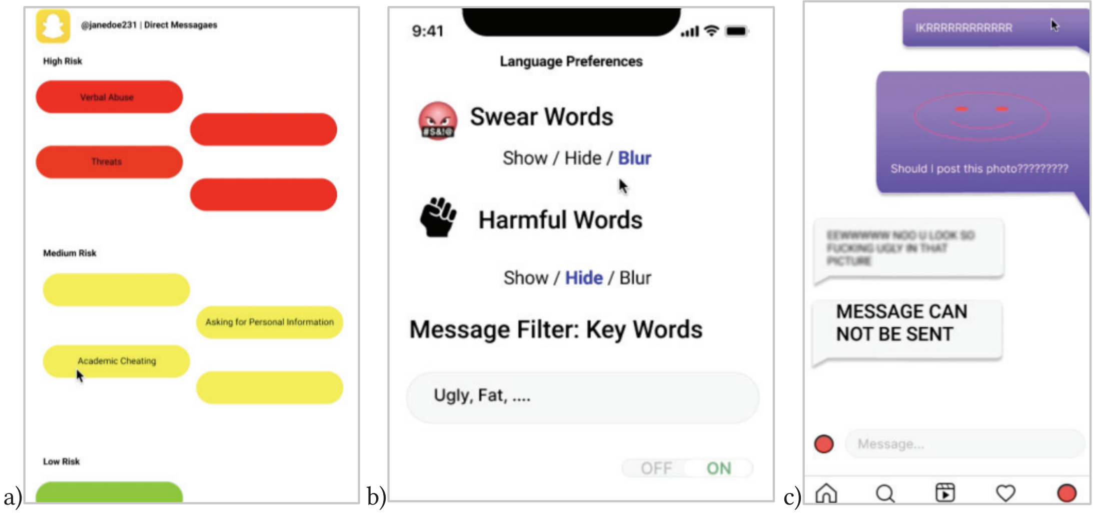
username1235, has viewed your location!

Would you like to continue sharing your location?

Anyone on your blacklist is unable to post photos of you but can still see your photos. Can be used to brag to your enimies without risk of revenge

Fig. 11. a) Alert showing prevented adult messages, b) nudge for disabling location access from certain users, c) a list of users with limited access to the teens' content.

Sensitivity Filters for Risk Prevention. Several teens also (47.6%, N=10) designed personalized ways to filter out or censor negative and unsafe content that made them feel uncomfortable. A few of these teens designed automatic censorship features where the app detected the unsafe content and censored it without much intervention from the teen. For example, some teens designed an option in the privacy settings that allowed them to toggle between turning swear words and harmful words on or off. Apart from harmful words, teens also designed for censoring explicit photos and content from spam links that disabled links with harmful content from being opened within the app. P3 designed a space where filtered content from all the different apps and platforms would be stored in a central location and risks were categorized based on risk type and risk level (**Fig. 12a**).



Swear Words

Show / Hide / **Blur**

Harmful Words

Show / **Hide** / Blur

Message Filter: Key Words

Ugly, Fat, ...

OFF ON

Should I post this photo?????????

EEWWWWW NOO U LOOK SO FUCKING UGLY IN THAT PICTURE

MESSAGE CAN NOT BE SENT

Fig. 12. a) A centralized location for all risks sorted by risk level and type, b) Settings for personalized filters of harmful words, c) Alert on perpetrator's side when message is filtered.

Other teens envisioned features for controlled censoring and personalized filtering of online risks, that gave them greater control over the type of content they did not want to see. These included options in the app settings which allowed them to enter specific key words of their choice that they wanted to be blurred or hidden. When the teen would be sent a message containing the specified unsafe words, the message would be censored, and they would receive a warning about the risk. For example, P6 designed filters in the settings that allowed teens to enter key words offensive to them, which would get censored, and the risk perpetrator would receive an alert informing them their message was blocked from being sent (**Fig. 12b, 12c**):

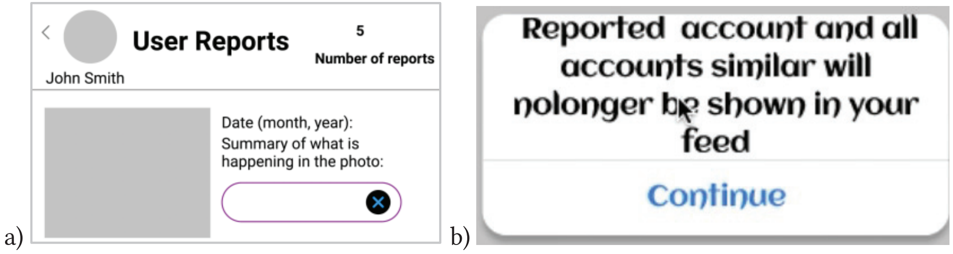
> "*Something that would prevent harmful stuff from getting to people...You get to choose... to hide it completely or you just want to blur it...*" *–P6 (15-year-old, Female)*

Majority of these teens designed filtering and censoring features in a way that still allowed the teen to view the message if they wanted to, after being warned. P21 created double warnings to ensure that the teen had considered more than once before choosing to view the risk. On the other hand, P20 created a unique filtering feature that allowed unsafe and uncomfortable messages to be filtered for the teen and sent to the parent instead. The parent had the ability to view only the unsafe or flagged content sent to the teen and approve or disapprove it from being sent to the teen. The parent also had the ability to chat with the teen via a "Family Chat" feature to discuss the filtered content before approving it. Overall, teens envisioned personalized ways of filtering sensitive or unsafe content that blocks the risk perpetrator and protects them from harm.

Robust Risk Reporting. Some teens (43%, N=9) extended the report feature to have improved functionality that facilitates safer interactions in the future. Multiple teens (N=6) designed features within reporting to upload proof of the risk, in the form of a screenshot or descriptions. Teens believed that providing proof of the encounter made the risk reports more reliable and may lead to greater accountability after reporting. A few other teens (N=2) reimagined user reports to be similar to a rating or review system where each user had a publicly accessible score based on their actions online (**Fig. 13a**). This score would be negatively impacted every time their account got reported. P4 described this feature:

> *"Once you click on his profile ... you see he has a lot of reports, you can click on user reports and then it'll show you like the number of reports he has" –P4 (17 yr. old Female*)

Teens envisioned this score to reduce online risks as it directly impacted people's reputation, as many people online get away with risks because their public facing image remains unaffected. Another teen, P9, designed a report feature that recorded the type of unsafe or uncomfortable interaction faced and key characteristics of the user that the risk was faced with. Then, the feature provided recommendations for auto-reporting other similar users, to avoid future risks (**Fig. 13b**).



Date (month, year):
Summary of what is happening in the photo:

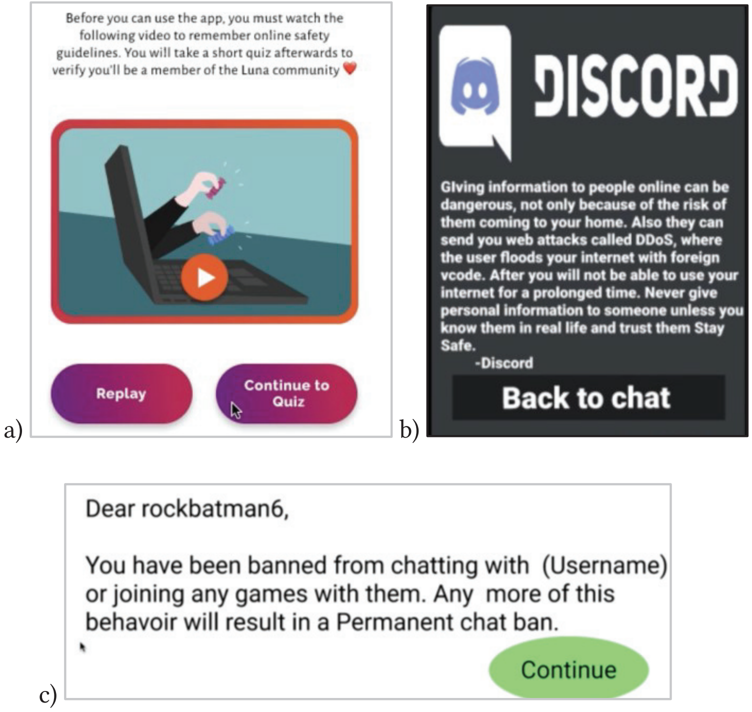
Reported account and all accounts similar will nolonger be shown in your feed

Fig. 13. a) Public user reports feature showing the total number and metadata of reports b) Prompt for auto-reporting similar users.

Educational Warnings and Penalty. Some teens (38%, N=8) designed educational warnings or guidelines to promote safer actions online. These included general online safety education (N=4) through mandatory tutorials, informational pop-ups, virtual assistants, and/or knowledge checks on online risks and safety (**Fig. 14a**). For instance, P8 designed a video tutorial about online safety and potential consequences of online risks that teens would have to watch before they start using a social media app. After watching the tutorial, teens had to take a screening test to ensure that they paid attention and understand the guidelines provided on online safety. P8 explained that there was a need to establish guidelines before teens got access to social media:

> *"Establishing guidelines on what can or should or shouldn't be talked about in these social media settings is important"* –P8 (15 yr. old, Female)

Other teens came up with guidelines within the app that provided reminders on the visibility of their actions by content moderators or potentially their school/employer (**Fig. 14b**). P19, designed an informational assistant on Discord that provided timely education relevant to the risk encountered. For example, the assistant reminded the teen about the dangers of sharing personal information with strangers when asked for his address. Some teens also designed these informational nudges for educating the risk perpetrator which might influence them to change their behavior or avoid perpetuating risks in the future. Often, these nudges were designed with additional features that may hold the risk perpetrator accountable (**Fig. 14c**). For example, several teens (N=4) came up with penalty features that warned the risk perpetrator that they may lose access to the site or be penalized in other ways if they continued with the risk. P10 designed a three-step warning, in which the penalties increased incrementally if the user repeated the unsafe behavior; first they would get a 5-minute penalty, then a 3-day penalty, finally a permanent ban.



Before you can use the app, you must watch the following video to remember online safety guidelines. You will take a short quiz afterwards to verify you'll be a member of the Luna community ❤️

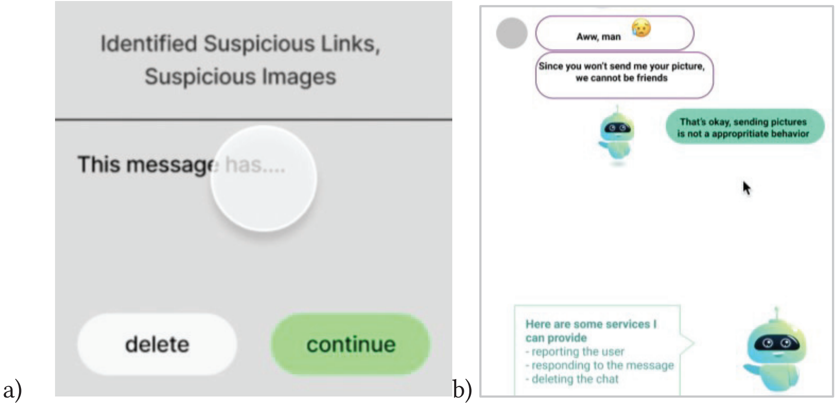
GIving information to people online can be dangerous, not only because of the risk of them coming to your home. Also they can send you web attacks called DDoS, where the user floods your internet with foreign vcode. After you will not be able to use your internet for a prolonged time. Never give personal information to someone unless you know them in real life and trust them Stay Safe.
-Discord

Dear rockbatman6,

You have been banned from chatting with (Username) or joining any games with them. Any more of this behavoir will result in a Permanent chat ban.

Fig. 14. a) Video Tutorial, b) Educational guidelines and quiz, c) Incremental penalty and ban.

Guided Actions. Many of the alerts or prompts designed by teens included guided actions for safety (33%, N=7). Recommendations for safe actions ranged from blocking, reporting, deleting, to getting help from friends, or education about the risks (**Fig. 15a**). In addition to options for safety actions, some teens designed features for live help with responding safely during an unsafe conversation. Teens especially felt the need for this live help when the risks happened with friends or people they personally knew. This included a virtual assistance that could take over the conversation if the teens wanted, and the virtual assistance would respond to the unsafe or uncomfortable encounter on behalf of the teen (**Fig. 15b**). Other ideas were focused on automated suggestions for responses that the teens may choose and modify as needed. Alternatively, P21 designed live help for teens via a parent portal through which teens could connect with parents when feeling unsafe and get their advice on how to respond safely.



Identified Suspicious Links, Suspicious Images

This message has....

Aww, man 😥

Since you won't send me your picture, we cannot be friends

That's okay, sending pictures is not a appropritiate behavior

Here are some services I can provide
- reporting the user
- responding to the message
- deleting the chat

Fig. 15. a) General recommendations for safe actions, b) Virtual assistant providing optional automated responses to the risk encounter.

5 DISCUSSION

In this section, we discuss the implications of our findings and provide recommendations for designing new online safety solutions.

5.1 Designing Evidence-based Online Safety Interventions for When Teens Need Them the Most (RQ1)

Our study revealed the types of interactions that make teens feel uncomfortable and unsafe online. Most of the unsafe interactions storyboarded by teens were information breaches in private messages that involved strangers. An explanation for this might be that teens may be conditioned to a "stranger danger" mentality through prevention programs and parents [6,35]. From a young age, children are often taught to stay away from strangers because they can bring them harm. Therefore, teens might need to be made aware that risks can also involve closer relationships, like family and friends. Similarly, in responding to risks, teens mostly used these approaches (e.g., block, report, delete) with strangers and less often used with people they knew. Similar to offline interactions [10], teens hesitated to take actions against or stand up to acquantances and friends online. This might be a result of teens feeling peer pressure and need to feel accepted and valued by their friends, which aligns with Hartikainen et al.'s work on how teens' feel pressured into risk-taking behavior, like sexting, despite saying no to their friends or partners [27]. Overall, our findings indicate the need to empower teens with features to recognize and deal with risks, even more when it happens with people they care about.

Moreover, the strategies implemented currently by teens were not effective in mitigating the risks they described. Despite using multiple responses to a risk, teens stated that their responses often led to more persistent and continued risks. Teens expected their responses (e.g., block, report) to not just protect them in that moment, but to provide continued safety for similar situations. Moreover, teens expressed a sense of distrust with the existing safety mechanisms such as reporting. Teens wanted increased accountability and evidence-based mechanisms for actions like reporting, and wanted platforms to respond with the specific actions taken for safety. Such protective mechanisms have been developed in the context of intimate partner violence, where apps allow recording and reporting evidence of abuse that is admissible in the court of law [64]. We emphasize for adolescent online safety interventions to provide similar evidence-based responses to online risks that can lead to greater accountability and overall safety for teens.

However, implementing evidence-based approaches is challenging as teens faced risks mostly in private conversations, which are often end-to-end encrypted (E2E) [20] on many of the largescale direct message platforms. Although E2E is the best standard security practice, it is at odds with ensuring online safety for teens, as it cannot detect most online risks faced by teens that happen in private direct messages. Recently, the EU government released a document that indicates it may require social media companies to compromise end-to-end encryption to scan private messages for the purpose of detecting child sexual abuse online [13]. The current debate on this topic has far-reaching consequences for policymakers, social media companies, and teens online safety. We implore future adolescent online safety researchers to investigate solutions for implementing online safety interventions without compromising privacy and security, in innovative ways so that the benefits outweigh the risks, and teens online safety is prioritized.

5.2 Implications for Design: Towards Risk Prevention and Collective Safety (RQ2)

Our study proposes a shift in the way research and practitioners view online safety solutions, as most of the existing interventions and proposed solutions aimed towards teens are ineffective. As such, our research calls for new ways to address adolescent online safety at the root cause. Therefore, we provide the following implications and comparisons with prior work for designing interventions to address the escalating unsafe scenarios teens face online:

- **Design Proactive Solutions to Safeguard Against Risk Exposure, Not Just Help Teens After-the-Fact.** Researchers and designers must identify and create solutions that can protect teens from risk exposure **before** and **during** the risk, without compromising on teen's autonomy and control. Current solutions primarily target reactive responses that happen once an unsafe interaction occurs and are often ineffective [24,25,50]. Therefore, to protect teens from these unwanted interactions, we propose new avenues for dealing with risks at the root cause and source, instead of the receiving end. For example, social media platforms could implement different levels of safety prevention towards the perpetrator, including a) warnings, b) education, c) blocking, and d) punishment, where the severity of the prevention level may depend on the frequency of risk behavior. In terms of punishment, this could be incremental penalties (e.g., removing app access or permanent ban from the app) that could be given to perpetrators if similar unsafe actions are repeated several times.

- **Design for Guidance, Not Just Risk Warnings**. In the future, designers and practitioners should design for live assistance and help teens with responding to the person after the risk. This may be in the form of intelligent recommendations that guide the teen on how to respond to an unsafe interaction, such as disengage with the conversation, block or report the person, or inform a parent or friend. The novelty of this design emerges from the types of recommendations the system will make. Systems should move beyond generic recommendations of reporting and blocking recommended by prior work [6], to more timely suggestions on *how to respond* back to the person. This could be guidance on saying "No," since they often struggled with this. It could also be features that work similar to autocorrect recommendations, but instead provide teens with suggestions for safe responses. These recommendations may further be facilitated by explanations on why that may be the best course of action in that moment, to help teens make the right choice. At the same time, teens do not want the automated systems to make the decisions for them [6], therefore we must design the tool in a way they will remain in control over their actions. For example, assistance should be provided in a way that teens have the ability to edit and rephrase the recommendations. Assistance can also be redirected to people that teens trust, such as prompting them to reach out to a parent or a friend for advice. Overall, assistance in unsafe situations needs to go beyond risk warnings, as teens want guidance on safe actions to take.

- **Design for Good Digital Citizenship and Empowerment, Not Victim Protection**. To help teens go beyond personal safety to ensure collective safety within their community, social media platforms should provide mandatory education for online safety that can help create awareness about the real-world consequences of unsafe behaviors online. While big technology companies have began to create trainings focused on Digitial Citizenship [42], they are often tailored for school curriculums or more broader audiences (e.g., parents, teachers, and youth). Instead, we should take on more novel approaches that embed training within social media platforms and are tailored towards adolescents. These trainings will not only teach teens how to manage their own online experiences, but they will also teach teens them about the consequences of their actions and the impact their behaviors can have on their broader community. Additionally, we need to design for accountability, through tracking safe and unsafe actions and more public-facing actions if users did not adhere to the community guidelines. For example, a public rating and score system for users that could hold them accountable for their actions, with negative impact on scores that is transparent to the community. Moreover, penalties may be included for perpetuating harm online, in a way that warns people to rethink their actions. In designing for community protection, teens should be provided the ability to nudge their friends about potentially risky user if they faced unpleasant experiences with them. Moreover, features such as "Close Friends" should be promoted for teens who want to engage in safer, trusted circles online. Overall, we need to design for online safety by prioritizing accountability, community safety, and negative consequences for harmful actions.

5.3 Online Safety Nudges: Moving Beyond the Status Quo (RQ3)

The design-based solutions teens developed in our study demonstrated their desires for a holistic approach towards online safety interventions. Rather than focusing solely on reactive solutions that only work *after* a risk has been encountered, teens also designed for novel preventative approaches (i.e., active measures that stopped the risk *before* it happened). Teens ultimately wanted interventions at every stage of the risk; ***before***, ***during***, and ***after*** the risk.

Unlike prior co-design work with children [6] that proposed automated assistance in the form of suggestions for blocking or reporting, the teens in our study envisioned "smarter" ways for automated assistance beyond the traditional safety actions. For example, they designed a real-time virtual assistant that could take over the conversation or suggest auto-responses in unsafe or uncomfortable situations. This assistance would be contextualized to the conversation, rather than generic risk warnings or alerts. Another difference between the children's and teens' designs was that children typically designed for prevention through parental control features, while the teens in our study preferred solutions that did not include parental involvement and monitoring. Instead, teens wanted to be empowered and assisted to make safer choices independently [25,49].

Additionally, these resilience-based solutions designed by teens went beyond prior work by adding a layer of control and personalization. For example, teens wanted to customize their social media experience by specifying who could contact them online and the type of content they consider risky.

Comparing our results with prior co-design work on cyberbullying [5,11], some proposed designs and interventions such as warning the cyberbully about the consequences and filtering offensive content, were similar. However, the novelty of our work is that many of the solutions in prior work were about consequences and coping *after* the risk. Whereas, in our work, along with penalties and consequences after the risk, most of the teens wanted to prevent risks *before* they occurred, through multiple warnings to the risk perpetrator and some recommendations for forcefully halting the unsafe action. The teens in our study also designed ways to be protected from risk exposure *during* the risk. Features designed for this intermediate stage, between sending and receiving of a risk, were aimed towards cautioning the teen of the risk received and discouraging them from viewing it. Unlike prior work [5], the teens in our study preferred filtering to be based on a personalized list of words, rather than a pre-defined dictionary of abusive words. The teens in our study considered the possibility that something they find offensive may not be offensive to another teen.

Additionally, our work is the first to propose co-designed interventions with teens that provide protection beyond their own personal safety to ensure collective safety within their community. Teens designed strategies that could prevent unsafe interactions to happen to others, such as public user reports and features that allowed nudging friends and family about unsafe accounts. Teens also wanted to create awareness about real-world consequences of unsafe behaviors online by providing communities mandatory education for online safety. For example, teens can receive online training about strategies or coping mechanisms that can be used to confront an unsafe or uncomfortable situation online. Prior work has shown that some users are organically coming together as a collective to address online risks [65], but this is not facilitated by social media. Therefore, we propose new system level designs that connect and strengthen the community towards online safety.

On the other hand, research has primarily focused on designing for protection and education for victims of online risks [8]. While recognizing the importance of these efforts, we call for holistic approaches to online safety education where perpetrators can be provided corrective training on online behaviors that are considered unsafe. By providing education to both the perpetrator and the teen, we reach the intermediate stage, between sending and receiving of a risk, in which we can help the teen navigate an unsafe situation, but also discourage a perpetrator from creating harm. These strategies towards collective safety present a new perspective to adolescent online safety in which responsibility shifts from the individual to the community.

5.4 Evaluation and Feasibility of Design-based Interventions for Online Safety

We recognize that the solutions presented in this paper cannot be considered effective until they have been implemented or evaluated. In our future work, we plan to implement nudges based on our findings from this study and evaluate them with teens in a realistic social media simulation, to understand the effectiveness of nudges in realistic scenarios. Yet, the implementation of online safety nudges is not simple. The real-time interventions recommended in our work rely on accurate detection of risk at the right time *before* teens are victimized. Razi et al.’s work on sexual risk detection emphasizes this need as they identified that currently most sexual risks are identified after-the-fact, and promote approaches for early detection that leads to risk prevention [41]. Moreover, for effectively intervening for online safety, we need a deeper understanding on how online risks evolve and the context in which these risks happen.

Apart from developmental constraints, nudges that influence users' behaviors, or interventions that prevent users' from sending or viewing certain offensive content, and others from viewing such content also pose several ethical challenges. The most prominent ethical concern around nudging is that it compromises individual's autonomy. On the other hand, risk prevention requires content moderation and censorship which are considered a breach of freedom of speech in many contexts. Yet, much of the debate around such interventions compromising freedom of choice and speech comes from a lack of control over actions or content [37]. In the context of adolescent online safety, many teens in our study proposed ways with controlled and personalized ways of filtering unsafe content, which respected their decision-making autonomy. On the risk perpetrator side, teens believed that moderating harmful content was necessary, and that compromising on individual's freedom of speech was reasonable if it protected minors from harm. Based on our findings, we recommend nudges that incrementally warn and control for harmful content being promoted or sent to teens, and eventually prevent the unsafe action. This incremental process allows the perpetrator to reconsider their actions, along with the freedom to continue, but not without consequences. Based on our findings, we recommend prioritizing online safety over freedom of speech, especially for youth as a vulnerable population, through controlled sensitivity filtering, and incremental warnings before blocking unsafe content.

5.5 Limitations and Future Work

While we provide several actionable recommendations through our co-designed interventions for online safety with teens, we recognize the limitations of our study. First, since we worked with pairs and groups of teens, their ideas for online safety nudges may be subject to social desirability bias or groupthink. Moreover, some of the teens' ideas may be biased through the material and examples provided in the training activities. For example, in the storyboarding training, teens were shown an example storyboard of a cyberbullying risk faced on a public post on Instagram, which may have encouraged them to share more about risks faced on this platform. Therefore, it is not our intention to call out any particular platform as the risk scenarios described by teens could happen on any online platform that affords similar means of communication. Further, while we had an ethnically diverse group of teens, our sample mostly consisted of relatively privileged and skilled teens with access to education, technology, and interest in UX design. This implies that our results may not represent teens from all socioeconomic backgrounds or those with less access to technology and technical skills. As such our findings may not be generalizable to all youth populations, particularly those outside of the United States. We recommend future researchers leverage co-design using similar methods with a larger, more diverse groups of youth to iterate upon these ideas for more generalizable results.

6 CONCLUSION

Our work addresses the need for strength-based online safety solutions that are designed with and for teens to empower them in their unsafe online interactions. Through novel UX bootcamps, we trained 21 teens to co-design online safety interventions in the context of relevant unsafe online interactions. Our findings show how teens found that existing ways of dealing with online risks often fail in ensuring their online safety. In moving towards their ideal safe social media experiences, teens co-designed interventions that challenge the direction of online safety, by emphasizing for features aimed towards preventing the risk sender from perpetuating harm, rather than weighing down victims with the sole responsibility of ensuring online safety. Instead, they designed for empowering teens with personalized features for risk filtering and guided actions, that help them when they are in crisis about responding to a risk, without compromising their need for control and autonomy. Lastly, teens went beyond designing for individual safety, and called for features that support collective action and accountability for a safer online world. Overall, this work calls for a radical change in how we look at, design for, and promote adolescent online safety, by proposing a shift from victim protection to risk prevention at the root cause.

ACKNOWLEDGMENTS

This research was supported by the William T. Grant Foundation (#187941, #190017) and National Science Foundation under grants CHS-1844881. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our sponsor. We are thankful for our participants**.** We would also like to thank Luke Shirley, Zinan Zhang, Oluwatomisin Obajemu, Alice Zhang, Fabrizio Martins, Camila Acevedo, Ariane Avendano for helping run the sessions, transcribe and/or analyze the data, as well as all teens who participated in the study.

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* (CHI EA ’20), 1–8. https://doi.org/10.1145/3334480.3383073

[2] Zainab Agha, Reza Ghaiumy Anaraky, Karla Badillo-Urquiola, Bridget McHugh, and Pamela Wisniewski. 2021. ‘Just-in-Time’ Parenting: A Two-Month Examination of the Bi-directional Influences Between Parental Mediation and Adolescent Online Risk Exposure. In *HCI for Cybersecurity, Privacy and Trust* (Lecture Notes in Computer Science), 261–280. https://doi.org/10.1007/978-3-030-77392-2_17

[3] J. Alemany, E. del Val, J. Alberola, and A. García-Fornes. 2019. Enhancing the privacy risk awareness of teenagers in online social networks through soft-paternalism mechanisms. *International Journal of Human-Computer Studies* 129: 27–40. https://doi.org/10.1016/j.ijhcs.2019.03.008

[4] Giuseppe Amato, Paolo Bolettieri, Gabriele Costa, Francesco la Torre, and Fabio Martinelli. 2009. Detection of images with adult content for parental control on mobile devices? In *Proceedings of the 6th International Conference on Mobile Technology, Application & Systems* (Mobility ’09), 1–5. https://doi.org/10.1145/1710035.1710070

[5] Zahra Ashktorab and Jessica Vitak. 2016. Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI ’16), 3895–3905. https://doi.org/10.1145/2858036.2858548

[6] Karla A. Badillo-Urquiola, Diva Smriti, Brenna McNally, Evan Golub, Elizabeth Bonsignore, and Pamela J. Wisniewski. 2019. Stranger Danger!: Social Media App Features Co-designed with Children to Keep Them Safe Online. In *IDC*. https://doi.org/10.1145/3311927.3323133

[7] Karla Badillo-Urquiola, Chhaya Chouhan, Stevie Chancellor, Munmun De Choudhary, and Pamela Wisniewski. 2020. Beyond Parental Control: Designing Adolescent Online Safety Apps Using Value Sensitive Design. *Journal of Adolescent Research* 35, 1: 147–175. https://doi.org/10.1177/0743558419884692

[8] 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: 231:1-231:46. https://doi.org/10.1145/3432930

[9] Yara Barrense-Dias, Lorraine Chok, Sophie Stadelmann, André Berchtold, and Joan-Carles Suris. 2022. Sending One’s Own Intimate Image: Sexting Among Middle-School Teens. *Journal of School Health* 92, 4: 353–360. https://doi.org/10.1111/josh.13137

[10] Karen Bouchard, Camilla Forsberg, J. David Smith, and Robert Thornberg. 2018. Showing friendship, fighting back, and getting even: resisting bullying victimization within adolescent girls' friendships. *Journal of Youth Studies* 21, 9: 1141–1158. https://doi.org/10.1080/13676261.2018.1450970

[11] Leanne Bowler, Eleanor Mattern, and Cory Knobel. 2014. Developing Design Interventions for Cyberbullying: A Narrative-Based Participatory Approach. https://doi.org/10.9776/14059

[12] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology.* 57–71.

[13] Jon Brodkin. 2022. "War upon end-to-end encryption": EU wants Big Tech to scan private messages. *Ars Technica.* Retrieved May 20, 2022 from https://arstechnica.com/tech-policy/2022/05/war-upon-end-to-end-encryption-eu-wants-big-tech-to-scan-private-messages/

[14] Jennifer E. Copp, Elizabeth A. Mumford, and Bruce G. Taylor. 2021. Online sexual harassment and cyberbullying in a nationally representative sample of teens: Prevalence, predictors, and consequences. *Journal of Adolescence* 93: 202–211. https://doi.org/10.1016/j.adolescence.2021.10.003

[15] Rikke Friis Dam and Teo Yu Siang. 5 Stages in the Design Thinking Process. *The Interaction Design Foundation.* Retrieved October 13, 2021 from https://www.interaction-design.org/literature/article/5-stages-in-the-design-thinking-process

[16] Arianna Davis. 2020. Co-Designing "Teenovate": An Intergenerational Online Safety Design Team. *Honors Undergraduate Theses.* Retrieved from https://stars.library.ucf.edu/honorstheses/847

[17] Katie Davis, David P. Randall, Anthony Ambrose, and Mania Orand. 2015. 'I was bullied too': stories of bullying and coping in an online community. *Information, Communication & Society* 18, 4: 357–375. https://doi.org/10.1080/1369118X.2014.952657

[18] Allison Druin. 2002. The role of children in the design of new technology. *Behaviour & Information Technology* 21, 1: 1–25. https://doi.org/10.1080/01449290210147484

[19] Lee B. Erickson, Pamela Wisniewski, Heng Xu, John M. Carroll, Mary Beth Rosson, and Daniel F. Perkins. 2016. The boundaries between: Parental involvement in a teen's online world. *Journal of the Association for Information Science and Technology* 67, 6: 1384–1403. https://doi.org/10.1002/asi.23450

[20] Ksenia Ermoshina, Francesca Musiani, and Harry Halpin. 2016. End-to-End Encrypted Messaging Protocols: An Overview. In *Internet Science* (Lecture Notes in Computer Science), 244–254. https://doi.org/10.1007/978-3-319-45982-0_22

[21] World Leaders in Research-Based User Experience. The Definition of User Experience (UX). *Nielsen Norman Group.* Retrieved October 13, 2021 from https://www.nngroup.com/articles/definition-user-experience/

[22] Maite Garaigordobil and Vanesa Martínez-Valderrey. 2015. Effects of Cyberprogram 2.0 on "face-to-face" bullying, cyberbullying, and empathy. *Psicothema*, 27.1: 45–51. https://doi.org/10.7334/psicothema2014.78

[23] Douglas A. Gentile, Amy I. Nathanson, Eric E. Rasmussen, Rachel A. Reimer, and David A. Walsh. 2012. Do You See What I See? Parent and Child Reports of Parental Monitoring of Media. *Family Relations* 61, 3: 470–487. https://doi.org/10.1111/j.1741-3729.2012.00709.x

[24] Arup Kumar Ghosh, Karla A. Badillo-Urquiola, Heng Xu, Mary Beth Rosson, John M. Carroll, and Pamela Wisniewski. 2017. Examining Parents' Technical Mediation of Teens' Mobile Devices. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17 Companion), 179–182. https://doi.org/10.1145/3022198.3026306

[25] Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J. LaViola Jr, and Pamela J. Wisniewski. 2018. Safety vs. Surveillance: What Children Have to Say about Mobile Apps for Parental Control. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–14. https://doi.org/10.1145/3173574.3173698

[26] Arup Kumar Ghosh, Charles E. Hughes, and Pamela J. Wisniewski. 2020. Circle of Trust: A New Approach to Mobile Online Safety for Families. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), 1–14. https://doi.org/10.1145/3313831.3376747

[27] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. &#x2018;If You Care About Me, You'll Send Me a Pic&#x2019; - Examining the Role of Peer Pressure in Adolescent Sexting. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, New York, NY, USA, 67–71. Retrieved December 15, 2021 from https://doi.org/10.1145/3462204.3481739

[28] Monique M. Hennink, Bonnie N. Kaiser, and Vincent C. Marconi. 2017. Code Saturation Versus Meaning Saturation: How Many Interviews Are Enough? *Qualitative Health Research* 27, 4: 591–608. https://doi.org/10.1177/1049732316665344

[29] Alexis Hiniker, Sarita Y. Schoenebeck, and Julie A. Kientz. 2016. Not at the Dinner Table: Parents' and Children's Perspectives on Family Technology Rules. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW '16), 1376–1389. https://doi.org/10.1145/2818048.2819940

[30] Sonia Livingstone and Peter K. Smith. 2014. Annual Research Review: Harms experienced by child users of online and mobile technologies: the nature, prevalence and management of sexual and aggressive risks in the digital age. *Journal of Child Psychology and Psychiatry* 55, 6: 635–654. https://doi.org/10.1111/jcpp.12197

[31] Sonia Livingstone, Mariya Stoilova, and Svetlana Smirnova. 2021. Can the internet be age appropriate, or at least not inappropriate or harmful? The promise of age verification and parental control tools. *EuConsent*. Retrieved July 9, 2022 from https://euconsent.eu/can-the-internet-by-age-appropriate-or-at-least-not-inappropriate-or-harmful-the-promise-of-age-verification-and-parental-control-tools/

[32] Edward J. Markey. 2021. Text - S.2918 - 117th Congress (2021-2022): KIDS Act. Retrieved July 7, 2022 from http://www.congress.gov/

[33] Hiroaki Masaki, Kengo Shibata, Shui Hoshino, Takahiro Ishihama, Nagayuki Saito, and Koji Yatani. 2020. Exploring Nudge Designs to Help Adolescent SNS Users Avoid Privacy and Safety Threats. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), 1–11. https://doi.org/10.1145/3313831.3376666

[34] Brenna McNally, Priya Kumar, Chelsea Hordatt, Matthew Louis Mauriello, Shalmali Naik, Leyla Norooz, Alazandra Shorter, Evan Golub, and Allison Druin. 2018. Co-designing Mobile Online Safety Applications with Children. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 1–9. https://doi.org/10.1145/3173574.3174097

[35] Raymond G. Miltenberger, Victoria A. Fogel, Kimberly V. Beck, Shannon Koehler, Rachel Shayne, Jennifer Noah, Krystal McFee, Andrea Perdomo, Paula Chan, Danica Simmons, and Danielle Godish. 2013. Efficacy of the Stranger Safety Abduction-Prevention Program and Parent-Conducted in Situ Training. *Journal of Applied Behavior Analysis* 46, 4: 817–820.

[36] Monica Anderson. 2018. A Majority of Teens Have Experienced Some Form of Cyberbullying. *Pew Research Center: Internet, Science & Tech*. Retrieved September 14, 2021 from https://www.pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/

[37] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11: 4366–4383. https://doi.org/10.1177/1461444818773059

[38] Anthony Pinter, Pamela Wisniewski, Heng Xu, Mary Beth Rosson, and John Carroll. 2017. Adolescent Online Safety: Moving Beyond Formative Evaluations to Designing Solutions for the Future. 352–357. https://doi.org/10.1145/3078072.3079722

[39] Afsaneh Razi. 2022. Instagram Data Donation: A Case Study on Collecting Ecologically Valid Social Media Data for the Purpose of Adolescent Online Risk Detection. 9.

[40] 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* (CHI '20), 1–13. https://doi.org/10.1145/3313831.3376400

[41] 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. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2: 465:1-465:38. https://doi.org/10.1145/3479609

[42] sadadow. Digital citizenship: Prepare today's learners for online success - Training. Retrieved October 12, 2022 from https://learn.microsoft.com/en-us/training/courses/digital-citizenship-prepare-todays-learners

[43] Wonsun Shin and Hyunjin Kang. 2016. Adolescents' privacy concerns and information disclosure online: The role of parents and the Internet. *Computers in Human Behavior* 54: 114–123. https://doi.org/10.1016/j.chb.2015.07.062

[44] Richard H. Thaler and Cass R. Sunstein. 2008. *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press, New Haven, CT, US.

[45] Ashley Walker, Yaxing Yao, Christine Geeng, Roberto Hoyle, and Pamela Wisniewski. 2019. Moving beyond "one size fits all." *Interactions.* Retrieved January 16, 2020 from https://dl.acm.org/doi/abs/10.1145/3358904

[46] B. Wirth, Doctoral Student, Nora J. Rifon, Ph D, Robert Larose Ph D, Melissa L. Lewis, and Doctoral Student. Promoting Teenage Online Safety with an i-Safety Intervention: Enhancing Self-efficacy and Protective Behaviors.

[47] Pamela Wisniewski. 2018. The Privacy Paradox of Adolescent Online Safety: A Matter of Risk Prevention or Risk Resilience? *IEEE Security Privacy* 16, 2: 86–90. https://doi.org/10.1109/MSP.2018.1870874

[48] Pamela Wisniewski, Arup Kumar Ghosh, Heng Xu, Mary Beth Rosson, and John M. Carroll. 2017. Parental Control vs. Teen Self-Regulation: Is there a middle ground for mobile online safety? In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17), 51–69. https://doi.org/10.1145/2998181.2998352

[49] Pamela Wisniewski, Haiyan Jia, Na Wang, Saijing Zheng, Heng Xu, Mary Beth Rosson, and John M. Carroll. 2015. Resilience Mitigates the Negative Effects of Adolescent Internet Addiction and Online Risk Exposure. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 4029–4038. Retrieved September 14, 2021 from https://doi.org/10.1145/2702123.2702240

[50] 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 *CHI 2016 - Proceedings, 34th Annual CHI Conference on Human Factors in Computing Systems*, 3919–3930. https://doi.org/10.1145/2858036.2858317

[51] Jason Yip, Elizabeth Foss, and Mona Guha. 2012. *Co-Designing with Adolescents.*

[52] Marc A. Zimmerman. 2013. Resiliency Theory: A Strengths-Based Approach to Research and Practice for Adolescent Health. *Health Education & Behavior* 40, 4: 381–383. https://doi.org/10.1177/1090198113493782

[53] 2014. Developing Design Interventions for Cyberbullying: A Narrative-Based Participatory Approach. In *iConference 2014 Proceedings.* https://doi.org/10.9776/14059

[54] 2021. Safety Resources for Parents, Guardians, and Caregivers. *TikTok*. Retrieved July 6, 2022 from https://www.tiktok.com/safety/en/guardians-guide/

[55] 2021. New Teen Safety Features and "Take a Break" on Instagram. *Meta.* Retrieved April 24, 2022 from https://about.fb.com/news/2021/12/new-teen-safety-tools-on-instagram/

[56] 2022. Safety by design to keep children safe online. Retrieved July 9, 2022 from https://www.terredeshommes.nl/en/latest/safety-by-design-to-keep-children-safe-online

[57] Parental Guide for Teens on Instagram | About Instagram. Retrieved July 9, 2022 from https://about.instagram.com/community/parents

[58] How to cope and build online resilience? - LSE Research Online. Retrieved May 24, 2022 from http://eprints.lse.ac.uk/48115/

[59] Nudge Theory Examples In Online Discussions. *OpenWeb*. Retrieved April 24, 2022 from https://www.openweb.com/blog/openweb-improves-community-health-with-real-time-feedback-powered-by-jigsaws-perspective-api

[60] AhaSlides - Live Polls, Quiz and Q&A with your audience for FREE. *AhaSlides.* Retrieved October 12, 2021 from https://ahaslides.com/

[61] What is Ideation? *The Interaction Design Foundation.* Retrieved October 13, 2021 from https://www.interaction-design.org/literature/topics/ideation

[62] FigJam is an online whiteboard for teams to explore ideas together. *Figma.* Retrieved October 12, 2021 from https://www.figma.com/figjam/

[63] Figma: the collaborative interface design tool. *Figma.* Retrieved October 12, 2021 from https://www.figma.com/

[64] DocuSAFE Documentation and Evidence Collection App. *Technology Safety.* Retrieved July 13, 2022 from https://www.techsafety.org/docusafe

[65] The Fibreculture Journal: 22 | FCJ-157 Still 'Searching for Safety Online': collective strategies and discursive resistance to trolling and harassment in a feminist network. Retrieved October 12, 2022 from https://twentytwo.fibreculturejournal.org/fcj-157-still-searching-for-safety-online-collective-strategies-and-discursive-resistance-to-trolling-and-harassment-in-a-feminist-network/

Received July 2022; revised October 2022; accepted January 2023.