Tricky vs. Transparent: Towards an Ecologically Valid and Safe Approach for Evaluating Online Safety Nudges for Teens

Zainab Agha
Vanderbilt University
Nashville, Tennessee, USA
zainab.agha@vanderbilt.edu

Jinkyung Park
Vanderbilt University
Nashville, USA
jinkyung.park@vanderbilt.edu

Ruyuan Wan
University of Notre Dame
South Bend, USA
rwan@nd.edu

Naima Samreen Ali
Vanderbilt University
Nashville, USA
naima.samreen.ali@vanderbilt.edu

Yiwei Wang
Vanderbilt University
Nashville, USA
yiwei.wang.1@vanderbilt.edu

Dominic DiFranzo
LeHigh University
Bethlehem, USA
djd219@lehigh.edu

Karla Badillo-Urquiola
University of Notre Dame
South Bend, Indiana, USA
kbadillou@nd.edu

Pamela J. Wisniewski
Vanderbilt University
Nashville, Tennessee, USA
pam.wisniewski@vanderbilt.edu

ABSTRACT

HCI research has been at the forefront of designing interventions for protecting teens online; yet, how can we test and evaluate these solutions without endangering the youth we aim to protect? Towards this goal, we conducted focus groups with 20 teens to inform the design of a social media simulation platform and study for evaluating online safety nudges co-designed with teens. Participants evaluated risk scenarios, personas, platform features, and our research design to provide insight regarding the ecological validity of these artifacts. Teens expected risk scenarios to be subtle and tricky, while also higher in risk to be believable. The teens iterated on the nudges to prioritize risk prevention without reducing autonomy, risk coping, and community accountability. For the simulation, teens recommended using transparency with some deceit to balance realism and respect for participants. Our meta-level research provides a teen-centered action plan to evaluate online safety interventions safely and effectively.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

Adolescent Online Safety, Nudges, Interventions, Behavior Change, Evaluations, User Personas, Simulations, Ecological Validity

ACM Reference Format:
Zainab Agha, Jinkyung Park, Ruyuan Wan, Naima Samreen Ali, Yiwei Wang, Dominic DiFranzo, Karla Badillo-Urquiola, and Pamela J. Wisniewski. 2024. Tricky vs. Transparent: Towards an Ecologically Valid and Safe Approach for Evaluating Online Safety Nudges for Teens. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3613904.3642313

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642313

1 INTRODUCTION

There have been numerous efforts to design and develop adolescent online safety solutions to help teens deal with common online risks, such as cyberbullying [11], information breaches [2], explicit content [61], and sexual risks [67], amongst others. These efforts have resulted in interventions ranging from parental controls and monitoring approaches [35, 54], to more strength-based solutions, such as real-time nudges that help teens self-regulate their online risks without compromising on their decision-making autonomy [2, 25, 51]. A commonality amongst these approaches is that they are mostly centered on *designing* interventions for online safety, rather than *evaluating* these solutions for their real-world viability [63]. Developing new 'ways of knowing' is an important contribution within the Human-Computer Interaction (HCI) research community [59], and also an important endeavor in moving forward adolescent online safety research [51, 91]. While designing is an essential first step, in order for these designs to be beneficial, there is a need to evaluate these solutions, in a way that accurately depicts teens' responses to these interventions when faced with online risks.

Evaluations of technology-based interventions in other related fields, such as networked privacy and security, have emphasized the importance of leveraging 'experimental realism' [5]. Prior work simulated authentic experimental environments to evoke participants' unbiased responses [48, 92], which is a promising approach for evaluating adolescent online safety interventions targeted to youth. Yet, teens have unique developmental needs [10] and online experiences that require further investigations to ensure experimental realism and ecological validity of such research prior to implementation. As leaders in the field of HCI caution, we should

always start with observation before doing intervention or experimentation [73], particularly when our research involves vulnerable populations, such as teens [15, 81].

An underlying challenge for evaluating online safety interventions is the trade-offs between ecological validity (i.e., realism [53] and teen safety, as it is difficult to simulate online risks (e.g., sexual solicitations, cyberbullying) realistically without putting teens at risk. To overcome this challenge, researchers within the HCI community are increasingly advocating for more meta-level research (or *"research on research"*), especially when working with vulnerable populations [15, 81]. Meta-level research refers to the study of research methodologies themselves with the aim to evaluate and improve research practices to ensure effective outcomes [43]. In this study, we conducted a meta-level research study with teens to assess whether the design probes (e.g., personas, risk scenarios, and nudges) that we created as study artifacts for a future evaluation were acceptable to teens. We used these artifacts with teens to create generalizable guidelines for evaluating online safety interventions for teens in a semi-controlled (i.e., 'Wizard of Oz' approach [37]) open-source social media environment. The primary goal of this study was to understand how to expose teens to realistic risk scenarios without exposing them to real risks. The larger impetus of the study was to inform researchers in the field of online safety how to conduct such 'tricky' research studies in the future. Therefore, we pose the following high-level research questions:

- ***RQ1:*** *a) What are the contextual factors teens look for when identifying risky people online (i.e., personas)? b) What are the social cues teens use to decode risky situations (i.e., risk scenarios)?*
- ***RQ2:*** *What user goals should be supported when designing nudge-based interventions for teens to mitigate online risks within social media?*
- ***RQ3:*** *What approaches do teens recommend for designing research studies to evaluate online safety outcomes for nudge-based social media interventions?*

To answer these questions, we conducted focus groups via Zoom with teens (N = 20) between the ages of 13-18, in the United States. We used design probes to solicit teens feedback on effective online safety evaluations, including a) user personas and risk scenarios that would trigger a nudge (RQ1), b) nudges based on previously co-designed interventions with teens (RQ2), and c) the research design and interface of a social media environment (RQ3). Using the analytical lens of the Social Information Processing [84] framework, we found that teens co-designed risk scenarios that were subtle and higher in risk to be believable, perpetuated by risky personas that tricked the teen by establishing trust or shared context (RQ1). Teens recommended nudges for risk prevention through personalized sensitivity filters, with the autonomy to view the risk. Additionally, teens wanted proactive coping mechanisms, accountability for perpetrators and community guidelines for education (RQ2). To evaluate these nudges, most teens recommended measuring actual behavior changes resulting from nudges within a realistic social media environment, in a way that balances the ecological validity of the research while ensuring teen's unbiased responses and well-being (RQ3).

A key contribution of our work is that it is the first to take a meta-level research approach to deeply understand how to effectively and safely evaluate online safety interventions with teens. Moreover, our work plays a pivotal role in advancing the field of adolescent online safety toward ecologically valid evaluations by leveraging Social Information Processing (SIP) theory [84] to underscore the importance of realistic, nuanced risk scenarios and nudges that maintain youth autonomy and community accountability. More importantly, we contribute to the broader CHI community by providing evidence-based best practices and guidelines for conducting ethical, yet effective and realistic intervention-based research on sensitive topics with at-risk populations.

2 BACKGROUND

In this section, we synthesize the literature on adolescents' online risk experiences and efforts to design nudge-based interventions that promote adolescent online safety.

2.1 Understanding the Context of Adolescent Online Risk Experiences

Adolescent online risk experiences have been studied extensively in the HCI community (e.g., [8, 33, 61, 70, 89]). Teens experience various online risks, such as cyberbullying [80], exposure to explicit content [67], and problematic internet use [62]. Immediate reactions to these risks included restrictive and authoritarian approaches, such as parental controls [56], which teens find privacy-invasive as they desire independence online [50], leading to mistrust between parents and teens [68, 87]. Therefore, scholars have called for moving from restrictions towards more strength-based approaches that help build resilience among teens to self-regulate their online safety [19, 60, 88]. To design such teen-centered solutions, prior work emphasizes on using participatory design [72] as it empowers teens to have a voice in the design of online safety solutions [11, 23]. Yet, it is challenging to understand whether these strength-based approaches effectively help teens be safe online. In this regard, Pinter et al. [63] emphasized the need to conduct more summative evaluations of interventions to ensure that the solutions indeed improve teens' online safety outcomes.

One approach for evaluating interventions is to conduct a real-world randomized control trial, commonly used in the field of medicine [74]. However, as HCI researchers, we tend to be more conservative in that the solutions we build could have unintended consequences and are not often as critical in directly saving lives (e.g., COVID-19 vaccine trials). Therefore, we prefer to take a more modest approach, such as testing interventions in a semi-controlled or simulated environment, which is possible to do with modern-day technologies [13, 48]. At the same time, a main challenge with evaluating interventions is simulating the online risks. Specifically, we need to understand the context of the risk, the characteristics of the bad actors, and the social cues teens identify to simulate risks in a way that is realistic without putting teens in more danger. To do this, we build upon prior research recommendations to involve teens as partners in the design of research, by getting feedback on user personas and risk scenarios based on commonly reported risk experiences of teens in prior work [2, 89].

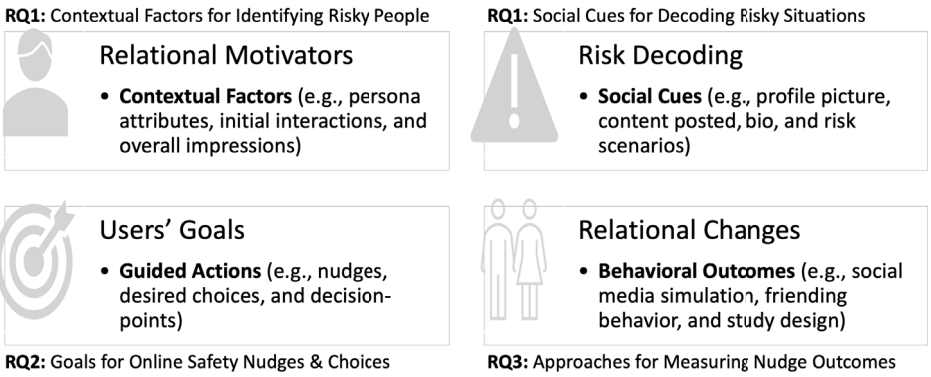


Figure 1: How the Social Information Processing (SIP) Framework guided the formation of our RQs and qualitative analyses.

2.2 Designing Nudge-based Interventions for Online Safety

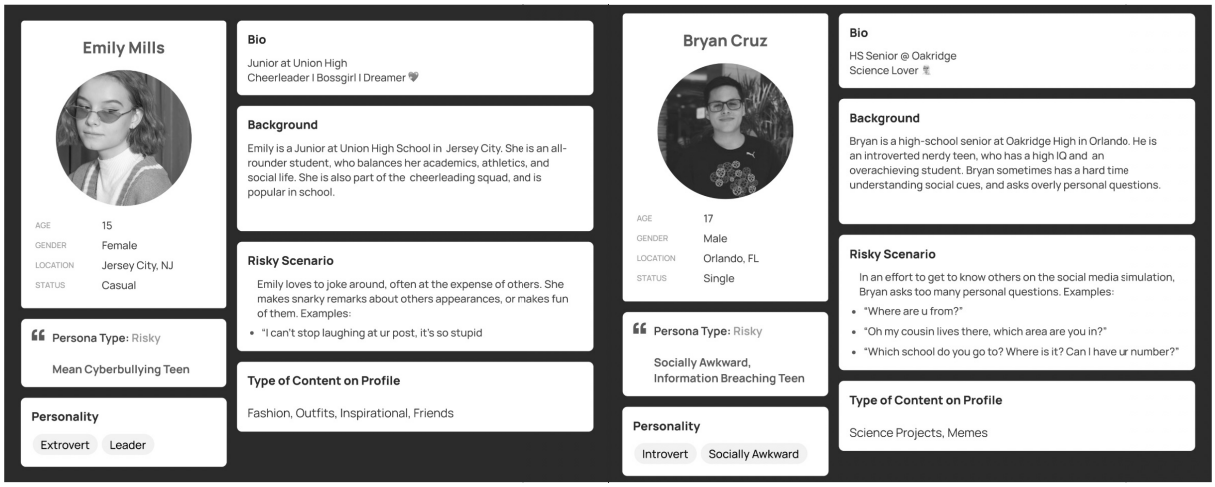
Nudges are subtle persuasive cues that aim to influence people's behavior without compromising their decision-making autonomy [77], commonly used in the fields of online privacy and security [36, 69, 75, 85]. For instance, Wang et al. [85] conducted a field study with Facebook users and found that privacy nudges that remind users to reconsider the audience of a post can effectively assist users in avoiding unintentional information sharing. While most nudges have been designed for general populations, recently researchers have investigated nudges for adolescents' online safety, for risks such as information breaches [44, 51], cyberbullying [78] and sexual risks [79]. For instance, Masaki et al. [51] found that teens envisioned nudges that can help reduce information disclosure, especially when the nudge is negatively framed and emphasizes the risk. Most recently, Agha et al. [2] conducted co-design workshops with teens and designed nudges for commonly experienced online risks. They found that teens wanted to prevent risk perpetrators by prompting them to reconsider their actions or penalizing them for perpetuating harm. Teens also designed sensitivity filters, educational guidelines, and risk alerts with guidance to help manage the risk. Yet, HCI researchers emphasize the use of co-design as an *iterative* process in which designs evolve with each cycle of feedback [14]. Therefore, there is a need to further refine previously co-designed nudges with teens before they can be implemented for evaluation. To do this, we conceptually grouped the nudge design ideas from Agha et al. [2] into four nudges broadly covering the common online risks, to understand how teens critically improve these designs for evaluation.

2.3 Research Considerations to Evaluate Nudges for Online Risk Prevention

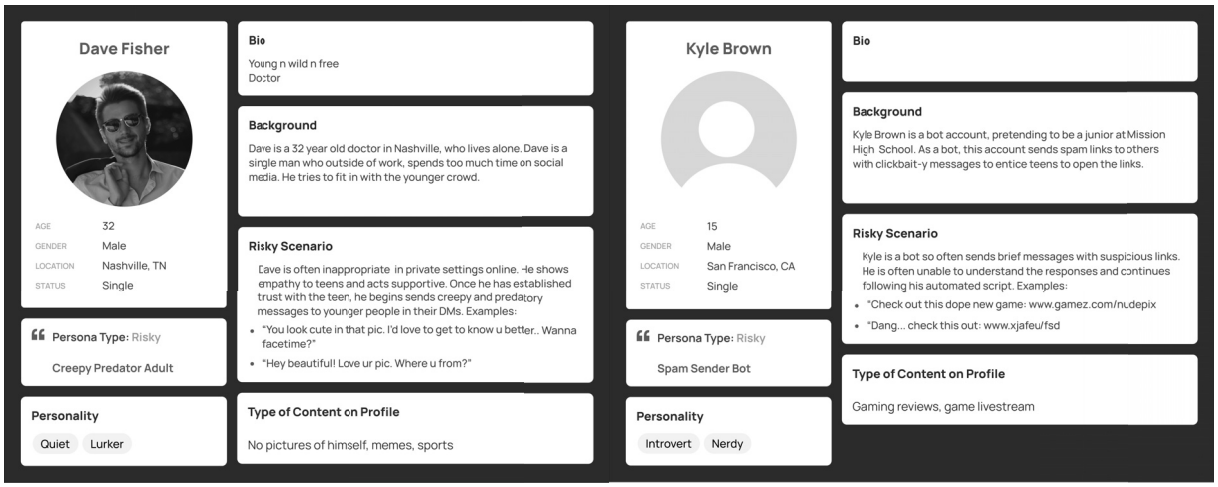
While several adolescent online safety researchers have focused on designing interventions [2, 3, 16, 17, 54], there still remains a gap in evaluating the real-world efficacy of these solutions. Although survey-based feedback is commonly used [28, 51] and provides valuable insights, they may not align with participants real-world responses [57] as they often rely on teens' hypothetical feedback on how they would potentially respond to an intervention [28, 51]. Consequently, there is a lack of ecologically valid evaluations that demonstrate teens' genuine behavior change in response to nudges [63]. Some research efforts have been made in the privacy and security domain to address this gap by evaluating interventions in more realistic settings [12, 48, 92]. For instance, Zinkus et al. [92] developed a fake social network platform for teens to learn about privacy protective choices within a realistic social media environment. However, conducting ecologically valid evaluations and simulating realistic risk scenarios for teens introduces unique ethical challenges (e.g., how can the scenarios be realistically risky, while ensuring that the research does not harm teens as a vulnerable population?) [15, 64, 81]. Additionally, researchers need to carefully conceptualize the research design in a way that does not bias the participants [27, 34]. To ensure that the research meets teens' needs, Badillo-Urquiola et al. [15] called for meta-level research in which adolescents take center stage and co-direct the research to meet their unique needs and experiences. According to Ioannidis et al. [43], meta-level insights can be beneficial at various stages of research, including studying methods, reporting, evaluation, reproducibility, and incentives of research. Prior works encouraged researchers to include vulnerable populations in such meta-research, in our case teens (ages 13-18), focusing on ethical considerations to ensure 'beneficence,' so that the benefits of the research outweigh the potential risks [15]. Extending prior work, we involved teens in co-designing ethical research practices focused on simulating online risks, applying effective nudges, and designing an ecologically valid evaluation for online safety nudges.

3 APPLYING A THEORETICAL LENS OF SOCIAL INFORMATION PROCESSING

A key aspect of simulating adolescent online risks for experimentation is to understand how adolescents process *social cues* and identify bad actors online. Therefore, in this work, we leverage the theoretical framework of Social Information Processing (SIP) defined by Walther [83], which is an extension of SIP [29] within computer-mediated communication (CMC). This framework focuses on understanding how people process social cues online to form initial impressions of other users (e.g., non-verbal and textual cues), form knowledge regarding other users' motivations and goals, as well as how they change or manage their relations with others



(a) Cyberbullying (b) Information Breaching



(c) Predatory Sexual Remarks (d) Scam Risks

Figure 2: Risky Personas Presented to Teens for Feedback

over time. Prior research has applied SIP theory in understanding adolescent behaviors on online dating platforms [90] and social tendencies of sex offenders [24]. However, a gap remains in the application of SIP to understand how youth navigate potentially risky online situations on social media when forming new relationships online. We address this gap by grounding our qualitative analyses by the SIP framework as shown in Fig. 1 and further described in our methods. We combine our SIP framing with online safety nudges built upon prior work [2], which serve as in-situ "teachable moments" [40], to examine how such interventions may affect teens' social information processing when faced with online risks. Further, we expanded this framework by developing design implications for effective evaluations of adolescent online safety that can enhance teens' self-regulated social information processing.

4 METHODS

4.1 Study Overview

We conducted focus groups with 20 adolescents (ages 13-18) via Zoom to co-design a study in which teens are exposed to risks in a simulated social media setting with other users, which would trigger nudges to be evaluated. To elicit meaningful feedback, we used design probes, which is a useful method in HCI as it provides tangible artifacts to generate insights [82]. Probes included user personas, risk scenarios, and nudge-based interventions adapted from prior co-design work with teens [2]. The probes were embedded in an adapted version an open-source social media environment called "Truman," originally developed by DiFranzo et al. [27]. Screenshots of this system were also presented as probes to solicit teens' feedback on designing a simulated environment for evaluating nudges. Teens provided annotated and verbal feedback on these designs using a collaborative online whiteboard (i.e., FigJam) so that they

Table 1: Summary of Characteristics and Scenarios for Risky, Neutral and Positive Personas

Persona Type	Scenario	Characteristics	Background
Risky	Asks too many **information breaching** questions	**Bryan** (17, M) Location: Orlando, FL Personality: Introvert, Awkward	Overachieving student, struggles with understanding social cues
	Cyberbullies by making snarky remarks about others	**Emily** (15, F) Location: Jersey City, NJ Personality: Extrovert, Leader	All-rounder student, popular and social cheerleader
	Scam bot that sends suspicious, enticing links	**Kyle** (15, M) Location: San Francisco, CA Personality: Introvert, Nerdy	Bot pretending to be a high school student with suspicious profile
	Sends creepy and **predatory messages** to teens	**Dave** (32, M) Location: Nashville, TN Personality: Quiet, Lurker	Spends too much time online and tries to fit in with youth
Neutral	**Limits personal info** and shares about hobbies online	**Sarah** (16, F) Location: Atlanta, GA Personality: Introvert, Artsy	Part of a band as a guitarist and shy in new people
	Stays **reserved** on social media and rarely posts	**Frank** (17, M) Location: Orlando, FL Personality: Introvert, Sporty	Sporty teen who loves soccer and trying skateboarding tricks
	Loves to **share about travel** and food experiences	**Maria** (16, F) Location: New York City, NY Personality: Introvert, Artsy	Loves planning trips with her friends and is a good student
Positive	Talks to and **appreciates everyone** online	**Brenda** (14, F) Location: Boston, MA Personality: Extrovert, Helper	Public speaker and loves helping her community
	Shares funny content with friends and family	**Josh** (18, M) Location: Salt Lake City, UT Personality: Extrovert, Funny	Knows how to make everyone laugh and is street smart
	Posts helpful resources and tries to be a supportive peer	**Alice** (15, F) Location: Chicago, IL Personality: Introvert, Agreeable	Writes for the school paper and helps others with their problems

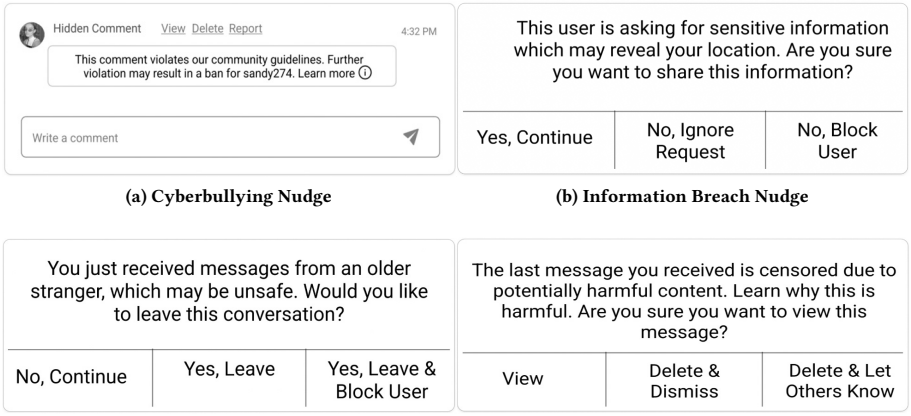
could provide direct feedback, while collaboratively generating deeper conversations to iterate beyond the designs that were provided. Each session included 1-3 teens and lasted for about two hours in total. This study was approved by the authors' Institutional Review Board (IRB) and parental consent was required for participants under the age of 18.

4.2 Study Procedure and Design Probes

We started with an introduction and icebreaker, followed by introduction to the concept of online safety interventions with examples of nudges from prior work. Then, the researchers led a group discussion on challenges of evaluating nudges effectively and realistically which can help with teens online safety. Next, we presented the idea of a social media simulation to evaluate nudges, specifying that the goal of the current study is to obtain teens feedback on different components of the simulation (e.g., user personas, nudges, platform). In the following subsections, we summarize our study procedure and provide an overview of the design probes.

4.2.1 User Personas and Risk Scenarios. Personas have been commonly used within the HCI and the User Experience (UX) research communities to help understand target users' goals, needs, and behavioral patterns through the creation of fictional but realistic characters [22]. Personas are designed in various ways, such as through grounded theory [30], empirical data [41], or assumptions based on common user traits [65] to create baseline characters refined through iterative feedback. Personas are developed following a systematic approach involving: a) defining a salient problem or goal, b) defining characteristics of the user, and c) describing a supportive narrative for the user [30]. When designing personas for adolescents, with limited data and access to this unique group, Antle [9] recommended that personas should emphasize diverse representation through pictures, personalities, backgrounds, and hobbies. Additionally, the personas should relate background experiences to adolescent needs (e.g., safety) or scenarios, which should be validated through iterative feedback from the users.

Following this process, we developed risky personas based on four prominent online risks found in Agha et al.'s [2] co-design study with teens, where teens created storyboards regarding their past online risk experiences. These risks guided our persona development, including information breaching (Bryan), sexually inappropriate messages from adult strangers (Dave), cyberbullying risks in



(a) Cyberbullying Nudge **(b) Information Breach Nudge**

(c) Predatory Risk Nudge **(d) Private Scam & Explicit Content Nudge**

Figure 3: Nudge Design Prompts

public posts (Emily), and suspicious scam from bots (Kyle), which mapped to each of the four risky personas (Fig. 2). While the main goal was to ensure that risky scenarios and personas are realistic, we added three neutral and three positive personas each to ensure a balanced and diverse set of user types to reduce negative bias in teens feedback. These six user personas covered neutral or positive characteristics based on commonly encountered social media users, with a diverse range of personality traits (e.g., friendly, shy, helpful, agreeable) and interests (e.g., traveling, sports, arts). Each persona included a profile picture, bio, personality traits, background, and risk scenarios, as detailed in Table 1. The scam risk persona (Kyle) was the only persona without a profile picture, as it represented a fake bot account. As recommended by prior research [9, 65], these personas served as a baseline with the goal of validating these personas through teens feedback.

We presented these personas to teens for the design activity, while explaining that the goal of designing user personas with them is to understand realistic scenarios and different types of users they encounter online (RQ1). This led to the first activity, in which participants were asked to redesign at least one or more of the risky personas with high-level feedback for the neutral/positive personas based on their preference, using FigJam [31]. We asked participants to provide feedback on the characteristics of the user, the risk scenarios, and social cues or quotes that fit the persona, encouraging them to redesign unrealistic scenarios that they may think are missing from the set. Feedback was provided through design annotations on FigJam, along with verbal discussions.

4.2.2 Online Safety Nudges. Next, participants were asked to provide feedback on four nudges for online safety including the nudge design, how they would respond to the nudge, and how they would change the options provided (RQ2). The nudges (Fig. 3) were aligned with each of the risky user personas, and presented with the risk scenarios. These four nudges conceptualized key ideas from prior co-design research with teens [2]. The public cyberbullying nudge (Fig. 3a) filtered a risky comment and highlighted community guidelines while giving options to view, delete, or report the risk. The private information breaching nudge (Fig. 3b) warned users of requests for location-revealing sensitive information, allowing them to continue, ignore, or block. The private predatory risk nudge (Fig. 3c) warned about inappropriate messages from a stranger, with options to continue, leave, or block the sender. The private scam & explicit content nudge (Fig. 3d) used filters to censor the risk, with choices to view, delete, and inform others.

4.2.3 Social Media Environment. Lastly, participants were asked to provide feedback on the research design and interface of a social media evaluation. We presented screenshots of a web-based social media environment, to get teens' feedback on the interface and study design (RQ3). This system builds upon Truman, developed for experimental social media research [27]. Participants were presented with the tasks of the study, the profile/bio page, a friending feature, a feed for posting and interacting with content, a researcher interface for switching between personas, as well as a chat interface for exchanging messages with other users (Fig. 4). Alongside, teens were asked to provide feedback on research design choices, including observation practices, deceptive research, and privacy concerns. For instance, we asked participants how they feel about doing a think-aloud while completing the tasks or being deceived about the study. At the end of each activity, the researchers summarized the ideas shared by teens. After the conclusion of the research session, participants were asked to complete a brief demographic survey.

4.3 Data Analysis Approach

The data collected included audio and video recordings, design annotated whiteboards on FigJam, and demographic survey data. The recordings were transcribed using Zoom transcription and manually checked for errors. The data was analyzed qualitatively using thematic qualitative analysis to answer the research questions [20], as it is a suitable approach for generating new themes and insights from the data. To answer RQ1, RQ2, and RQ3, we analyzed the redesigned and annotated user personas, nudges, and social media simulation interfaces, along with verbal feedback from teens. Our qualitative coding scheme was informed by different aspects

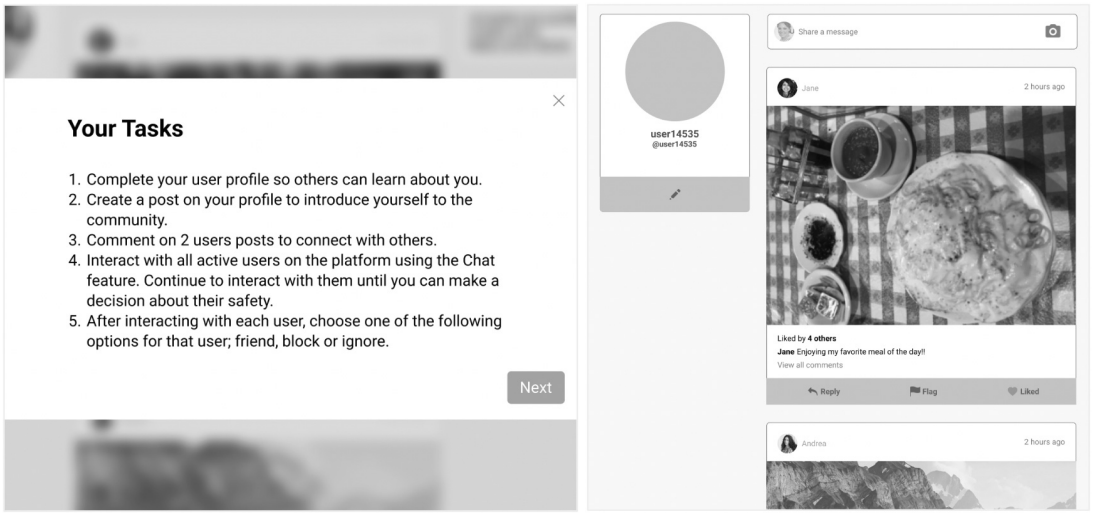


Figure 4: Tasks & Feed Interface from the Social Media Environment

of the SIP framework for each RQ (Fig. 1). To answer RQ1, we developed codes for teens feedback on *contextual factors* of a risky user (e.g., predators hide their age) and *social cues* that helped them decode the risk (e.g., bots have profile pictures to be believable). For RQ2, we coded for how teens wanted nudges to help in relation to their *user goals* (e.g., autonomous risk prevention) and how they informed *decision-making choices* (e.g., multiple options for safety). For RQ3, we coded for teens feedback on study design choices for measuring *online safety outcomes* of nudges (e.g., using deception) and how they wanted to incorporate *relational changes* over time (e.g., changes in privacy settings). The first author reviewed the transcripts and designs for the first few sessions to create the initial codebook informed by SIP. As sessions continued, the first author added new codes, merged similar codes, and completed the coding with frequent check-ins and consensus building among all co-authors. By the last session, we reached theoretical saturation, as no new codes emerged from the data, and hence, we concluded our data collection. We then further refined our codes and conceptually grouped them to create our final codebook, mapped to the SIP framework, as presented in Table 3.

4.4 Participant Recruitment and Demographics

Participants were mainly recruited through existing contacts with youth-serving organizations in the U.S., schools, and social media advertisements. We fostered these long-term relationships by ensuring that our research studies offered tangible benefits to participants, such as skill development (e.g., UX design skills) and resume building activities (e.g., advisory board roles), as well as by engaging in community engaged scholarship through presenting our research at local schools for the benefit or parents and teens. These organizations were initially contacted via email, call, and/or distributing flyers to them. The session lasted for about 2-3 hours and participants were compensated with $20 Amazon gift cards for participation. During the informed consent process, participants were reminded that their privacy and anonymity would be protected, and they could withdraw from the study at any time. A total of 20 teens completed the study who had to complete an eligibility survey to confirm that they are from the United States, between the ages of 13-18 years old, and have access to reliable internet, and video-calling capabilities. Parental consent was acquired for teens under the age of 18 before participation. The majority of the teens were between the ages of 16-17 (55%), some early teens between the ages of 13-15 (30%), and a few 18-year-old participants (15%). We had a balanced gender representation with 12 male (60%) and 8 female (40%) participants. The majority of the participants identified themselves as Asian (55%), Black/African American (25%), followed by Hispanic/Latino (15%), and White/Caucasian (5%) (Table 2).

5 RESULTS

In this section, we use illustrative quotes and annotated design probes as artifacts to illustrate the main emerging themes for answering our research questions.

5.1 Online Risks Need to Be Realistic, Tricky and Subtle, Perpetuated by Deceptive Users (RQ1)

Overall, teens were forthcoming in telling us that: 1) social media risk scenarios need to be more realistic to be believable, 2) that the characteristics of the perpetrators should more closely match the risk, and 3) that the risk posed need to be tricky and subtle, rather than overtly obvious to users. Below, we further unpack the themes and feedback teens gave us for decoding and designing for risky situations on social media.

5.1.1 Risk scenarios need to be realistically risky to be believable. Overall, many of the teens (65%, n = 13) indicated that the risk scenarios needed to be more believable and relevant to their lived experiences. For instance, almost half of the teens (40%, n = 8) thought that cyberbullying in real-life is often harsher than what

Figure 5: Summary of Teen's Design Changes for Dave's Persona

we depicted and typically about one's physical appearance. As such, many participants considered "Emily's" (cyberbullying persona and risk scenario) remarks, "*I can't stop laughing at ur post, it's so stupid*" to be too generic to be considered risky. Teens thought that such banter was common among teens, especially in friend groups. Instead, teens recommended harsher forms of cyberbullying, such as condescending remarks about one's physical appearance or body image or making fun of someone's weight, outfit, or eating disorders, namely, body-shaming. For instance, P16 recalled a cyberbullying situation where her friend was cyberbullied for being anorexic:

> *"It's always like body image, like someone claimed that they were anorexic. And then someone was like, well, you literally eat everything, you can't be."* - P16 (17-year-old, female)

Other teens believed that cyberbullying sometimes felt harsher, even if it was not targeted towards them. For instance, some cyberbullies made them feel unsafe by judging and "backbiting" others, such as *"R u actually friends with (someone), aren't they annoying?"* (P9, 13-year-old, female). A few teens also redesigned the cyberbullying persona to indicate higher social status, such as Emily being rich and bullying others by looking down on them about their social status. Some teens (20%, $n = 4$) also redesigned "Dave" (predatory persona and risk scenario) by having him share explicit content with teens. Their rationale was that sending compliments on one's appearance may seem creepy but did not reach the threshold of being an actual threat. Instead, some teens felt that the predatory risk would be more believable if the predator shared explicit content with requests for sensitive information, such as their phone number or location to meet in-person (Fig. 5). In the same vein, a few teens (15%, $n = 3$) critiqued that instead of asking generic questions (e.g., "where are you from?"), "Bryan" (information breach persona and risk scenario) should ask questions for more specific and targeted sensitive information, such as their location information, which could escalate into a predatory situation or offline risks. P20 explained,

> *"The information should at least risky enough to the point where it's not just Oh, where are you generally from? It's like, Where do you live? Can I have your phone number?"* - P20 (14-year-old, male)

Similarly, teens (20%, $n = 4$) redesigned "Kyle" (scam bot persona and risk scenario) to be higher risk through targeted scam based on teens' interests, which would feel riskier due to personalization. At the same time, teens acknowledged that such personalization might not be possible for the purposes of research and therefore recommended sending phishing links related to popular brands (e.g., Starbucks) that everyone knew, as a feasible alternative. Teens also wanted to increase the believability of Kyle by having the scam bot reshare content that was trending, making Kyle seem less suspicious, as they had seen fake accounts that stole content from smaller creators to draw less attention. In summary, teens increased

the believability of the personas and risk scenarios by increasing the severity of risk, but they also considered alternatives, so that the risk scenarios were not harmful to teens in a research study.

5.1.2 Online risks should be tricky and subtle, not obvious. Teens' feedback indicated that our risk scenarios were too obvious. Instead, online risks often happen in subtle ways that aim to trick people. For instance, many teens (55%, n = 11) suggested that Kyle (scam bot persona and risk scenario) should send scam and phishing links in a deceptive and subtle way. They found Kyle to be too phishy to be a real person, as the account did not have a photo or bio and sent scam links that were clearly suspicious. Therefore, they redesigned Kyle to be deceptive by including a photo, bio, and content on a profile. These teens also suggested that Kyle should first attempt to interact with users similar to real human accounts and then send malicious content. Moreover, they recommended that Kyle should send personalized click-bait to match the type of scam they receive online and to make the link more deceiving such as *"Hey, is this you?"... "No? can you at least check this out."* Other teens suggested making the scam link more enticing by offering money, giftcards, or gaming points, such as *"Congrats, you've won our giveaway from Target! Click here to redeem..."* (P10, 15-year-old, male). A few teens commented that such scam links often come from hacked accounts of their friends, which increased their chances of clicking the links. P12 recalled similar targeted phishing risks,

> *"I see those a lot like someone tagged me, oh, you want a giveaway, press the link and then they ask for your credit card information. My friend gave her Social Security once."* - P12 (14-year-old, female)

Additionally, almost half of teens (45%, n = 9) recommended that Dave (predatory persona and risk scenario) should trick victims by building trust first, as they felt that such risks often fell into two categories; a) stalkers who messaged you inappropriate comments out-of-the-blue, or b) predators with a an ulterior motive who slowly built trust with the teen, and befriended them before risky behavior. As such, teens recommended different strategies for trust building for Dave. For instance, a few teens recommended that such predators often try to gain the victims' sympathy by sharing personal problems or relatable "rants" (e.g., about their job), and later revealing their risky motives (e.g., requests to meet) (Fig. 5). For instance, P14 added a risky message for Dave, *"I really enjoy talking to you and would love to get to know you better...want to meet up?"* (P14, 16-year-old, male). Other suggestions included making Dave a supportive person for the teens who helps them to gain their trust, as P19 suggested,

> *"I would make him like more a counselor. Someone like who listens to people's feelings and pretends they want the best for them"* - P19 (17-year-old, male)

A few teens thought that predators often try to offer incentives to attract teens, such as showing off their belongings, career, or social life (Fig. 5). Other teens believed that Dave would make personalized comments about the teen's photos, instead of generic remarks, to get a response; they explained that teens are at a vulnerable age where they are often impressed by such compliments. Relatedly, some teens (35%, n = 7) thought that Bryan (information breach persona and risk scenario) was asking for information in very obvious ways, by immediately jumping to ask the teen's address. In contrast, they believed that such risks were often perpetuated ambiguously and redesigned Bryan to ask for information subtly, for instance, based on mutual factors (e.g., location, interests), *"Hey, did you go to Oakridge, u look kinda familiar"* (P11, 18-year-old, female). Moreover, some teens believed that such risks were often built up over time, with established rapport and shared context with the teen, before asking for their personal information. P16 explained,

> *"If they immediately put, Where is it? What's your number, that's automatic block for me. But I would give it a second thought if they're in my area or around my age."* - P16 (17-year-old, female)

Lastly, teens (30%, n = 6) changed Emily's (cyberbullying persona and risk scenario) persona to be more tricky as they thought that condescending remarks happened in subtle ways through back-handed compliments about physical appearance. For instance, one of the participants suggested that Emily should make a sarcastic comment, *"OMG that outfit would look so much better on me :)"*. Teens explained that such sarcastic comments can leave the victims confused and disturbed. Additionally, they thought that such back-handed bullying often comes from people they know. For instance, P11 questioned the intention of such cyberbullies online, as she thought that their aim is often to hurt the other person while protecting themselves without saying something too obviously harmful. P11 explained,

> *"Many people are really smart with how they say things, they won't be too direct so that they can back out and they want you to think about it so that it hurts"* - P11 (18-year-old, female)

Overall, half of the teens shared how risk severity often lies in subtlety, where risks and positive interactions are not mutually exclusive (50%, n = 10). Online risks were considered tricky, as interactions that start as positive may end up with a breach of trust or information.

5.1.3 Characteristics of the perpetrator should match the risk. The majority of teens (90%, n = 18) in our study wanted the characteristics of the perpetrators to match the risk in order to improve realism, based on experience with users they met online. Overall, teens were thoughtful about how the personalities would play a critical role in the type of interaction they expected to have with the personas. Teens paid attention to several characteristics of the risky personas, including their personalities, backgrounds, occupations, and content. For instance, several teens (40%, n = 8) pointed out that "Dave" should have a more realistic and relaxed occupation because an adult who is busy with their job would not spend that much time on social media (Fig. 5). P12 explained,

> *"It's unrealistic that someone with a busy profession would spend so much time outside of work on social media"* - P12 (14-year-old, female)

Teens also considered Dave's apparent age problematic; several teens (25%, n = 5) thought that he should hide his real age in an attempt to deceive teens and fit in with the younger crowd. A few teens (15%, n = 3) thought that his character should be narcissistic to match the type of creepy users they encounter online, by posting more photos of himself and oversharing about himself (Fig. 5). Being

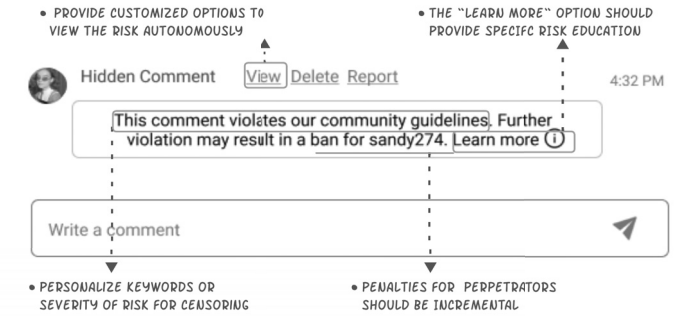


Figure 6: Summary of Feedback on the Cyberbullying Nudge

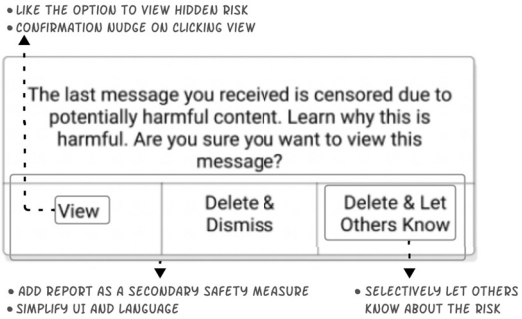


Figure 7: Summary of Feedback on the Scam Nudge

an adult, teens also imagined Dave to have a different texting style than teens, such as using too many emojis or not being familiar with their slang. Teens also changed the personality traits of "Bryan" (information breach) to better match the risk scenario. For instance, some teens (25%, n = 5) wanted Bryan to act more oblivious and naive, as an introverted teen, who would use innocence to trick the victim. This was because they thought that Bryan would benefit from *seeming* like someone who does not understand social cues and was unaware of what was appropriate to ask, leading others to give him the benefit of the doubt. P10 explained,

> *"It can leave a person who's talking with this persona thinking 'Oh, well, they're a bit naive. Maybe they didn't realize that at first.' which might lead someone to trusting them more."* - P10 (15-year-old, male)

In contrast, some teens (25%, n = 5) wanted to change Bryan's personality completely to be extroverted, as they considered that an introverted person would be less likely to ask such direct and invasive questions. Importantly, a few teens were particularly off-put by the awkward nature of Bryan's persona, as they strongly felt that socially awkward individuals should not be depicted as unsafe. This was insightful feedback from an autistic teen who participated in the study that we will take to heart. Overall, teens felt that Bryan should be deceptive by either faking being a naive teen or being an upfront extroverted personality to ask direct questions.

Teens also thought that risky users may show opposing sides of their personality in different contexts. Several teens thought that popular people care about their reputation and do not cyberbully others publicly for fear of getting "canceled." Therefore, many teens (35%, n = 7) redesigned Emily to be more "two-faced", who would bully in private, while pretending to be supportive to others in public. Moreover, some teens thought that such cyberbullies often have influential personalities and often act as the "leader" of the group. Due to the peer pressure, they often get support from others, leading to ganging up on a victim, as no one would stand up to the cyberbully. P3 described this persona,

> *"If she says something, everyone will follow. No one's brave enough to stand up ."*- P3 (16-year-old, female)

Finally, we looked at the big picture, by viewing all personas and risk scenarios (risky, neutral, positive) together. Teens were quite vocal about not agreeing with our categorizations as they thought that social media users should not be boxed into black-and-white categories like "risky," "positive," or "neutral" as many of these traits co-exist. Instead, some teens (35%, n = 7) recategorized personas based on their relationship with the person (e.g., safe users they know and trust, acquaintances, and untrusted strangers). In contrast, a few teens did not think that the people they knew were always safe (e.g., positive users can have risky traits). A few teens (20%, n = 4) decided on the safety of a user depending on past interactions (e.g., someone who supports them would be safe). Therefore, teens considered the online safety of users to be a convoluted concept, where safety was on a spectrum, rather than discrete categorizations.

5.2 Teens Redesigned Nudges for Autonomous Risk Prevention, Guidance and Accountability (RQ2)

In this section, we summarize teens' feedback to understand their mental models and goals for effective nudges, using design probes of nudges for public cyberbullying, private information breaching, predatory risk, and scam bot risk.

5.2.1 Most teens wanted proactive risk prevention but with the autonomy to override decision-making. Overall, most teens wanted proactive ways for risk prevention before the risk. For instance, most teens liked the options for sensitivity filters in the cyberbullying nudge (100%, n = 20) which hid risks in a public comment (Fig. 6), as well as nudges for hidden scam bot risks (65%, n = 13) in private chat. This is because teens wanted an automated layer of protection from the risk. In addition, several teens (50%, n = 10) wanted control of risk prevention with the option to personalize keywords for sensitivity filters based on their preferences, risk tolerance or risk severity. P7 summarized,

> *"Have a list of what words or phrases that the algorithm recognizes as harmful and that list could be edited by the user. So that could get flagged."* - P7 (16-year-old, male)

Yet, teens wanted the ability to override decision-making of nudges, as they did not want the platform to enforce censorship. In this regard, most teens liked the option to autonomously view the hidden risk for both cyberbullying (70%, n = 14) and scam bot

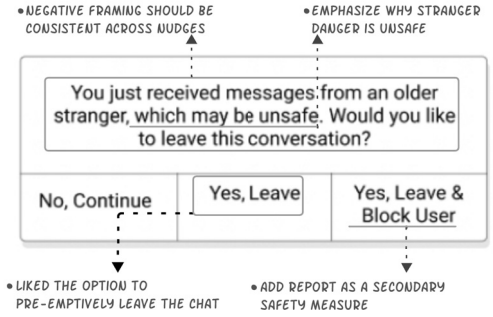


Figure 8: Summary of Feedback on the Predatory Risk Nudge

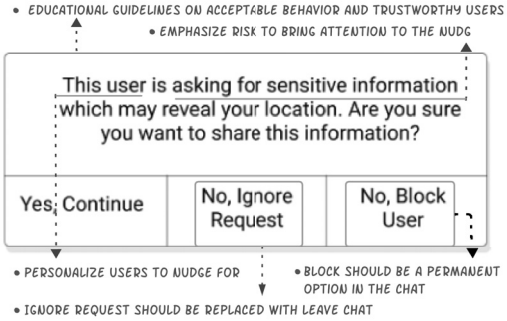


Figure 9: Summary of Feedback on the Information Breach Nudge

risks (50%, n = 10). While some teens wanted to impose censorship for public cyberbullying, in contrast, many of them wanted the option to view the risk within private chat as they believed that there was no risk of public humiliation and they would still be cautious about the risk. On the other hand, some teens wanted control by customizing the option to view the risks (35%, n = 7). For instance, some teens wanted to remove the option to view the risk as they thought that it would make the receiver curious and lead to risk exposure. Whereas, others wanted to vary the view option based on the frequency of risky behavior from other users. For example, if the comment or message is from someone who is routinely reported for harmful behavior, then view should not be presented to the victim. Others wanted granular control over the types of users that they would get nudged for (25%, n = 5), such as nudging for strangers only, as they did not want nudges to interrupt interactions with family or friends.

> *"You could have the nudge only if it's somebody that you don't know. And like somebody random that you don't follow."* - P3 (16-year-old, female)

Moreover, many teens (45%, n = 9) wanted to ensure that the risk would remain hidden from other users, regardless the teen decision to view it, to avoid public humiliation. Some teens (20%, n = 4) recommended a confirmation nudge on clicking view or a limited-timed view after which the message would disappear, acting as a second layer of safety. A few teens (10%, n = 2) also suggested having a comment approval system, where any potentially risky users' comments would require approval from the receiver, as a preemptive measure. Lastly, a few teens wanted to prevent the risk by nudging earlier before the risk escalates (10%, n = 2).

5.2.2 Most teens wanted guidance on what actions to take for risk coping through multiple safety mechanisms. Overall, many teens wanted nudges to provide guidance on coping mechanisms after the risk with multiple safety options. For instance, for both information breaching (60%, n = 12) and predatory risks (65%, n = 13), many teens wanted mechanisms to safely cope with the situation after the risk (Fig. 7, Fig. 8). Many teens liked the option to leave the chat and get nudged earlier to avoid the risk (25%, n = 5). Some (25%, n = 5) preferred to replace the "Ignore Request" option with the option to leave the chat, as they did not find ignoring to help with their safety. Other teens (20%, n = 4) wanted more permanently available ways to block the user, rather than just having it as a pop-up option in the nudge so that they could block at any time. Another idea involved blocking all associated users with a risky user. Similarly, teens wanted to add blocking and reporting as simultaneous safety measures to increase accountability for all risks including information breaches (15%, n = 3), predatory risks (50%, n = 10), cyberbullying (15%, n = 3) and scam (30%, n = 6), as blocking or reporting alone was often insufficient.

> *"I agree with report being a part of the block options, since it helps, not only you, but also others on the platform if you are targeted by the same person."*- P10 (15-year-old, male)

Some teens wanted proactive ways beyond blocking and reporting. For instance, a few participants (15%, n = 3) suggested that nudges should provide automated messages to reject requests for sensitive information. Alternatively, a few teens considered block to be an extreme measure at first and wanted the option to mute the user initially, with the option to block if the risk is repeated (15%, n = 3). A few teens (10%, n = 2) recommended getting parental support for younger teens who may find it harder to respond to online risks. Additionally, many teens wanted to promote community safety through specific educational guidelines early on, regarding acceptable behavior online related to cyberbullying (30%, n = 6), safe information sharing (20%, n = 4), identifying predatory (25%, n = 5) and scam/bots (10%, n = 2). Teens also wanted to emphasize the risk through attention-grabbing visual cues or color coding.

5.2.3 Most teens wanted nudges to enforce accountability. In addition, most teens wanted to enforce accountability through nudges to encourage long-term behavior change among risk perpetrators. For instance, many teens (45%, n = 9) liked the option to let others know about the scam bot risk as it helped increase responsibility. Some teens (30%, n = 6) wanted to selectively let others know about the risk, such as close friends and family, as they did not feel comfortable announcing the risk to everyone in their network. On the other hand, a few teens (15%, n = 3) wanted to let everyone in their network know as a public safety message, after further verification. For instance, a few teens (10%, n = 2) wanted to introduce *public* user reports to discourage scam risks (Fig. 9). Yet, some teens were

concerned that such public humiliation would cause more conflict with the perpetrator:

> *"So I wouldn't have an option for the let others know. I'd rather just screenshot it and send it to people. But if you do it publicly, and if that user sees the post, won't that create more fire?"* - P4 (17-year-old, female)

To overcome this, a few of them wanted to anonymously report users or screenshot the risk to share with trusted users. Likewise, a few teens (15%, n = 3) wanted improved accountability through reporting features as part of the cyberbullying sensitivity filter, such as supporting evidence-based anonymous reporting (15%, n = 3) that led to action against perpetrators. Alternatively, some teens suggested features for penalizing the perpetrator for cyberbullying (30%, n = 6), such as incremental penalties for the risk perpetrators on repeating offenses, leading to a ban (20%, n = 4).

5.2.4 Limitations with Nudging for Autonomous Risk Prevention, Proactive Coping & Community Safety. While teens wanted nudges for autonomous risk prevention, proactive risk coping, and accountability, many teens were cognizant of the limitations of their ideas. For instance, some teens thought that public accountability can have negative effects (35%, n = 7) such as causing more conflict with the perpetrator or increasing curiosity about the risk. Similarly, a few teens (15%, n = 3) were concerned about nudges feeling privacy-invasive and disruptive, especially during private conversations. Also, a few teens discussed the long-term effectiveness of nudges, and suggested that risk detection should evolve over time, especially with realistic bots and Artificial Intelligence approaches that may be able to bypass risk detection. Lastly, several teens recognized that nudges with too little emphasis may be easy to ignore, but those with too many options can feel overwhelming. Therefore, teens recommended simplifying the interface and the language used within nudges. A few teens (20%, n = 4) wanted to improve the choice architecture of the nudges by making the safer options more appealing, such as leading with the safe options or follow-up nudges to split up the number of choices. Additionally, a few teens wanted to have options to delete or view the message in one dedicated place on the nudge, with one choice leading to a single action (e.g., separating "delete and dismiss").

5.3 The Simulation should Mimic Existing Platforms and Balance Realism with Transparency (RQ3)

In this section, we summarize our key findings related to meta-research on how to effectively design research for evaluating adolescent online safety nudges.

5.3.1 Teens wanted transparency and assurance prior to engaging in the research. When asked for feedback regarding the study design, we found that most teens wanted transparency regarding the purpose of the study (60%, n = 12), to ensure that the research met its goals while also prioritizing teens' well-being. For instance, some teens (35%, n = 7) wanted clear and accessible instructions about interacting with other users, for how long, and what the final actions (e.g., friend or block) would mean. Some other concerns included the length of the experiment (a day vs. a week) to ensure timely completion of tasks. Some teens also wanted clarity regarding who would initiate the interactions and whether they should expect to receive messages from other strangers. For these reasons, teens also wanted researchers or additional information about the tasks to be accessible at all times for getting help during participation.

> *Overall having someone who's on stand by like a moderator, to remind you of the rules and kind of explain them whenever you have questions, just even if it's a prompt, just somewhere where you can get assistance"* - P11 (18-year-old, female)

Moreover, teens wanted assurance related to their privacy and safety prior to participation (25%, n = 5). For instance, they wanted researchers to be honest that they were not interacting on a real social media platform and that it was part of a simulated environment to increase comfort with participation. Some teens considered it important to have transparency about data collected and recorded to address any privacy concerns related to their participation. While teens were not asked to provide feedback directly on IRB consent forms, when answering questions about the use of deception in research settings, some leveraged their knowledge of the assent process to suggest adding clear explanations on data protection as part of the informed consent for the future nudge evaluation study. Other teens wanted the researchers to provide such reminders as part of the platform, through privacy and community guidelines that can remind teens to not post any sensitive information as they interact with others. Similarly, a few teens suggested that researchers should minimize the personal information required for the study (e.g., during sign-up and creating their profile/post) and mark the mandatory fields with asterisks. Lastly, a few teens recommended trigger warnings and resources to help those who may be sensitive to certain online risks.

5.3.2 Teens saw the need for deception to increase the realism of study tasks. Several teens weighed the benefits and drawbacks of using deception in the evaluation study (40%, n = 8) in order to increase realism. On one hand, some teens wanted to be informed about the risks presented by the study so that they are not surprised while participating. At the same time, teens acknowledged the need to conceal some information from participants, as disclosing all information would bias their responses to the risk scenarios and nudges. For instance, teens thought that participants could be informed that they will be participating in using a fake social media system with other "real" teen users. Many of them believed that it is important for teens to be deceived about the realness of other social media users on the platform, to ensure that their responses are genuine. P18 explained,

> *"It's definitely important that they don't know that, like their like responses are being tracked, because that will definitely add some bias to it."* - P18 (16-year-old, male)

Some teens thought that participants should know that they are participating in a simulation, but they should not be informed that the other users on the platform are "fake" or "actors". While teens proposed informing participants that they would be interacting with other teens for the purposes of deception, they recognized that this could cause more risks in an uncontrolled environment. Instead, teens liked the idea of other researchers acting as social media users on the platform, for safer interactions. As such, most

teens (60%, n = 12) liked including deception in the study using a Wizard-of-Oz approach where researchers were behind-the-screen and played the role of other social media users.

5.3.3 Teens insisted that the simulated social media platform mirror the ones they actually used. A majority of teens (80%, n = 16) wanted the social media simulation system to be realistic, usable, and mirror other social media platforms they are familiar with. For instance, teens emphasized the importance of imitating features of existing popular social media platforms (35%, n = 7), such as Instagram, by having the same commenting, replying, and sharing post features, as well as similar community guidelines and reporting. Moreover, several teens wanted additional privacy settings on the platform (65%, n = 13), such as the option for making your account private or public depending on your preference. They explained that this is often one of the first social cues they processed about another user that helped them understand their personality (i.e., whether a person is very private or open to public sharing). P14 explained,

> *"Whether or not the account is private or public gives a lot about the person, because if it's public, maybe they're more outgoing - which gives an indicator about their personality"* - P14 (14-year-old, male)

Teens also suggested ways to improve the interface for seamless private chat interaction between the researcher and participant. For instance, some teens (25%, n = 5) suggested ways to make user discovery simpler on chat by having one tab for all chat messages, with the recent messages on top, and the ability to search for users, similar to the Instagram direct messages. In addition, some teens (40%, n = 8) liked the option for a researcher interface that would allow researchers to seamlessly switch between different user personas, whereas a few others thought that such an interface may not be necessary if researchers can login from different browsers.

5.3.4 Teens balanced the need for gaining research insights with the awkwardness of being observed. When asked about the research design, most teens thought that the study should balance gaining insightful research findings with natural responses during the research (85%, n = 17). For instance, many teens (85%, n = 17) weighed the benefits and drawbacks of using a think-aloud approach for giving feedback in real-time. Some teens considered a "think-aloud" approach while participating in the simulation to be awkward, as it would make them conscious and could potentially bias their responses. Instead, they suggested interviewing teens right after the tasks would be more effective in ensuring natural responses and getting in-depth explanations for their choices. On the other hand, many teens thought that it was essential for the researchers to get responses in-the-moment through a think-aloud approach which may not be possible through a post-interview as participants may be unable to recall the rationale for their choices. P18 elaborated,

> *"I feel like the researchers would gain more insight from doing it on zoom, but that would kind of alter the results a bit, because they [participants] probably feel like they're being observed"* - P18 (16-year-old, male)

Other feedback regarding the research design was related to the types of online risks covered in the experiment, for which a few teens (20%, n = 4) recommended presenting a variety of risk scenarios that range in the nature of the risk as well as the severity of the risk, to get insights on the effects of nudges on different types of online risks. Yet, teens realized that as part of the research, it may not be possible to put teens at higher risk (e.g. sharing explicit content). To overcome this, one suggestion was to conceal the risky content through meta-data, without exposing teens to the explicit material to balance research insights with teen well-being.

> *"There could be word-based symbolism where he sent an image and then instead of an actually graphic image, it was just some textual description of what it's actually supposed to be"* - P1 (18-year-old, male)

Overall, teens assessed the benefits and drawbacks of the methodological choices and wanted to strike a balance between unbiased, natural responses and insightful evaluations for nudges.

5.3.5 Teens found it valuable to study actual behavior. Overall, all teens (100%, n = 20) liked the idea of a simulated social media environment and considered this to be the most viable approach for evaluating online safety nudges. Yet, many teens (85%, n = 17) were concerned that the effectiveness of nudges should be determined by measuring actual behavior and that future implementation was based on successful behavior change. For instance, several teens came up with the idea of evaluating nudges in a realistic setting, where different interventions could be compared to assess what works best. Teens recommended several ways to measure behavior change, such as tracking behavioral patterns, app usage, frequency of using a nudge, or A/B testing where different versions of a feature can be compared, similar to what they had seen on social media platforms. P13 explained,

> *"I've seen social media platforms like implement certain things to see how it works and then take them away later. So that's a good way to gauge what works and what does not."* - P13 (16-year-old, female)

At the same time, many of them understood that researchers may not have access to test nudges within large-scale social media platforms. Therefore, teens recommended ways to mimic that experience through high-fidelity prototypes or simulated social media environments. Moreover, some teens (25%, n = 5) wanted researchers to assess how the success of a nudge would be determined and how teens would indicate that they actually felt safe because of a nudge. They wanted additional clarity around what the final actions of safety mean for each user, such as whether blocking would equate to treating a user as unsafe, or friending would mean that a user is safe. A few teens pointed out that safety is a subjective concept and understanding the nuances of each participant's perceived safety is important to determine how effective a nudge was for their safety.

6 DISCUSSION

In this section, we discuss the implications of our findings and provide recommendations for evaluations of adolescent online safety interventions and nudge designs.

6.1 Shades of Grey: Risky People and Situations are Not Black and White

Overall, our research indicates a nuanced challenge: adolescents prefer risk scenarios that are more severe to consider them as credible

or realistic (e.g., involving intense cyberbullying like body shaming or offensive scam content). This presents an ethical conundrum; to gain insights into effective interventions for severe online risks, researchers may need to expose young individuals to scenarios that are both realistic and potentially harmful. To prioritize teens' well-being in such high-risk, high-reward research [15], we need to find creative ways to compromise by simulating higher risks in safe and ethical ways. Instead of directly exposing teens to explicit content, one way of simulating realistic and higher-level risks is by concealing the risk through meta-data [7] or links that block explicit content when clicked. Alternatively, since we cannot directly cyberbully or body-shame participants, they could be exposed to second-hand risks such as using a cyberbully persona who body-shames someone else. Our findings underscore the nuanced nature of online risks, which often manifest subtly through context, trust, and shared interests. This suggests that effective study designs should incorporate personalized, subtle risks, drawing on teens' profiles and shared content to establish rapport before introducing risk elements. Longitudinal approaches are essential, as these nuanced risks tend to escalate over time and cannot be adequately captured through cross-sectional analysis [42].

Correspondingly, our work provided a deeper understanding of how teens envision online risk perpetrators. Most teens believed that perpetrators should match the characteristics of the risks while leveraging deceptive tactics to trick the teen. For instance, teens designed predators who build trust with the teen and would not have a busy career as teens imagined predators to have enough time to spend on social media. While some of these traits align with the different stages of cybergrooming defined by prior literature (e.g., trust-building, friendliness) [38, 55], at times teens' assumptions about predators were too stereotypical (e.g., they have a lot of free time). For example, efforts from Perverted-Justice [46], an anti-predator organization online, show that predators can come from diverse backgrounds and have varying occupations, conflicting with teens understanding of predators. As such, some teens may find it harder to recognize concealed or unusual traits of perpetrators. Our findings depict the need for nudges to help teens decode such risky users, through educational guidelines that could help teens dispel common myths about dangerous users. Our findings inform that as the SIP framework suggests, the social cues and characteristics of a perpetrator are key in determining whether an interaction will evolve into a risk and nudges can help identify these traits early on. Lastly, teens called us out on how risk was treated as binary in our personas. Instead, they challenged the traditional conceptualization of online risks by pointing out that risky people are not always unsafe in their interactions. Therefore, the real risk that requires interventions lies in subtlety, where it is unclear whether someone should be trusted, as perpetrators often mask themselves as relatable and trustworthy. Therefore, teens perceived online risks beyond black-and-white categories of risky or safe behaviors, and were well-versed in thinking about the grey areas of online risk in mature ways. As such, using personas as design probes and the analytical lens of the SIP framework allowed us to look beyond online risks as fixed categories and helped us delve deeper into contextual characteristics of perpetrators and social cues that make online risks nuanced and tricky for teens to recognize. We call future researchers to further build upon the SIP framework to understand how implemented risk scenarios and nudges help teens decode the grey areas of online risk early and effectively.

6.2 Giving Back Control: Tailored Nudges for Autonomous Risk Prevention and Community Safety

A key finding of our work is that teens want proactive risk prevention through sensitivity filters which challenges the narrative that teens are risk-seeking [52]. In fact, teens in our study wanted to shield themselves from online risks in more than one way, through multiple safety mechanisms, as they thought that blocking alone was insufficient in defending them from perpetrators. Across these different nudges, a fundamental difference in our findings is that teens extended these ideas towards personalization of nudges with an emphasis on full control of their decisions. While prior work [4, 51] and social media platforms like Instagram and Twitter include automated sensitivity filters for explicit content [47], these nudges mostly provide generalized warnings regarding violations of community guidelines, with a fixed set of options for safety. Yet, the teens in our study wanted personalization and control at different stages of nudges; i.e., the specific types of risks/triggers of nudges, users they want to be nudged for, and personalized options for safety. The SIP framework allowed us to understand teens' goals of risk prevention and decision-making autonomy within nudges, which aligns with concepts from developmental psychology on how teens seek independence [32]. For instance, teens wanted options to view the risk, and customized keywords and audiences for nudges. One way that social media platforms can grant control and tailor nudges to teens' unique preferences is through online safety questionnaires and granular risk settings when a teen signs up on an app. In this regard, our work contributes to a longstanding debate regarding the ethics of nudging. While nudges have been questioned on manipulating a user's freedom [58], our findings provide a teen-centric perspective to reinforce prior research that puts responsibility in the hands of choice architects to design nudges that are autonomy-granting and supplement decision-making, without undermining freedom of choice [49, 76].

Moreover, teens wanted to move beyond generic community guidelines, towards tailored online safety education that helped with identifying tricky online risks early on. Overall, our findings depart from prior interventions that focused on restrictive or parent-centric approaches for adolescent online safety [45, 78]. Instead, the teens in our study build upon nudges as a strength-based approach, that serves as "teachable moments," providing teens with tips for online safety to self-regulate their experiences, at the right moment [40]. Similarly, teens wanted ways to protect one another, such as warning others or public user reports to hold risk perpetrators accountable. While the broader literature in technology accountability largely focuses on algorithmic accountability [1, 39, 86] or accountability from big-tech companies [18, 21] regarding privacy and data use, we found that teens want to shift accountability back to the users to have shared responsibilities to protect themselves from risks. For instance, Abdul et al. [1]'s literature review on technology accountability presents a shift from gaining individual trust in algorithms towards building systematic and social accountability in intelligent systems. Our work parallels this from

a human perspective, by emphasizing human accountability, calling for consequences and collective transparency regarding users' actions online. Our findings most closely align with those from Xiao et al. [91] in which teens wanted long-term transformation of online spaces through accountability and restorative justice. Yet, teens in our study went beyond the current recommendations for community safety [2, 91] to recognize limitations of accountability nudges, such as misuse through false accusations, increased conflicts or challenges with risk detection as people or bots learn to bypass risks detection. In this regard, recent work in online risk detection for youth calls for continued improvement of automated risk detection using approaches that consider the context of the risk, the human perspective, and multiple modes of data [6, 66]. Therefore, our work calls for nudges that provide control to teens and promote good digital citizenship, triggered by accurate risk detection and designed with careful consideration, so they do not unintentionally put teens in harm's way.

6.3 A Social Media Sandbox as a Happy Medium between Experimental Design and 'In the Wild' Research

Our study highlights the complexities of evaluating nudges, as teens identified tensions between research design, realism, and safety. For instance, teens wanted transparency about the research goals while realizing that some deceit is necessary for unbiased responses during the research. Teens also recognized the challenge that thinking aloud during the experiment would be awkward while acknowledging that it may help researchers obtain in-depth insights. Prior work presents this dilemma as a three-horn problem, i.e., it is challenging to attain all three aspects in experimental research, namely, experimental realism, precision, and generalizability [53]. Our work demonstrates that these methodological challenges are further amplified when working with vulnerable populations.

Using the analytical lens of SIP, we uncovered the different compromises teens envisioned for resolving these tensions for designing approaches to measure online safety outcomes of nudges. For instance, they suggested using a light-weight think-aloud, with follow-up questions to avoid recall bias, supplemented by retrospective post-interviews. Teens also envisioned realistic simulations that mimicked social media features and allowed relational changes over time (e.g., changes in privacy settings). For deception, teens settled on being informed about participating in a simulation, while hiding information about *who* they would be interacting with. Instead of interacting with other teens in an uncontrolled environment, teens liked the idea of measuring actual behavior change using a Wizard-of-Oz approach with researchers. While this approach has been previously used [37, 71], prior work highlights some challenges with Wizard-of-Oz simulations, such as ensuring consistency and overcoming human errors [71]. Our findings imply that ecologically valid approaches to evaluating online safety outcomes of nudges with teens cannot be perfect; by increasing realism through a simulation in a partially controlled environment, we will reduce precision and consistency [5, 53]. Yet, creating a social media sandbox is a suitable middle ground based on teens feedback and a safer approach than a field study where teens can interact with others "in-the-wild" which can escalate into uncontrolled higher risks. A similar social media sandbox was employed by DiFranzo et al. [26] for online safety education, as they used a social media environment with educator-facilitated classroom lessons to teach youth how to be safe online. Our research calls for developing upon their efforts to create open-source research tools that allow integration of online safety interventions and can serve as an interactive playground for online safety experimentation.

6.4 Design Implications for Nudges

Based on our findings, we provide design guidelines for future researchers who aim to use strength-based nudges for promoting adolescent online safety:

- Design for ***personalized proactive risk prevention***, through sensitivity filters and granular settings to specify nudge preferences, so that prevention happens according to the users (e.g., strangers only) and risks (e.g., explicit content only) that teens are sensitive to.
- Provide more than one path to safety, with ***multiple options as coping mechanisms*** for the risk, through a combination of traditional approaches (e.g., blocking, reporting), as well as guided approaches (e.g., automated response suggestions, risk filter, comment approval system).
- Leave the ultimate decision-making control in the hands of teens. For instance, nudges should ***provide control to override sensitivity filters*** to view risks, dismiss a nudge, or turn off nudges.
- Provide explicit ways to ***support disengaging from the risk***. While this action may seem obvious, teens preferred explicit reminders to disengage, ignore, or leave the conversation without consequences.
- ***Educate beyond generic community guidelines***, to assist teens with tips on identifying social cues leading to risk.
- Provide ways to hold the perpetrators accountable for their harmful actions (e.g., notifying select others about the risk). Yet, future work is needed to find ways that ***balances accountability while protecting the victim*** from further conflict or backlash.
- Maintain a ***simple choice architecture*** by using follow-up nudges and confirmation prompts as ways to present multiple options in a minimized way.

6.5 Heuristic Guidelines for Evaluating Technology Interventions and Protecting Vulnerable Users

We summarize our findings to develop the following heuristic guidelines for evaluating technology-based interventions with vulnerable users for promoting their online safety:

- Provide ***transparency about the larger research goals*** while concealing specific details about the experiment that would bias the participant responses (e.g., fake interactions, exposure to nudges).
- Consider ***simulating higher risks in public spaces*** and targeted towards another user (e.g., body shaming), rather

than to the participant themselves so that the risk is visible but not harmful to the vulnerable participant.
- ***Assure participants about data protection*** during the informed consent process to increase comfort with participation and genuine interactions.
- Consider using a ***lightweight think aloud*** with a retrospective post-interview to balance in-depth insights captured at the moment, with minimal observation and interruption that would bias the participant responses.
- Consider using a realistic ***simulated environment with semi-controlled interactions*** to strike a meaningful compromise between experimental realism and precision [53].

6.6 Limitations and Future Research

While we provide several actionable recommendations for designing and evaluating adolescent online safety interventions with teens, we recognize the limitations of our study. We had a racially diverse group of teen participants; yet, our sample size was small, therefore, diverse opinions from underrepresented groups of teens may not be reflected in our insights. As such, our findings may not be generalizable to all youth populations, such as those from low socio-economic backgrounds or non-Western contexts. Moreover, since we worked with teens using focus groups, their ideas may be subject to social desirability bias or groupthink. Additionally, a challenge in our study was that sometimes the feedback from teens conflicted, hence, we had to do our best to resolve disagreements. For instance, while many teens thought that Bryan (the information breaching persona) should portray a naive, socially awkward personality, the feedback from an autistic teen was poignant enough to prioritize redesigning this persona in the future for implementing an information breaching persona that is extroverted and inquisitive in nature, based on feedback from the teens. Finally, while our work contributes towards *how* to evaluate adolescent online safety nudges, we encourage future researchers to leverage our meta-level insights to implement these guidelines and conduct empirical research to validate them.

7 CONCLUSION

Our work is the first-of-its-kind to involve teens as co-designers of research for evaluating adolescent online safety interventions, realistically and safely. Teens taught us that online risks are not black and white, as the real risk often lies in nuanced grey areas. Overall, for evaluating nudges to address these nuanced situations, teens carefully weighed the pros and cons of different approaches and reached a middle ground of using a simulated social media environment, with some deception, light-weight think-aloud, and assurances regarding data privacy. Our work contributes to the larger HCI community, by providing design and heuristic guidelines for conducting ecologically valid evaluations for technology-based interventions with vulnerable populations. We call future researchers to rely on our work as a stepping stone to move beyond designing interventions towards such high-risk, high-reward evaluations that can advance adolescent online safety.

ACKNOWLEDGMENTS

This research was supported by the William T. Grant #187941 and National Science Foundation Award IIS-2333207. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our sponsors.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

[2] 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. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.

[3] Zainab Agha, Kelsey Miu, Sophia Piper, Jinkyung Park, and Pamela J Wisniewski. 2023. Co-designing user personas and risk scenarios for evaluating adolescent online safety interventions. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 249–253.

[4] Zainab Agha, Zinan Zhang, Oluwatomisin Obajemu, Luke Shirley, and Pamela J. Wisniewski. 2022. A Case Study on User Experience Bootcamps with Teens to Co-Design Real-Time Online Safety Interventions. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–8.

[5] Zhila Aghajari, Eric PS Baumer, Jess Hohenstein, Malte F Jung, and Dominic DiFranzo. 2023. Methodological Middle Spaces: Addressing the Need for Methodological Innovation to Achieve Simultaneous Realism, Control, and Scalability in Experimental Studies of AI-Mediated Communication. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–28.

[6] Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Joshua Gracie, Munmun De Choudhury, Pamela J Wisniewski, and Gianluca Stringhini. 2022. Understanding the digital lives of youth: Analyzing media shared within safe versus unsafe private conversations on Instagram. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.

[7] Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Chen Ling, Munmun De Choudhury, Pamela J Wisniewski, and Gianluca Stringhini. 2023. Getting Meta: A Multimodal Approach for Detecting Unsafe Conversations within Instagram Direct Messages of Youth. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–30.

[8] Ashwaq Alsoubai, Jihye Song, Afsaneh Razi, Nurun Naher, Munmun De Choudhury, and Pamela J Wisniewski. 2022. From'Friends with Benefits' to'Sextortion:'A Nuanced Investigation of Adolescents' Online Sexual Risk Experiences. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–32.

[9] Alissa Nicole Antle. 2006. Child-personas: fact or fiction?. In *Proceedings of the 6th conference on Designing Interactive systems*. 22–30.

[10] Jeffrey Jensen Arnett. 2000. Emerging adulthood: A theory of development from the late teens through the twenties. *American psychologist* 55, 5 (2000), 469.

[11] 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*. 3895–3905.

[12] Erinn Atwater, Cecylia Bocovich, Urs Hengartner, and Ian Goldberg. [n. d.]. Netsim: Network simulation and hacking for high schoolers. ([n. d.]).

[13] Erinn Atwater, Cecylia Bocovich, Urs Hengartner, and Ian Goldberg. 2017. Live Lesson: Netsim: Network simulation and hacking for high schoolers. In *2017 USENIX Workshop on Advances in Security Education (ASE 17)*.

[14] Minja Axelsson, Raquel Oliveira, Mattia Racca, and Ville Kyrki. 2021. Social robot co-design canvases: A participatory design framework. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 1 (2021), 1–39.

[15] Karla Badillo-Urquiola, Zachary Shea, Zainab Agha, Irina Lediaeva, and Pamela Wisniewski. 2021. Conducting risky research with teens: co-designing for the ethical treatment and protection of adolescents. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–46.

[16] Karla 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 *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. 394–406.

[17] Naulsberry Jean Baptiste, Jinkyung Park, Neeraj Chatlani, Naima Samreen Ali, and Pamela J Wisniewski. 2023. Teens on Tech: Using an Asynchronous Remote Community to Explore Adolescents' Online Safety Perspectives. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 45–49.

[18] Kean Birch, DT Cochrane, and Callum Ward. 2021. Data as asset? The measurement, governance, and valuation of digital personal data by Big Tech. *Big Data & Society* 8, 1 (2021), 20539517211017308.

[19] Leanne Bowler, Eleanor Mattern, and Cory Knobel. 2014. Developing design interventions for cyberbullying: A narrative-based participatory approach. *iConference 2014 Proceedings* (2014).

[20] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[21] Ben Clements. 2022. The Big Tech Accountability Act: Reforming How the Biggest Corporations Control and Exploit Online Communications. *W. New Eng. L. Rev.* 44 (2022), 5.

[22] A Cooper. [n. d.]. The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How To Restore the Sanity,(1999). *Indianapolis: SAMS* ([n. d.]).

[23] Arianna J Davis. 2020. Co-Designing" Teenovate": An Intergenerational Online Safety Design Team. (2020).

[24] Brittany Cook Davis. 2012. *Social information-processing deficits in adolescent sex offenders.* Alliant International University.

[25] Katie Davis, Petr Slovak, Rotem Landesman, Caroline Pitt, Abdullatif Ghajar, Jessica Lee Schleider, Saba Kawas, Andrea Guadalupe Perez Portillo, and Nicole S Kuhn. 2023. Supporting Teens' Intentional Social Media Use Through Interaction Design: An exploratory proof-of-concept study. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 322–334.

[26] Dominic DiFranzo, Yoon Hyung Choi, Amanda Purington, Jessie G Taft, Janis Whitlock, and Natalya N Bazarova. 2019. Social media testdrive: Real-world social media education for the next generation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–11.

[27] Dominic DiFranzo, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D Wherry, and Natalya N Bazarova. 2018. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[28] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 3 (2012), 1–30.

[29] Kenneth A Dodge and Nicki R Crick. 1990. Social information-processing bases of aggressive behavior in children. *Personality and social psychology bulletin* 16, 1 (1990), 8–22.

[30] Shamal Faily and Ivan Flechais. 2011. Persona cases: a technique for grounding personas. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2267–2270.

[31] Figma. 2023. FigJam Turn possibilities into plans. https://www.figma.com/figjam/

[32] Manuela Fleming. 2005. Adolescent Autonomy: Desire, Achievement and Disobeying Parents between Early and Late Adolescence. *Australian Journal of Educational & Developmental Psychology* 5 (2005), 1–16.

[33] Diana Freed, Natalie N Bazarova, Sunny Consolvo, Eunice J Han, Patrick Gage Kelley, Kurt Thomas, and Dan Cosley. 2023. Understanding Digital-Safety Experiences of Youth in the US. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[34] Darren Gergle and Desney S Tan. 2014. Experimental research in HCI. In *Ways of Knowing in HCI*. Springer, 191–227.

[35] 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*. 1–14.

[36] Isha Ghosh and Vivek Singh. 2022. "Not all my friends are friends": Audience-group-based nudges for managing location privacy. *Journal of the Association for Information Science and Technology* 73, 6 (2022), 797–810.

[37] Thomas Grill and Manfred Tscheligi. 2013. The ConWIZ protocol: a generic protocol for wizard of Oz simulations. In *International Conference on Computer Aided Systems Theory*. Springer, 434–441.

[38] Aditi Gupta, Ponnurangam Kumaraguru, and Ashish Sureka. 2012. Characterizing pedophile conversations on the internet using online grooming. *arXiv preprint arXiv:1208.4324* (2012).

[39] Samantha Hautea, Sayamindu Dasgupta, and Benjamin Mako Hill. 2017. Youth perspectives on critical data literacies. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 919–930.

[40] Robert J Havighurst. 1953. Human development and education. (1953).

[41] Niels Hendriks, Frederik Truyen, and Erik Duval. 2013. Designing with dementia: Guidelines for participatory design together with persons with dementia. In *Human-Computer Interaction–INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part I 14*. Springer, 649–666.

[42] Jina Huh-Yoo, Afsaneh Razi, Diep N Nguyen, Sampada Regmi, and Pamela J Wisniewski. 2023. "Help Me:" Examining Youth's Private Pleas for Support and the Responses Received from Peers via Instagram Direct Messages. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[43] John PA Ioannidis, Daniele Fanelli, Debbie Drake Dunne, and Steven N Goodman. 2015. Meta-research: evaluation and improvement of research methods and practices. *PLoS biology* 13, 10 (2015), e1002264.

[44] Carrie James, Emily Weinstein, and Kelly Mendoza. 2019. Teaching digital citizens in today's world: Research and insights behind the Common Sense K–12 Digital Citizenship Curriculum. *Common Sense Media* (2019).

[45] MAP Jayawardena, MHFM Mahadi Hassan, MIA Aflal, WAAS Weerathunga, SMB Harshanath, and UU Samantha Rajapaksha. 2022. Monitoring System for Underage Smart Phone Users. In *2022 4th International Conference on Advancements in Computing (ICAC)*. IEEE, 228–233.

[46] Perverted Justice. 2023. Perverted Justice. http://www.perverted-justice.com/

[47] Shawn Knight. 2017. Instagram is adding a "sensitive content" filter, enables two-factor authentication for all. https://www.techspot.com/news/68636-instagram-adding-sensitive-content-filter-enables-two-factor.html

[48] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. 2010. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)* 10, 2 (2010), 1–31.

[49] Tim-Benjamin Lembcke, Nils Engelbrecht, Alfred Benedikt Brendel, and Lutz M Kolbe. 2019. To Nudge or not to Nudge: Ethical Considerations of Digital nudging based on its Behavioral Economics roots.. In *ECIS*.

[50] Sonia Livingstone. 2008. Taking risky opportunities in youthful content creation: teenagers' use of social networking sites for intimacy, privacy and self-expression. *New media & society* 10, 3 (2008), 393–411.

[51] 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*. 1–11.

[52] Cheryl McCarty, Aimee D Prawitz, Linda E Derscheid, and Bette Montgomery. 2011. Perceived safety and teen risk taking in online chat sites. *Cyberpsychology, Behavior, and Social Networking* 14, 3 (2011), 169–174.

[53] Joseph E McGrath. 1981. Dilemmatics: The study of research choices and dilemmas. *American Behavioral Scientist* 25, 2 (1981), 179–210.

[54] 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*. 1–9.

[55] Miljana Mladenović, Vera Ošmjanski, and Staša Vujičić Stanković. 2021. Cyber-aggression, cyberbullying, and cyber-grooming: a survey and research challenges. *ACM Computing Surveys (CSUR)* 54, 1 (2021), 1–42.

[56] Kathryn L. Modecki, Rachel E. Goldberg, Pamela Wisniewski, and Amy Orben. 2022. What Is Digital Parenting? A Systematic Review of Past Measurement and Blueprint for the Future. *Perspectives on Psychological Science* 17, 6 (2022), 1673–1691.

[57] Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. 2014. Survey research in HCI. *Ways of Knowing in HCI* (2014), 229–266.

[58] Robert Noggle. 2018. Manipulation, salience, and nudges. *Bioethics* 32, 3 (2018), 164–170.

[59] Judith S Olson and Wendy A Kellogg. 2014. *Ways of Knowing in HCI*. Vol. 2. Springer.

[60] Lisa H Papatraianou, Diane Levine, and Dean West. 2014. Resilience in the face of cyberbullying: An ecological perspective on young people's experiences of online adversity. *Pastoral Care in Education* 32, 4 (2014), 264–283.

[61] Jinkyung Park, Joshua Gracie, Ashwaq Alsoubai, Gianluca Stringhini, Vivek Singh, and Pamela Wisniewski. 2023. Towards Automated Detection of Risky Images Shared by Youth on Social Media. In *Companion Proceedings of the ACM Web Conference 2023*. 1348–1357.

[62] Jinkyung Park, Irina Lediaeva, Amy Godfrey, Maria Lopez, Kapil Chalil Madathil, Heidi Zinzow, and Pamela Wisniewski. 2023. How Affordances and Social Norms Shape the Discussion of Harmful Social Media Challenges on Reddit. *Human Factors in Healthcare* (2023), 100042.

[63] Anthony T Pinter, Pamela J Wisniewski, Heng Xu, Mary Beth Rosson, and Jack M Caroll. 2017. Adolescent online safety: Moving beyond formative evaluations to designing solutions for the future. In *Proceedings of the 2017 Conference on Interaction Design and Children*. 352–357.

[64] Erika S Poole and Tamara Peyton. 2013. Interaction design research with adolescents: methodological challenges and best practices. In *Proceedings of the 12th International Conference on Interaction Design and Children*. 211–217.

[65] Rebecca M Quintana, Stephanie R Haley, Adam Levick, Caitlin Holman, Ben Hayward, and Mike Wojan. 2017. The persona party: Using personas to design for learning at scale. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 933–941.

[66] Afsaneh Razi, Ashwaq AlSoubai, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J Wisniewski. 2023. Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–29.

[67] 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*. 1–13.

[68] Tara L Rutkowski, Heidi Hartikainen, Kirsten E Richards, and Pamela J Wisniewski. 2021. Family Communication: Examining the Differing Perceptions of Parents and Teens Regarding Online Safety Communication. *Proceedings of the*

ACM on Human-Computer Interaction 5, CSCW2 (2021), 1–23.
[69] Florian Schaub, Rebecca Balebako, Adam L Durity, and Lorrie Faith Cranor. 2015. A design space for effective privacy notices. In *Eleventh symposium on usable privacy and security (SOUPS 2015)*. 1–17.
[70] Diane J Schiano, Christine Burg, Anthony Nalan Smith, and Florencia Moore. 2016. Parenting Digital Youth: How Now?. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 3181–3189.
[71] Stephan Schlögl, Gavin Doherty, and Saturnino Luz. 2015. Wizard of Oz experimentation for language technology applications: Challenges and tools. *Interacting with Computers* 27, 6 (2015), 592–615.
[72] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.
[73] Ben Shneiderman. 2016. *The new ABCs of research: Achieving breakthrough collaborations*. Oxford University Press.
[74] Kenneth Stanley. 2007. Design of randomized controlled trials. *Circulation* 115, 9 (2007), 1164–1169.
[75] Peter Story, Daniel Smullen, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. 2020. From intent to action: Nudging users towards secure mobile payments. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. 379–415.
[76] Cass R Sunstein. 2015. Nudging and choice architecture: Ethical considerations. *Yale Journal on Regulation, Forthcoming* (2015).
[77] Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
[78] Lee Jia Thun, Phoey Lee Teh, and Chi-Bin Cheng. 2022. CyberAid: Are your children safe from cyberbullying? *Journal of King Saud University-Computer and Information Sciences* 34, 7 (2022), 4099–4108.
[79] Kathleen Van Royen, Karolien Poels, Heidi Vandebosch, and Philippe Adam. 2017. "Thinking before posting?" Reducing cyber harassment on social networking sites through a reflective message. *Computers in human behavior* 66 (2017), 345–352.
[80] Emily A Vogels. 2022. Teens and Cyberbullying 2022. (2022).
[81] Ashley Marie Walker, Yaxing Yao, Christine Geeng, Roberto Hoyle, and Pamela Wisniewski. 2019. Moving beyond'one size fits all' research considerations for working with vulnerable populations. *Interactions* 26, 6 (2019), 34–39.
[82] Jayne Wallace, John McCarthy, Peter C Wright, and Patrick Olivier. 2013. Making design probes work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3441–3450.
[83] Joseph B Walther. 1992. Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication research* 19, 1 (1992), 52–90.
[84] Joseph B Walther. 2008. Social information processing theory. *Engaging theories in interpersonal communication: Multiple perspectives* 391 (2008).
[85] Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Faith Cranor. 2013. Privacy nudges for social media: an exploratory Facebook study. In *Proceedings of the 22nd international conference on world wide web*. 763–770.
[86] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 1–18.
[87] 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*. 51–69.
[88] 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*. 4029–4038.
[89] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, Daniel F Perkins, and John M Carroll. 2016. Dear diary: Teens reflect on their weekly online risk experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3919–3930.
[90] Pai-Lu Wu and Wen-Bin Chiou. 2009. More options lead to more searching and worse choices in finding partners for romantic relationships online: An experimental study. *CyberPsychology & Behavior* 12, 3 (2009), 315–318.
[91] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. 2022. Sensemaking, support, safety, retribution, transformation: A restorative justice approach to understanding adolescents' needs for addressing online harm. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
[92] Maximilian Zinkus, Oliver Curry, Marina Moore, Zachary Peterson, and Zoë J Wood. 2019. Fakesbook: A social networking platform for teaching security and privacy concepts to secondary school students. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. 892–898.

A PARTICIPANTS DEMOGRAPHIC INFORMATION

Table 2: Participants' Demographic Information

Group	ID	Age	Gender	Ethnicity	Favorite Social Media Platform	State
Group 1	1	18	M	White/Caucasian	TikTok, Youtube	Florida
	2	17	M	Hispanic/Latino	Youtube	Florida
Group 2	3	16	F	Asian	Instagram	Florida
	4	17	F	Black/African American	Instagram	Florida
	5	18	F	Asian	Instagram	Texas
Group 3	6	17	M	Asian	Instagram	Florida
	7	16	M	Hispanic/Latino	Instagram	Florida
Group 4	8	16	M	Asian	Youtube	Massachusetts
	9	13	F	Asian	Instagram	Massachusetts
Group 5	10	15	M	Hispanic/Latino	Instagram, Youtube	Tennessee
	11	18	F	Black/African American	TikTok	New York
Group 6	12	14	F	Asian	Instagram	California
	13	16	F	Asian	Instagram	New York
	14	14	M	Asian	Snapchat	New Jersey
Group 7	15	14	M	Black/African American	TikTok	New Jersey
	16	17	F	Asian	TikTok, Instagram	Florida
	17	16	M	Black/African American	TikTok, Instagram	New York
Group 8	18	16	M	Asian	Instagram	Virginia
	19	17	M	Asian	Instagram	Texas
Group 9	20	14	M	Black/African American	Instagram	Massachusetts

B QUALITATIVE CODEBOOK SUMMARIZING KEY FINDINGS

Table 3: Final Codebook and Themes mapped to SIP Framework

Dimension	SIP Framework	Theme	Codes
The Risky Situations and People Teens Expect to Encounter When Using Social Media (RQ1)	**Social cues** for decoding online risks	**Risk scenarios need to be realistically risky to be believable** (65%, n = 13)	• Bullying is harsh and often about appearance (40%) • Predators send higher risk offensive content (20%) • Information breaching is specific and sensitive (15%) • Bots often reshare targeted spam (20%)
		Online risks should be tricky and subtle, not obvious (90%, n = 18)	• Bot risks are incentized (40%) and realistic (55%) • Predatory risks involve trust (45%) and incentives (20%) • Information breaching is indirect based on context (35%) • Bullying is often back-handed (30%)
	Contextual factors for risky users	**Characteristics of the perpetrators should match the risk** (90%, n = 18)	• Predators have free time (40%), hide their age (25%) and are narcissictic (15%) • Bullies are influential (25%) and protect their status (35%) • Information breachers act naive (25%) or are extroverted (25%)
How Teens Redesigned Nudge-based Interventions for Their Online Safety (RQ2)	How **teens goals** inform nudges	**Teens wanted proactive risk prevention with autonomy to override decisions** (100%, n = 20)	• Sensitivity filters for cyberbullying (70%) & scams (50%) • Personalize sensitivity filters for risk preferences (50%) • Autonomous customized options to view the risk (35%) • Hide risks from other users to avoid embarrassment (45%)
	Decision-making choices of nudges	**Teens wanted guidance on risk coping through multiple safety options** (65%, n = 13)	• Multiple safety choices (e.g., block, report or leave) (65%) • New safety choices (e.g., automated responses) (15%), comment approval system (15%) or parent support (10%) • Educate on specific ways to identify and avoid risks (35%)
		Teens wanted nudges to enforce accountability (45%, n = 9)	• Selectively inform others about the risk (30%) • Incremental penalties for risk perpetrators (20%) • Public risk messages (15%) and evidenced reports (15%)
Teen Considerations for Realistically and Safely Evaluating Online Safety Nudges (RQ3)	Study design choices for **online safety outcomes**	**Teens wanted transparency and assurance prior to engaging in the research** (60%, n = 12)	• Provide transparency & clarity about research goals (35%) • Assure data privacy and safety (25%)
		Teens saw the need for deception to increase realism of the study (60%, n = 12)	• Understood the need for deception about fake interactions (40%) • Liked the idea of researchers as "actors" using a Wizard-of-Oz approach (60%)
		Teens balanced the need for gaining research insights with awkwardness of being observed (85%, n = 17)	• Understood the benefits and awkwardness of think aloud (85%) • Recommended alternative solutions (e.g., interview after tasks) (45%) • Suggested simulating various risk types/severity (20%)
	Features that allow **relational changes** over time	**Teens insisted that the simulated social media platform mirror the ones they actually used** (80%, n = 16)	• Provide additional privacy settings (e.g., public/ private profile) (65%) • Mirror features from Instagram (e.g., comment reply, share post) (35%) • Improve user discovery on chat (25%)