Towards Enforcing Good Digital Citizenship: Identifying Opportunities for Adolescent Online Safety Nudges

OLUWATOMISIN OBAJEMU*, University of Florida, USA
ZAINAB AGHA*, Vanderbilt University, USA
FARZANA A. CHOWDHURY, Georgia Institute of Technology, USA
PAMELA J. WISNIEWSKI, Vanderbilt University, USA

With the prevalence of risks encountered by youth online, strength-based approaches such as nudges have been recommended as potential solutions to guide teens toward safer decisions. However, most nudging interventions to date have not been designed to cater to teens' unique needs and online safety concerns. To address this gap, this study provided a comprehensive view of adolescents' feedback on online safety nudges to inform the design of more effective online safety interventions. We conducted 12 semi-structured interviews and 3 focus group sessions with 21 teens (13 – 17 years old) via Zoom to get their feedback on three types of nudge designs from two opposing perspectives (i.e., risk victim and perpetrator) and for two different online risks (i.e., Information Breaches and Cyberbullying). Based on the teens' responses, they expressed a desire that nudges need to move beyond solely warning the user to providing a clear and effective action to take in response to the risk. They also identified key differences that affect the perception of nudges in effectively addressing an online risk, they include age, risk medium, risk awareness, and perceived risk severity. Finally, the teens identified several challenges with nudges such as them being easy to ignore, disruptive, untimely, and possibly escalating the risk. To address these, teens recommended clearer and contextualized warnings, risk prevention, and nudge personalization as solutions to ensure effective nudging. Overall, we recommend online safety nudges be designed for victim guidance while providing autonomy to control their experiences, and to ensure accountability and prevention of risk perpetrators to restrict them from causing harm.

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

Additional Key Words and Phrases: Adolescent online safety, Online risks, Nudges, Social media

ACM Reference Format:
Oluwatomisin Obajemu, Zainab Agha, Farzana A. Chowdhury, and Pamela J. Wisniewski. 2024. Towards Enforcing Good Digital Citizenship: Identifying Opportunities for Adolescent Online Safety Nudges. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 136 (April 2024), 37 pages. https://doi.org/10.1145/3637413

1 INTRODUCTION

Social media use comes with many benefits for all age groups, especially adolescents and teens for entertainment, growth, and connecting with people [50]. However, many teens experience online risks such as information breaches, sexual solicitations, cyberbullying, and harassment on social media [2, 65], and there are numerous studies and statistics that highlight this problem. For

*Both authors contributed equally to this research.

Authors' addresses: Oluwatomisin Obajemu, University of Florida, Gainesville, Florida, USA; Zainab Agha, Vanderbilt University, Nashville, Tennessee, USA; Farzana A. Chowdhury, Georgia Institute of Technology, Atlanta, Georgia, USA; Pamela J. Wisniewski, Vanderbilt University, Nashville, Tennessee, USA.

2573-0142/2024/4-ART136 $15.00
https://doi.org/10.1145/3637413

example, 59.9% of parents reported their children (ages 14-18) have been victims of bullying on the internet [49]. Similarly, a Pew Research study shows how a majority of teens (62%) desire to have a welcoming and safe online environment [37]. Recent research suggests that teens prefer to handle their online risks with strength-based approaches that allow them to manage their online safety issues independently [6, 63] as opposed to the traditional method of involving a third party (i.e., parent, guardian, or respected adult), which is often considered restrictive and privacy-invasive [61, 64]. Young users have also shown an affinity for "just-in-time" interventions that empower and help them (in that exact moment) when a social media risk occurs in comparison to safety features that can restrict or limit their online activity [2]. These interventions can be implemented via nudges, which are a resilient user-centric approach to online safety that augment a user's decision-making without taking away their autonomy [54].

Nudging has been successfully implemented in other fields (e.g., privacy, security, and other disciplines) to change users' behaviors towards positive choices for their privacy online [25, 39]. In the context of online safety, recent research supports that nudges seem to be a way forward in ensuring adolescents' online privacy and safety [2, 35]. For example, Masaki et al. conducted survey-based research with teens to ask how they would respond to different privacy nudges based on nine risky scenarios. They found nudges to be a promising approach in promoting privacy preserving choices for teens and recommended conducting further qualitative examinations of nudges in different contexts [35]. While this work provides relevant groundwork for adolescent online safety nudging, their nudges were based on a synthesis of the literature and not designed in partnership with teens. Currently, very few studies employ participatory design to involve teens in the process of designing online safety nudges, as they are often designed by other stakeholders such as parents or industry professionals [33]. As such, it is important to involve teens directly in co-design to gain a deeper understanding of effective nudges for teens, catering to their unique experiences. Recently, Agha et al. [2] addressed this gap by partnering with adolescents directly to co-design interventions including risk prevention and intelligent assistance to help teens cope with online risks. In this paper, we combine the strengths of the work of Masaki et al. [35] and Agha et al. [3] by having teens evaluate online safety nudges that were conceptualized with teens. Importantly, we extend prior work by moving beyond generative designs and survey-based feedback, to involving teens in an in-depth design critique to refine implementable online safety nudges for future field studies with youth.

We identified major themes across the prior work and developed three nudge types to represent trends within the literature, including those that were conceptualized by youth themselves. For instance, pop-up warnings have been commonly implemented [4, 59], and Masaki et al. [35] found that nudges with general descriptions and negatively framed warnings were preferred by teens. This informed the design of our first nudge; a) a **General Warning**, which provided a warning for the teen via a popup reminding them about the risk. Secondly, we included a **Sensitivity Filter**, which censored the risk along with a warning message. This nudge was based on prior work which recommended automatic censorship of risky content [5, 48], along with control and the ability to view the risk [2]. Lastly, prior work [2, 9] showed that teens want guidance and intelligent assistance during the risk which informed the design of c) a **Guided Actions** nudge, which provided automated actions and responses to the risk. Additionally, recent efforts have called for designing interventions for more than just teens as the victims [57], towards addressing online risks at the root cause by targeting nudges for risk perpetrators [2, 31]. Therefore, we added a novel aspect to this study by implementing all nudges from both the perspectives of a risk victim and a risk perpetrator. All nudges were applied in two risk scenarios based on what was commonly reported by teens [2], including a) an **Information Breach** scenario, where a stranger attempts to convince the teen to disclose their personal information, and b) a **Cyberbullying** scenario, where

a stranger bullies the teen with unwarranted offensive messages. To understand teens' design critique of the three nudges, we ask the following research questions.

RQ1: What nudges are considered most effective by teens for dealing with unsafe interactions online based on a) nudge approach, b) risk type, and c) user perspective? And why?

RQ2: What other key differences emerged that influenced teens' impressions of online safety nudges?

RQ3: What were the key challenges identified by teens when evaluating the nudges, and what were their recommendations to address these challenges?

To answer these research questions, we conducted three focus groups with three teens each, and twelve 1-on-1 interviews via Zoom. During the sessions, the teens were presented with two different online risks - Information Breach and Cyberbullying, where each risk had an implementation of the General Warning, Sensitivity Filter, and Guided Actions nudge approaches. The teens were then prompted with a series of semi-structured questions as a means of collecting their feedback on the nudges from two opposite perspectives (victim and perpetrator). The questions were asked to determine: (a) their impressions and perceptions of the nudges, (b) their suggestions to improve the nudges, (c) their concerns with the nudges, and (d) their overall preferences. We found out that teens preferred nudges that offered an actionable response to the risk (i.e., censoring, guided actions) over solely warning as overall, they ranked Sensitivity Filter and Guided Actions over General Warnings (RQ1a). When analyzing preferred nudges for each risk type, we found that teens preferred Guided Actions for Information Breaches, and they considered Sensitivity Filter to be more suitable for the Cyberbullying risk, while General Warning was ranked last for both risk types (RQ1b). When considering the perspective of the user (victim vs perpetrator), the teens preferred the Guided Actions Nudge for the risk victim while ranking the Sensitivity Filter Nudge best for the risk perpetrator (RQ1c). The teens also highlighted some key differences that may affect how these nudges may be perceived (RQ2), they include the risk medium, perceived risk severity, users' awareness of risk, and teens' age. Finally, they identified some challenges with online safety nudges such as being easy to ignore, inadvertently worsening the risk, or not being optimized enough to the risk and user. As solutions, teens recommended more convincing risk warnings, risk prevention measures, personalization of nudges, and the provision of safe actions to the user (RQ3). In summary, teens wanted customizable control over the types of nudges they receive while ensuring the risk is prevented from being sent and the perpetrator is held accountable.

Our contribution ties in with the long-standing investment of the SIGCHI and CSCW communities in providing critical insights into teens' online activities in the field of adolescent online safety [23, 65]. Significant contributions have been made that address the types of online risks teens face [27], the coping mechanisms employed [39], as well as current and proposed solutions [20, 38] for youth online well-being. In parallel, there has been a considerable focus on investigating real-time interventions such as nudges for privacy and safety. Our work is situated at the intersection of these research areas with novel contributions. As we move beyond the traditional online safety solutions (e.g., parental controls), we provide novel insights into designing and evaluating nudging interventions for teens' online safety. Based on our methods and findings, we make the following contributions to adolescent online safety research:

(1) A critical assessment of co-designed nudges with teens to iteratively refine online safety nudges for effective implementation.
(2) A holistic understanding for designing adolescent online safety nudges by analyzing both the perspectives of a risk victim and a risk perpetrator.
(3) Actionable guidelines that address key challenges identified for effectively nudging youth as a vulnerable and developmentally unique population.

2 RELATED WORK

In this section, we synthesize the trends and gaps in prior research which discusses a shift from restrictive online safety approaches to teen-centric solutions, prior nudge interventions, and intelligent assistance for online safety.

2.1 Towards Teen-Centric Online Safety Design

Most of the earlier research on online risks with teens assumes a risk-centric perspective [62, 64, 66], where common approaches focus on parental control or restricting the teen's access or usage of the platform to protect them from risks. There are multiple examples of online applications that utilize this risk-centric approach in dealing with adolescent online risks. For example, a study determined that 89% of 75 mainstream mobile applications utilize a form of risk-centric approach in the form of parental controls; parents have the ultimate power to monitor teen's online activity and in some cases, impose restrictions on teens' capabilities in the platform [62]. Empowerment given to the parents is seen as privacy-invasive by the young users for divulging access and control of their intimate social experiences [19, 61], which further discourages adolescents from openly sharing their issues and challenges with their parents.

Recent work concludes that these risk-centric models may hinder the teens' personal development and have negative implications for the parent-adolescent relationship, leading to a less-safe outcome in the long run [47, 62]. It is suggested that risk taking is a necessary element of developing into a young adult, and online safety features need to encapsulate a risk-taking element. This is further backed up by studies such as by Agatston et al. [1] where it is stated that youth have an anti-authoritarian preference as a cyberbullying coping strategy, and would prefer seeking help from their peers or by extension, autonomously. As a result, the research landscape has transitioned towards promoting teen-centric approach to self-regulating adolescent online risks [60, 61]. Yet, research on ways to actually design systems that empower youth to protect themselves online is still in its infancy. While there is a push towards teen-centric approaches, many of the recent solutions are still centered around parental controls [14, 53]. Recently, there have been efforts to empower teens in the process of co-designing online safety solutions [9, 38]. While this approach generates teen-centric ideas for online safety, these ideas are often developed using techniques such as blue sky thinking [58] or big ideas [9], which often do not involve critically analyzing the limitations and practicality of the solutions for real-world implementation. Therefore, there is a need for using a systematic approach to evaluate a comprehensive subset of ideas that are rooted in teens experiences and preferences. Additionally, prior work has shown that youth face a myriad of online risks [2, 65] that deserve further scrutiny. Therefore, we designed nudges contextualized for two risk scenarios; information breaches and cyberbullying, which were found to be two of the most commonly experienced risks by teens uncovered in prior work [2]. Overall, this study builds upon the prior literature by assessing resilience-based solutions (i.e., nudges) with teens that can empower them to effectively manage their online safety independently.

2.2 Design and Analysis of Nudge-based Interventions for Unsafe Online Experiences

A key characteristic of nudges is to preserve the user's autonomy in the decision making process through warnings and to help them make appropriate decisions based on the environment [10]. Such nudging techniques have been successful in several domains (e.g., time management, privacy and security, health, finance, entertainment). For instance, Jurczyk et al. found nudges to be effective in managing excessive screen-time by encouraging computer users to take breaks [28]. Within the fields of privacy and security, nudges have been frequently used to improve password strength [25, 59], reduce phishing or malware scams [40], and so on. While nudges in these related fields have

been frequently studied, nudges for promoting adolescent online safety remain understudied. This is because a majority of the research on nudging interventions within the adolescent online safety space has focused on designing interventions [5, 9], with few iterative approaches for critically analyzing the limitations and feasibility of the proposed designs for teen behavior change and online safety. Additionally, most of this design work has focused on involving parents, experts or adults in the design or research process, instead of involving teens as the primary stakeholders [62]. The few works that collected feedback on nudges for teens' online safety have largely relied on self-reported survey-based feedback from teens [35], which often used hypothetical scenarios or is subject to recall bias in some cases. While these prior works provided critical insights for adolescent online safety nudges, there is a need to have more in-depth feedback on nudges (beyond surveys), that involve teens in the process of understanding practical implications of nudges that help progress towards implementation and evaluations for online safety. We address this gap by involving teens in the in-depth critique and refinement of teen-designed adolescent online safety nudges that can be implemented in future field studies. In Table 1, we summarize how these trends in prior literature influenced our study design.

Table 1. How the Literature Informed Our Study Design

Trends in the Literature	**Our Study Design**
Early work focused on parental interventions to protect youth [62, 66]	Our work accepts this call-to-action by moving away from interventions for parents to analyzing nudge-based interventions targeted at youth.
Prior work called for teen-centric solutions to move away from parental controls to teen self-regulation [7, 19, 63]	
Masaki et al. used survey methods to evaluate online safety nudges informed by the literature [35]	Our work iterated upon online safety nudges co-designed by teens with in-depth qualitative feedback and critique of nudges.
Agha et al. co-designed online safety nudges with youth [2]	

2.3 Building Upon Nudges for Intelligent Assistance and Positive Behavior Change

Within the field of online safety and privacy, different approaches for nudging have been proposed for youth safety. A commonality amongst most of these approaches is that they aim to warn the users, while emphasizing on the harm from risk. This has frequently been done through risk alerts, visual aids or statistics that convey the consequences of the risk [4, 35]. In the context of adolescent online safety, most of the initial work focused on using risk alerts for authoritarian approaches that warned the parent or authorities when a teen encountered a risk [9, 26, 32]. For example, Jayawardena et al. [26] designed a monitoring system that would detect online risks faced by youth and alert parents in case of a risk. More recently, Masaki et al. showed how negatively framed general warnings that are aimed towards teens themselves may be effective for improving their online privacy [35]. Additionally, prior studies have largely focused on solely nudging the victim for risk coping to protect themselves [36, 43]. According to Vale et al. [57], many teens assume a cyber-double role, that is, they can be both the victim and perpetrator in one risk instance, supplementing the importance of studying both perspectives. Another study by Alemany-Bordera et al. [4] supplements this finding by demonstrating that nudges (through bold texts and images) may help risk perpetrators to understand the associated risks and reconsider their harmful actions.

Consequently, there has been a shift towards more teen-centric nudges, that focus on multiple perspectives and move beyond a risk-aversive approach that overemphasizes on victim protection. For instance, On Twitter, tweets exposing violent images or hateful comments can be censored and

marked as sensitive, with more rigorous and unyielding censoring policies put in place for younger (<18 years) users to restrict their access to risky content [56]. However, most of the censorship approaches are automated and imposed with little control in the hands of teens [48]. Relatedly, Badillo-Urquiola et al. found that children prefer to have intelligent assistance in dealing with strangers online [9], such as blocking or reporting to authorities. Apart from intelligent assistance, other common ways of encouraging teens towards positive behavior change have focused on providing incentives or rewards [5, 12]. Yet most of the incentive-based approaches have focused on encouraging risk victims to change their behavior, rather than addressing the risk at the root cause [2, 43]. Most recently, Agha et al. [2] co-designed online safety features with teens' such as intelligent assistance, auto-responses, personalized sensitivity filters and educational advice on how to manage online risks themselves independently. Moreover, they found that teens mostly designed for risk prevention by targeting perpetrators, rather than teens as victims by prompting the risk perpetrator to reconsider their actions through negative incentives. In Table 2, we show how the nudges presented to teens in our work build upon prior work, while most closely aligning with insights from Agha et al. [2]. While their work was mostly focused on co-designing with teens using a blue sky approach where anything was possible, we take a critical lens to analyze the limitations and improvements for these nudges through in-depth interview-based feedback on designs for both the risk victim and the perpetrator.

Table 2. Basing our Nudge Designs and Risky Scenarios on Prior Work

Nudge Dimensions	**Findings from the Literature**
Risk Types: Information Breach Cyberbullying	· Prior work suggested that youth face a myriad of online risks [2, 65] · Agha et al. found information breaches and cyberbullying as two of the most commonly experienced risks by teens [2]
Nudge Designs: General Warnings Sensitivity Filters Guided Actions	· Masaki et al. found negatively worded **general warnings** to be effective for risk prevention. [35] · Maheswaran et al. designed a **sensitivity filter** to censor risks for teens and notify parents. [48] · Badillo-Urquiola et al. co-designed intelligent assistance features for prompting parents and au- thorities for help. [9] · Agha et al. codesigned personalized sensitivity filter and **guided actions** that allow teens to manage online risks independently [2]
Audiences: Victim Perpetrator	· Alemany et al. emphasized on nudging for protecting teens as victims only [4] · Prior work studied perpetrators who were also victims in cyber-double roles [31, 57] · Agha et al. uncovered that teens designed distinct nudges for both risk pre-vention of perpetrators and coping for risk victims [2]

3 METHODS

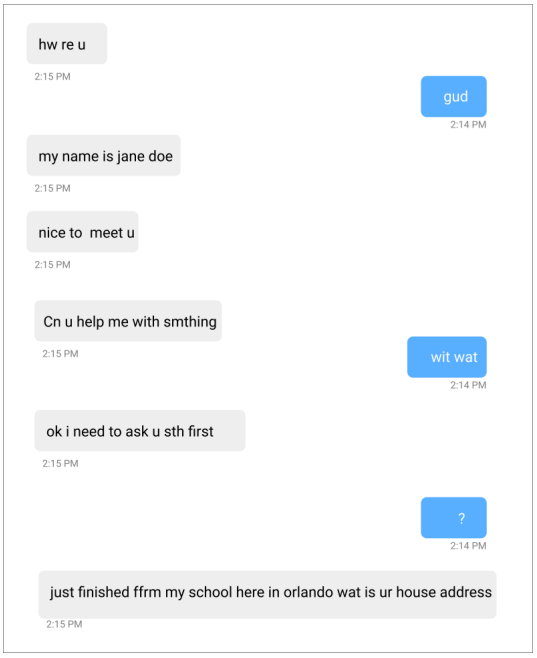
In this section, we describe the methods used to carry out this study, including the study design, nudge descriptions, recruitment strategy, and data analysis approach.

3.1 Study Overview

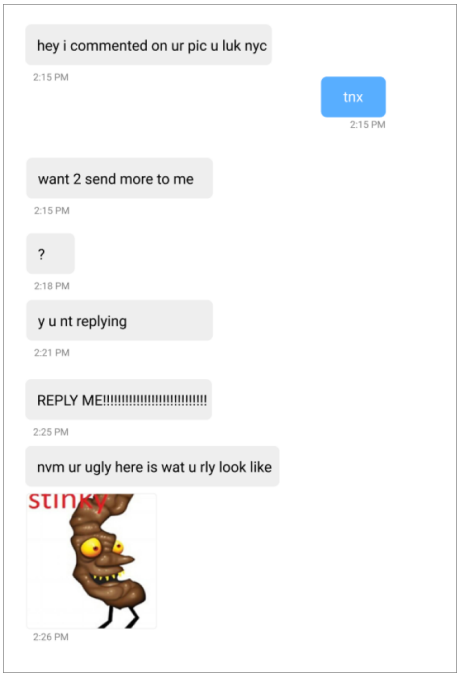
We conducted 15 sessions (3 groups and 12 interviews) with 21 teens from March to July 2022 to get their feedback on three online safety nudges, which were implemented in two risky scenarios. The three nudges were designed based on teens' ideas from previous co-design sessions [2, 3], as well

as current trends in adolescent online safety research and relevant platforms [35, 56]. The nudges include a) General Warning – a pop-up nudge warning the teen about the risk with an option to dismiss, b) Sensitivity Filter - a nudge that censors the received risky content while also giving a warning, and c) Guided Actions - a nudge that suggests risk responses to the user, implemented within two risk types. These nudges were also implemented from the perspectives of both the risk victim and risk perpetrator because the teens from the co-design sessions suggested them over the risk victim nudges.

To carry out the evaluation process, we conducted feedback sessions with 21 participants (aged 13-17) via a Zoom video call and a shared virtual whiteboard (FigJam) to understand their assessment and impressions of these nudges. The sessions were designed in a way that made the teens able to express their ideas beyond words by combining the traditional discussion elements of an interactive user study (interviews, focus groups) with the whiteboarding style of participatory design where they were able to mockup and sketch their feedback [3, 8]. The sessions were run virtually on Zoom in conjunction with FigJam (a collaborative virtual whiteboarding tool) with 1 or 3 teens per session, leading to a total of 21 teens. Around three researchers moderated each session, which lasted approximately 2 hours.



(a) Information Breach Risky Scenario

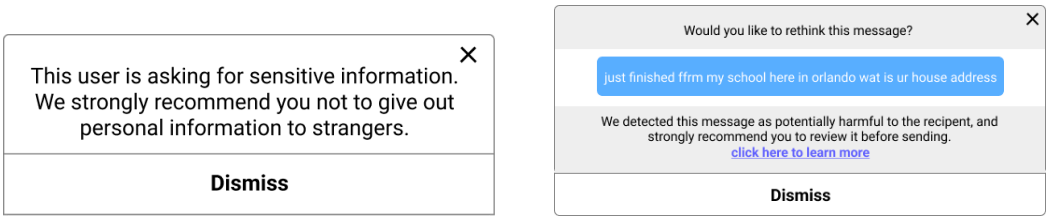


(b) Cyberbullying Risky Scenario

Fig. 1. Risky Scenarios

3.2 Risky Scenarios and Nudge Approaches

The risky scenarios and nudges used in this study are based on common trends in prior research [2], as illustrated in Table 2. To protect the participants, we made use of 2 low-level risky scenarios to implement all three nudges. The full implementations can be found in the appendices.



(a) for the information breach victim (b) for the information breach perpetrator

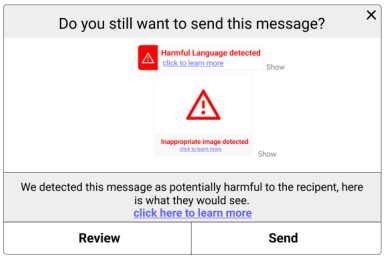
Fig. 2. General Warning Nudge Samples

3.2.1 Risky Scenarios. Information breach and Cyberbullying were one of the most commonly self-reported risks experienced by teens in prior work [2, 22]. As shown in Figure 1(a), the first scenario showed an information breach risk where a stranger sends a private message to the teen and requests them to disclose their personal information (house address). Figure 1(b) shows our implementation of a cyberbullying risk type, where a stranger sends a recurring series of messages to a user, then insults their appearance by calling them "ugly," and sending a vulgar image as a means of insulting them after being ignored [55].

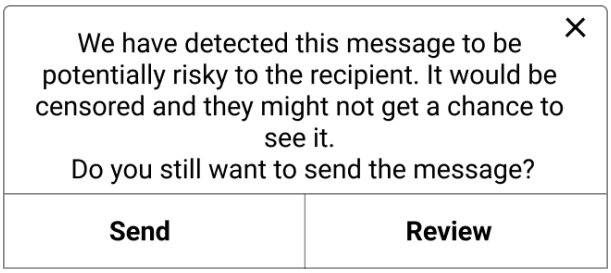
3.2.2 Risk Victim and Risk Perpetrator Perspectives. In this study, we designed nudges for both risk coping for victims and risk prevention for perpetrators based on prior online safety research [2, 31]. A major aspect of the feedback process involves the user as a victim in a hypothetical risky scenario to assess the effectiveness of the nudges in that context. However, for the same scenario, we also took a preventive approach of using nudges to deal with online risks by injecting the nudge to the risk perpetrator as a means of preventing the detected risk from being sent. We implemented both perspectives by introducing the teens to the risky scenario as a victim, giving them the prepared prompts, and getting their feedback, before finally asking them to provide feedback from the perspective of the risk perpetrator. In this process, we did not ask teens to play the role of a risk perpetrator, but to provide feedback based on whether they think the nudge would prevent the risk perpetrator from sending the risky content.

3.2.3 Nudge Designs. Prior literature implemented general warnings through risk alerts, emphasis on the risk, and negative framing may assist with online safety [4, 35, 59]. We implemented this in the form of a **General Warning** nudge which uses visual cues and a warning to let the teen know that they are experiencing a risk and urge them to be careful. This is implemented in figures 2(a) as a pop-up warning to a potential victim as soon as the risky message is detected. In figure 2(b), the risk perpetrator receives a similar pop-up message with a snippet of their message, informing them that the message has been detected as risky while recommending them to review it. Apart from the cropped samples of the nudges provided in Figures 2-4, the complete versions of each nudge can be found in the appendix.

Additionally, prior work co-designed censorship features that block offensive content [5, 12], with recent interventions promoting more control for teens with the ability to view the censored risk [2, 56]. This informed our second nudge, a **Sensitivity Filter** which censors the risky content (in addition to providing a warning) to the risk victim, while giving them the option to show or hide the content, as shown in figure3(a). In figure 3(b), the risk sender receives the Sensitivity Filter as a pop-up reprimand that the message has been tagged as risky, and the receiver would see a censored version, which functions as a disincentive to the perpetrator .

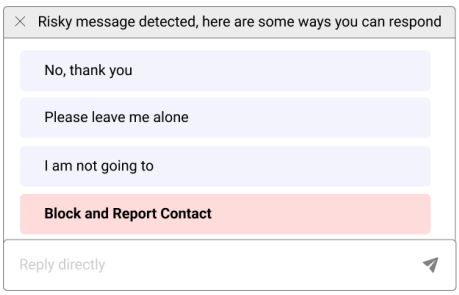


(a) for the cyberbullying victim



(b) for the information breach perpetrator

Fig. 3. Sensitivity Filter Nudge Samples



(a) for the information breach victim

(b) for the cyberbullying perpetrator

Fig. 4. Guided Actions Nudge Samples

Findings from prior research also recommended automated or intelligent forms of assistance [2, 9, 59] aiming to improve the users' online safety, by providing safe actions (e.g., block, delete) or responses to the user, which informed the design of our third nudge; **Guided Actions**. This nudge involves the platform generating safe responses and actions for the teens to deal with the risky scenario. When the teen receives a detected risky scenario, they are presented with a list of safe auto responses as a suggestion, with an option to respond directly. This is shown in figure 4(b). For the risk perpetrator, each detected risky content by the teen is underlined and marked red while they are typing with suggested replacement messages and a clear message option above the text box as shown in 4(a).

3.3 Session Procedure

The sessions were conducted online via a Zoom video conference and lasted approximately 120 minutes. At the start of each session, each participant was introduced to the researchers, the concept of nudging, how nudging is used in other domains, and an ice-breaking activity. The teens were then given a crash course on adolescent online safety and nudges to understand the goals of the session and what is expected of them. Considerable effort was also put into making the teens aware of the subjectivity of their responses. To promote interaction, the participants were asked to discuss what online safety means to them as a teen and to share an instance when a nudge influenced their decisions.

The feedback process included 12 nudges spread across (a) 2 perspectives, (b) of which each had 2 scenarios and (c) with 3 nudges each. The scenarios and their respective nudges were presented to the teens through a click-through high-fidelity prototype on Figma, this was done to allow them to clearly understand the flow of the scenarios and the functions of each nudge. To request feedback for the nudges, a researcher walked the teens through a prototype of the risky scenarios (in the order of Information Breach, Cyberbullying) and their associated nudges (in the order of General Warning, Sensitivity Filter, and Guided Actions), after which was a switch to the whiteboarding and discussion activity on FigJam to allow the teens provide their feedback. The researchers used sticky notes, drawing tools, and shapes to mock up the teen's feedback and ideas to improve the nudges, with a mix of verbal and design elaboration over the nudge prototype. The presentation and co-design activity processes were carried out for each nudge, and the teens were asked to rank the nudges in order of their personal preferences. We followed a semi-structured approach in giving the participants question prompts to generate feedback. This was done to allow the participants to express themselves while giving the researchers room to get the needed feedback. The general prompts given are highlighted below.

(1) Prompts provided per scenario (before nudge presentation):
   - What do you think about this conversation?
(2) Prompts provided for each of three nudges, per scenario:
   - What do you think about this nudge?
   - Does this nudge address or not address the risk, and why?
   - What are the pros and cons of this nudge in your opinion?
   - How can this nudge be improved? If you were the designer, what would you change?
(3) Prompts provided per scenario (after nudge presentation):
   - Rank all nudges and explain why
(4) Final Prompt (after all nudge and scenario presentation):
   - What is your best nudge overall?

3.4 Ethical Considerations For Teen Participants

This study was approved by the Institutional Review Board (IRB) at the University of Central Florida (UCF). Additional methods were also taken to ensure the safety of teen participants, as minors and a vulnerable population. We obtained parental consent as well as teen assent, where teens were encouraged to independently complete an assent form to confirm their interest in the study. Other protective measures included the use of low-level risks, allowing participants to interact privately with the researchers using the Zoom private chat feature, and data de-identification for analysis. Additionally, when reviewing the low-impact risk scenarios from the perspective of the perpetrator, participants were reminded to think about the perpetrator from a third person point-of-view, instead of directly playing the role of a risk perpetrator. We also compiled professional help resources based on the recommendations of UCF's Community Counseling Clinic, and UCF's Counseling and Psychological Services to share with participants in the case of any distress, and provided the parents/guardians and teens with an opportunity to withdraw from the study at any point. Overall, we complied with several guidelines provided by Badillo-Urquiola et al. [8] for conducting ethical research with teens on sensitive topics such as giving teens autonomy through teen assent, prioritizing data privacy and de-identification, and providing help resources. To our knowledge, none of the participants experienced harm or distress as a result of participating in this study.

3.5 Participant Recruitment and Demographics

After receiving IRB approval, the participants were recruited from youth-serving organizations, social media, and middle/high schools around the United States. The mode of recruitment included

Table 3. Participants' Demographic Information

Session	ID	Age	Sex	Ethnicity	Best Overall Nudge
Session 1	P1	14	F	Black/ African	Guided Actions
Session 2	P2	15	F	Black/ African	Sensitivity Filter
	P3	16	M	White/Caucasian	Sensitivity Filter
	P4	13	M	White/Caucasian	Sensitivity Filter
Session 3	P5	13	F	Black/African	Guided Actions
Session 4	P6	17	M	White/Caucasian	Sensitivity Filter
	P7	16	F	Asian	Sensitivity Filter
	P8	16	F	Asian	Sensitivity Filter
Session 5	P9	17	M	Asian	General Warning
Session 6	P10	13	M	Asian	Guided Actions
Session 7	P11	15	F	Asian	Sensitivity Filter
Session 8	P12	17	F	Asian	Sensitivity Filter
Session 9	P13	16	M	White/Caucasian	Guided Actions
	P14	16	M	White/Caucasian, Hispanic/Latino	Guided Actions
	P15	16	F	Black/African	General Warning
Session 10	P16	16	F	Black/African	Guided Actions
Session 11	P17	17	F	Asian	Guided Actions
Session 12	P18	16	F	Asian	Sensitivity Filter
Session 13	P19	16	F	Asian	Sensitivity Filter
Session 14	P20	17	M	Asian	Guided Actions
Session 15	P21	16	F	White/Caucasian, Hispanic/Latino	Sensitivity Filter

flyer distribution, phone calls, and emails. We supplemented our open recruitment strategy with existing contacts in our research lab's (STIR Lab) participant database from previous studies. All participants were aged 13-17, based in the United States, and able to communicate in English. The majority of the teens were 16 years (N=10, 47.6%) with the mean age and standard deviation being 15.6 and 1.29 respectively. The identified racial identities of the participants are as follows: Asian (47.6%), Black/African American (23.8%), Hispanic/Latino (9.52%), and White/Caucasian (28.5%). The participants comprised of 8 males (38%) and 13 females (64%) **(Table 3)**. They were compensated with a $20 Amazon gift card on completion of the study.

3.6 Data Analysis Approach

The data obtained from the teens' feedback via the whiteboard annotations and session transcripts were analyzed using Braun and Clarke's thematic analysis [13]. The sessions were transcribed to text using Otter.ai. The primary source of data is from the responses and discussions stemming from the question prompts given during the presentation of each nudge. The co-design whiteboard artifacts were also considered, as some of the annotations and sketches over the nudge design contained valuable information. Two researchers participated in the data analysis process, where there was an initial coding of the data to note emerging ideas that were grouped into the major themes. The codebook was initially coded along the dimensions of nudge response, nudge feedback, and areas for improvement. The two researchers had consistent meetings to merge their individual codes and resolve conflicts. The nudge ranking data was treated quantitatively and analyzed using cross-tabulation. For RQ1, each participant ranked each nudge type for the two risk scenarios and two user perspectives, which have been averaged for consistency. The quantitative rankings

helped answer the question of which nudges were preferred, while the qualitative thematic data answered why those nudges were preferred, as well as the challenges and recommendations. The final codebook in Table 4 consists of Nudge Challenges and Nudge Recommendations as dimensions.

4 RESULTS

In this section, we discuss the key findings from our qualitative analysis of the three focus group and twelve interview sessions to answer our research questions.

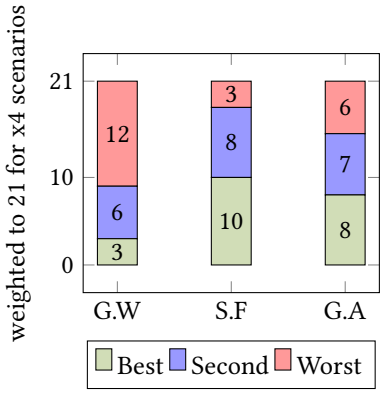
4.1 Most Effective Nudges in dealing with Unsafe Interactions Online (RQ1)

Overall, most teens (N=18, 85.7%) preferred either the Sensitivity Filter or Guided Actions nudges for dealing with the risk scenarios presented to them. The teens' preferences for those two nudges were seen consistently across both the perpetrator and the victim perspectives. In contrast, most teens (N=12, 57.1%) considered the General Warning nudges to be the least effective nudge for all risk scenarios and perspectives. The main determinant for their impressions of these nudges is the provided responses to the risk. The Sensitivity Filters were preferred for actively preventing the teens from viewing harmful risks, while the Guided Actions were preferred as they helped teens make safer choices in response to the risk and finally, the General Warning nudges were viewed unfavorably for not serving an additional function besides providing a warning message. The feedback is covered in more detail in the following sections. In (RQ1a), we present the comparison of the three nudge types based on the ranking by teens from most to least preferred. Next, we demonstrate which nudges were considered most effective by teens for dealing with the two risk types (RQ1b), then we illustrate which nudges were preferred based on the two user perspectives (RQ1c). Finally, all aggregate rankings were weighted to a total count of 21 for consistency.

4.1.1 Sensitivity Filters were generally considered to be the most effective nudge type as they prevented risk exposure for victims (RQ1a). While teens liked all the nudge types for providing a degree of risk awareness, their preferences depended on the functionality and choices provided by the nudges. When asked to rank the three different nudge approaches, most teens (N=10, 48%) considered Sensitivity Filter to be the most effective nudge as it censored the risk which prevented them from risk exposure while providing control to the user in case they wished to view the risk. Compared to other nudges, teens preferred the Sensitivity Filter as it provided complete protection from the risk and a more visual warning, which was found to help understand the gravity of the risk. The Guided Actions nudge was considered to be the second-best nudge type as most teens (N=8, 38%) found the actionable prompts (e.g., Block, Reply directly) to be helpful, especially in a situation where they may not know how to respond to the risky scenario and need guidance on what to say or the right action to take. For example, P20 explains the rationale behind their rankings with the favorable ones being Sensitivity Filters and Guided Actions due to their characteristic functionalities, while the General Warning nudge was ranked poorly because warning-only was not considered good enough.

> *"The Sensitivity Filter would actually be the top just because you don't really see the image or the message, and then I would do maybe the Guided Actions one next because the options of having like being able to choose block or - , and probably just the General Warning last because you can still see everything in the chat, and it doesn't do anything."*- P20. 17, M. (Cyberbullying Scenario)

However, some teens (N=6, 29%) ranked Guided Actions as the least effective nudge and criticized it for possibly making the risk worse for the user by encouraging victim-perpetrator communication, therefore prolonging the risky encounter. Lastly, the General Warning nudge was ranked last by



(a)
G.W - General Warning, S.F - Sensitivity Filter, G.A - Guided Actions

Fig. 5. Rankings - All Nudge Approaches as Ranked by Teens (RQ1a)

most teens (N=12, 57%) as it did not provide any real solution or actionable choices for the teen, making them feel like the nudge was not effective in helping during a risk.

Overall, these findings highlight that the teens believe Warnings to function as the baseline feature of an online safety nudge, where it is ineffective on its own without any additional features. While the most preferred feature to make a truly effective nudge depends on the context of the risk and individual (4.2), they mostly preferred censorship due to its ability to initially shield a victim from harmful risks and serve as a deterrent against a risk perpetrator.

4.1.2 While Guided Actions were considered more appropriate for Information Breaches, teens preferred Sensitivity Filters for Cyberbullying (RQ1b). Our findings showed that teens prefer different nudges for different risk types. While the danger of an Information Breach scenario is in responding and divulging personal information to the risk perpetrator, the risk of a Cyberbullying scenario lies in the exposure of the victim to the risky messages. While in RQ1a, the teens generally preferred the Sensitivity Filter overall, there was a difference in their most preferred nudge when grouped by scenario due to the nature of the risks as explained by P15.

> *"If someone says, 'hey what's your address?' that's not like affecting me directly, like I could just ignore it because like it's not that big of a deal, but if they are genuinely sending inappropriate pictures, maybe like explicit pictures, I feel like the youth can be a lot more susceptible to that."*- P15. 16, F.

Most teens considered Information Breaches to be a commonly encountered, low-impact risk that they did not mind being exposed to. Therefore, they wanted the ideal nudge to help them protect their personal information by mirroring the appropriate responses to the risk sender such as ignoring or blocking them, and this led to most teens (N=10, 48%) ranking Guided Actions Nudge first for allowing the teen to do both actions. Slightly fewer teens (N=9, 43%) ranked Sensitivity Filter best for incentivizing the teen to not respond to the risk perpetrator via censorship. However, more teens (N=7, 33%) ranked Guided Actions worst in comparison to Sensitivity Filter (N=4, 19%), which indicates a degree of mixed opinions on the Guided Actions nudge. This was due to the possibility of the provided message responses prolonging the risky scenario as discussed in 4.3.3.

(a) Information Breach (b) Cyberbullying

G.W - General Warning, S.F - Sensitivity Filter, G.A - Guided Actions

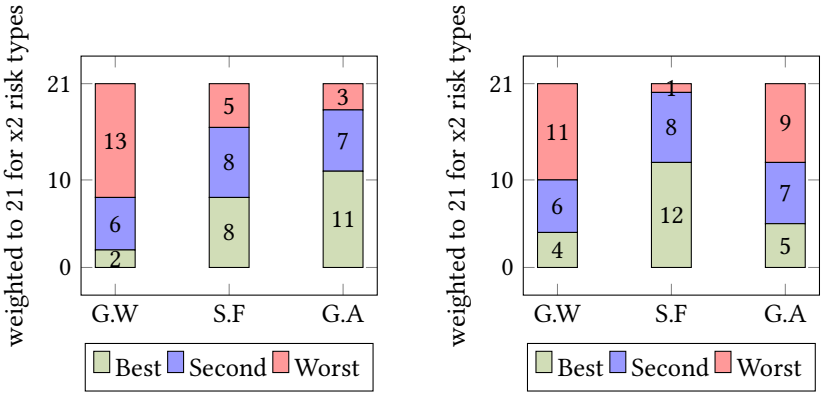
Fig. 6. Rankings - Nudges by Risky Scenario (RQ1b)

Teens had a different preference for the Cyberbullying scenario. In most cases, they did not want to be exposed to it as a higher-impact risk because it included personal and hurtful remarks or explicit content, which led to a more severe outcome for the teen victim. Therefore, most teens (N=12, 57%) ranked the Sensitivity Filter best due to its ability to censor and provide complete initial protection from the cyberbullying content. While some teens liked Guided Actions (N=6, 29%) for easily giving them the option to block the user, they still felt censorship was key to safety against a risk like Cyberbullying where solely exposure (without interaction) can leave a lasting negative effect on them.

Finally, teens' negative impressions of the General Warning nudge remained consistent in both the Information Breach scenario and the Cyberbullying scenario with most teens ranking it worst for both Information Breaches (N=10, 48%) and Cyberbullying scenarios (N=14, 67%) respectively. This is because warning-only as implemented in the General Warning Nudge was considered to be ineffective on its own in successfully solving both presented risk types.

In conclusion, the teens preferred having Sensitivity Filters for more explicit and interpersonal risks, such as Cyberbullying due to its function as a safety screen between the victim and harmful content. The automated actions and responses offered by the Guided Actions were considered more effective for risks perceived lower in severity, such as the Information Breach Risk where the teens felt that ignoring or blocking the perpetrator was sufficient for safety, and warning-only was considered inefficient for not implying an active solution or response to the risk.

4.1.3 Guided Actions were considered to work best for the risk victim, and Sensitivity Filters for the risk perpetrator (RQ1c). With the nudges being implemented from the perspective of both the risk sender and risk victim, most of the feedback remained the same, although they were for reasons unique to the target perspective. For the risk victim, most of the teens (N=11, 52%) ranked the Guided Actions nudge best for providing the right action to help the victim respond to the risk safely, and relatively fewer teens (N=8, 38%) ranked the Sensitivity Filter best for protecting the teen victim from being exposed to harmful content via censorship. Specifically, teens preferred Guided Actions over Sensitivity Filters for risk victims as the options of ignoring and blocking were more applicable to risks they encountered regularly online. Additionally, Guided Actions had



(a) Risk Victim (b) Risk Perpetrator

G.W - General Warning, S.F - Sensitivity Filter, G.A - Guided Actions

Fig. 7. Rankings - Nudges by User Perspective (RQ1c)

more proactive and permanent safety choices (such as block) in comparison to Sensitivity Filters which was preferred for preventing exposure to more infrequent, severe risks.

For the risk perpetrator, the Sensitivity Filter was ranked best by most of the teens (N=12, 57%), as the reprimanding language in the warning message (which states that the victim would receive a censored version of the message) was seen as the most effective tool in making the risk sender reconsider their harmful actions. Teens considered the enforced prevention of the risk to be the preferred approach for reducing online risks, especially for risk perpetrators that have an intent to harm who may not be easily influenced. Meanwhile, many teens (N=9, 43%) ranked Guided Action worst as they were concerned that the harmful intent of the perpetrator may lead them to disregard the safe recommended actions and messages, and the perpetrator would be free to continue the risky interaction. P17 explained,

> *"I like the victim's perspective because there were all the options to like delete it and block the person, whereas for this one [sender's perspective], the sender can still send the message and there's no way of like protecting the person who received the message."*- P17. 17, F.

Lastly, the General Warning nudge was ranked worst for both parties. It was ranked worst for the victim because it was considered easy to ignore without providing an actionable response to the risk, while it was ranked worst for the risk perpetrator because they are assumed to have a strong intent to cause harm which cannot be changed with a warning nudge. In conclusion, teens preferred actionable responses with Guided Action (block, report) for victim nudges to assist them in responding to the risk appropriately, while reprimanding language and risk prevention through Sensitivity Filters were preferred for risk perpetrators who may not be easily convinced to improve their behavior.

4.2 Key Differences that Influenced Teens' Perception of Nudges (RQ2)

Our results highlighted some differentiating emerging themes that affected teens' perceptions of online safety nudges. We elaborate on these themes in the following subsections.

4.2.1 Teens consider Sensitivity Filters to be more effective for image-based risks over text-based risks. While there are many kinds of online risks faced by adolescents, these risks can also be experienced

in different mediums including text, image, audio, and video [3]. Based on the feedback provided, the teens' responses identified the medium in which the risk was sent to be a major determinant of how the nudge would be perceived. This variation is seen in the Cyberbullying risk scenario (Figure 1b) where the risk occurred via two mediums: (1) text and (2) image. The teens independently identified this difference and expressed that they would be more wary of an image-based risk over a text-based risk, extending to a preference of an image-based sensitivity filter over a text-based one. Noticing this disparity, we followed up with them by asking how they would respond to the censorship of both risk mediums within the Cyberbullying risk instance. While the majority of teens (N=10, 47%) provided mixed or ambiguous responses to how they would respond to the nudge, more teens said they would uncensor the risky text (N=6, 29%), than the risky image (N=4, 19%). This is because teens found the image-based risks to be more severe as they contain explicit or offensive imagery, requiring more censorship than text-based risks, which teens commonly encounter online or find less offensive. For instance, P19 explained how they are more likely to view a harmful image as more dangerous than a harmful text.

> *"Usually, if I receive a harmful photo, I probably wouldn't want to see it and I would leave it censored. If I receive harmful words, I would just ignore the censor and view it, I think it is because I would be curious to see what the word says."*- P19. 16, F. (Cyberbullying Scenario)

In comparison, many (N=13, 62%) of the participants said they would view the censored material in the Information Breach scenario (which does not include an image-based risk) due to it being a text-based risk, which was also considered low severity in 4.1.2. In conclusion, our results suggest that even for the same risk type, teens' perceptions and responses to a nudge may change significantly due to the medium of the risk, where teens expressed a preference for censoring image-based risks over text-based ones with Sensitivity Filters to avoid exposure to explicit content.

4.2.2 Teens believe effective nudge approaches need to directly relate to the severity and type of risk being addressed. As previously noted in **RQ1b rankings** (Figure 6), our results show that teens can have different nudge preferences for different risk types. This is evident from teens' different preferences for nudges for Cyberbullying and Information Breaching risks, where teens considered Sensitivity Filters to be best for Cyberbullying but preferred Guided Actions for Information Breaches. We also observed some participants did not want a nudge for one risk but preferred it for another, based on their perceived severity of the risk, i.e., Cyberbullying being high severity, and Information breaches being low severity. For example, P1 described the Cyberbullying scenario as "another level" of a risk, when compared to the Information Breach risk scenario. They did not feel the need to have any nudge for the Information Breach attempt because it was considered a common occurrence that is to be reasonably expected in a social media interaction, whereas they found the General Warning nudge more useful for the Cyberbullying risk, as a higher-impact scenario.

> *R1 - "You said you did not like this kind of nudge (General Warning) for the previous information breach risky scenario. So, what is different?"*
> *P1. 14, F. (Cyberbullying) – "This is like a whole other level, because like, I'm pretty sure the person who is like asking for like more pics wanted like something explicit."*

This insight can be used to design nudges catered to different risk types, as well as the severity of the risk as low, medium, or high level. Teens also felt that the "strength" of the intervention should be proportional to the severity of the risk. For instance, commonly occurring risks such as Information Breaches may not warrant the need for a nudge which some feel could be handled without assistance. Similarly, teens preferred the Sensitivity Filter nudge only when the risk was

considered severe enough (i.e., explicit content or improper requests). In conclusion, the risk teens encounter online are varied and dynamic with multiple variables, which in turn requires dynamic online safety interventions to ensure the effectiveness, as each nudge may not be equally appropriate for all risk types.

4.2.3 Teens believe nudges with stronger language and stricter measures may be more appropriate for younger, less mature teens. While our participants included teens aged 13-17, some participants (N=5, 24%) mentioned a disparity in how younger and older teens might behave in an online risk instance. Teens suggested that older teens might have a greater degree of maturity and competence in dealing with their online safety, while younger teens might be more vulnerable to these online risks and would benefit from stricter safety measures because they might not have enough experience dealing with unsafe interactions online and can be easier manipulated, as explained by P2 in the session extract below,

> *"Older teens have somewhat of an idea of what to do, younger teens can feel pressure to be put in a corner to answer to adult [perpetrator] authority."*- P2. 15, F. (Information Breach Scenario, Guided Actions Nudge)

With most participants (N=17, 81%) being older teens ($\geq$ 15*years*), the vast majority of them (N=19, 90%) recommended the use of independent tools (safe actions) to help them respond to the risk, while only (N=3, 14%) teens suggested parental interventions as a solution to the presented online risks. Interestingly, like the quote from P16 below, teens often described the education level (high school, middle school, etc.) as a measure of how aware or vulnerable a potential teenager would be in an online risk indicating that even within adolescence, teens can have vastly different experiences, awareness and resilience in dealing with online risks, depending on multiple factors including their age, education level and prior experience with online risks. P16 explained,.

> *"If they have social media, and they're not in high school, they should definitely be censored. Because middle school and high school are two completely different things, and they live different experiences."* - P16. 16, F.

The participants also suggested other high-level behavioral differences between younger and older teens, such as younger teens being more sensitive to risks or more bound to making unsafe responses to the risk, which further emphasizes the need for differing online safety solutions for teens in different developmental stages, such as stronger warnings, stricter controls or parental help for younger teens.

4.2.4 Teens believe an unaware victim is significantly more vulnerable. This section explains how the teens identified a learning element as a key factor in online safety risk management, the teens indicated a teen victim's awareness of being in a risky situation determines their outlook of the scenario, as well as how relevant they would view the interventions addressing them. For example, a teen who has encountered fewer risks online may be more likely to find the content of an online safety intervention beneficial as opposed to a more risk-experienced teen who may find the intervention to be unnecessary. As an extension of awareness, the teens also mentioned the victim's ability to independently handle the risk (knowing how to respond to it) means a nudge may come off as obnoxious or unnecessary to them. As P10 explains below, solely providing a General Warning to a teen victim in the provided Cyberbullying instance is not needed and would be considered common sense by most teens, as the risk is already apparent to them, supporting the theme in 4.3 which shows that teens find nudges to be easy to ignore. This finding indicates that nudges will be less disruptive and more useful if they are tailored to the awareness and prior risk experiences of youth. Alternatively, nudges may be necessary for infrequent and severe risks that

are not commonly encountered by teens who may not have the ability to independently cope with the higher-level risks. P10 summarized,

> *"It tells you to be careful. But I guess it's just common sense to be careful that if you just see someone texting you something that looks like that, that they're just weird people or someone that's stupid."*- P10. 13, M. (Cyberbullying Scenario, General Warning Nudge)

Likewise, the teens also suggested that a risk perpetrator's perception of a nudge would depend on their intent or awareness of sending the risky material. For example, an adversarial and intentional risk perpetrator might have a more dismissive tendency towards a risk intervention and may require stricter preventative measures and enforcement, while someone who may accidentally perpetrate a risk might find a subtle nudge such as the General Warning to be a sufficient reminder to effectively change their behavior. The effects of this variable are discussed under 4.3.1.

4.3 Challenges and Recommendations Identified by Teens for Designing Effective Nudges (RQ3)

While giving feedback, teens identified challenges with online safety nudges and offered recommendations to improve the nudges. The challenges and recommendations are summarized in the codebook shown in Table **4**. The themes in the nudge challenges and recommendations sections are put in a 1-1 mapping and explained in the following subsections.

4.3.1 Designing Convincing and Tailored Nudges. Most teens (N=15, 71%) believed that a major challenge with the presented nudges was that they were **easy to ignore** and ineffective in altering a user's action. For example, teens believed that a risk sender with an intention and incentive to cause harm to a teen might not be easily nudged away from their action. Similarly, a risk victim who is convinced that they are not in a risky situation, either due to naivety or assuming a false-positive risk detection instance will remain uninfluenced by the nudge. In addition to nudges being easy to ignore due to the user's intent, the teens said nudges are also easy to ignore due to an intent that cannot be changed by nudging, such as a subjective desire to "not want to read" or other forms of biases. Specifically, the teens were concerned about warning fatigue (N=7, 33%), where the nudge loses its impact after being triggered multiple times for the user, making them redundant or disruptive. This issue is made worse by a majority of users having an experience of being nudged improperly in the past, i.e., being nudged too frequently, or for the benefit of the platform at the expense of their own interests, leading to one nudge leaving a bad impression that becomes associated with other nudges. Some teens (N=2, 10%), even likened the nudge popup format to a virus, i.e., a malicious JavaScript popup [30], that the teens might mistake for a malicious pop-up leading to malware, identity theft or other negative outcomes.

As a solution to nudges being easy to ignore, the teens wanted the nudges to **provide more convincing warnings**, and most teens (N=20, 95%) suggested emphasizing the risk harm using design cues. For instance, P11 recommended the use of visual or design elements such as bold text, which can help draw the teen's attention to the risk.

> *"Right now... it's just boring, I don't want to read. You should capitalize harmful, or somehow convince them how dangerous this image might be."*- P11. 15, F. (Cyberbullying, General Warning)

Other teens suggested using color coding, account flagging, and danger signs. With the teens' emphasis on design cues over text, many teens (N=17, 81%) still wanted the warning text on the nudges to justify or explain the need for the warning by being specific and tailored to address the intricacies of the detected risk; with the goal being to clearly communicate the underlying threats identified by the nudge to the user. In summary, the teens expressed a key challenge with

online safety nudges as their significant possibility of being ignored, either due to the user's intent or warning fatigue, and they recommended better, more effective nudges to mitigate this by emphasizing the risk via design cues and justifying the given warning.

4.3.2 Nudging for Timely Intervention. Many of the teens (N=15, 71%) considered online safety nudges to be **ill-timed or intrusive to regular use**, i.e., they may be disruptive to the flow of everyday conversations, or possibly triggered too late after the preventable harm had been done. To overcome the issue of disruptive nudges, teens suggested an option to disable online safety

Table 4. RQ3: Nudge Challenges and Design Recommendations

Dimensions	Themes	Sub-themes	Count	Exemplars
Nudge Challenges	Nudges are easy to ignore	The user's intent can not be changed by nudging	(15, 71%)	*"I don't think this would be helpful at all, because they already have **harmful intentions**"*- P7, 16, F
		Nudges might be prone to warning fatigue	(7, 33%)	*"After a while, teens will **get so used to that pop-up** that they'll sort of **ignore** it"*- P4, 13, M
	Nudges could be optimized better	Nudges might disrupt the user flow	(15, 71%)	*"It would get **tedious** to have that [nudge], like to have it **in every conversation**"* - P21, 16, F
		Nudges may appear too late	(9, 43%)	*"**Prevent** repeated unwanted messages **before** you got to the point of like, this image"* - P6, 17, M
	Nudges can have adverse effects	Censorship might make the risk more appealing	(15, 71%)	*"I don't feel like teens are going to ignore this like, 'Oh, harmful, harmful'..it might even get some teens to **see it more**"* - P3, 16, M
		Nudges might escalate the risk they address	(12, 57%)	*"The **idea** of like **continuing the interaction** [with the risk sender] could also be **harmful**"*- P21, 16, F
Nudge Recommendations	Nudges should warn the user better	Emphasize risk harm using design cues	(20, 95%)	*"**Capitalize** harmful...if they don't look at the message, maybe [they'll] look at the **bold words**"*- P11, 15, F
		Explain the need for a warning	(17, 81%)	*"It could **elaborate** on like **why** they don't recommend it...why should I care?"* - P9, 17, M
	Nudging for optimal timing and actionability	Provide safe actions for users	(19, 90%)	*"I feel like it should give you **an option to block the user** [risk perpetrator]"* - P4, 13, M
		Limit who safety features apply to	(15, 71%)	*"I would like [the nudge] to **not** be there **when I'm tal king to my friend**"* - P10, 13, M
		Require the user to confirm their action	(10, 48%)	*"There could be **another pop up** like, 'are* textbf *you sure?'like, 'this is sensitive information' "*- P2, 15, F
	Nudging to prevent the risk	Prevent the perpetrator from sending risk	(17, 81%)	*"You want to **prevent** them [risk perpetrator] **from like ever sending it out** instead of, they send it out and like someone [victim] just blocks it"*- P20, 17, M
		Include reprimands for the risk perpetrator	(17, 81%)	*"Make you [perpetrator] **aware** that the receiver end [victim] might **delete** it and **not get** a **chance to see it**"*- P12, 17, F
		Penalize perpetrator for their harmful actions	(5, 24%)	*"Maybe like a **strike**, and like[in] three strikes, you're**banned**"*- P9,17, M

nudges for trusted contacts, so their everyday conversations are not falsely flagged and nudged for unsafety. Alternatively, teens suggested designing nudges in a way that is subtle, seamless and does not detract from the user flow so that the nudges do not overwhelm the teen, or significantly add more steps to the communication process. Additionally, the teens also wanted those solutions to be triggered promptly before the risk escalates, especially for high-level risks such as the cyberbullying or explicit content exposure. Several teens (N=9, 43%), including P9 below, criticized the nudges implemented in the Cyberbullying victim scenario for being triggered too late because they believed the nudges should have been triggered at the initial point of contact with the risk perpetrator, as a means of preventing the risky scenario from reaching a more serious level. These challenges are not a conceptual problem with the nudges, but rather, an opportunity that **nudges could be optimized for timing and actionability** to provide a more pleasant experience. P17 elaborates,

> *"Have filters in place to prevent repeated unwanted messages before you got to the point of like, this image."*- P6. 17, M. (Cyberbullying)

Many teens (N=19, 90%) also expressed a dislike for some of the nudges that provided no actions for the user to take in response to the risk - which suggests an emphasis on online safety nudges to provide safe actions or responses in addition to notifying the user about the risk. They found warning-alone to be ineffective in addressing their online safety issues, because it was insufficient for only notifying the teen about the risk, without providing an easy to access call to action that can help the teen cope with the identified risk. For example, P8 says,

> *"There's like no proactive option that you can take, like okay, 'This is risky,' you tell us that, but there is no option to act on it."*- P8. 16, F. (Information Breach, Sensitivity Filter)

Finally, especially in dire situations, the teens (N=10, 48%) suggested a follow-up "are you sure?" nudge to require the user to confirm their action in case they ignore the initial nudge and plan to continue engaging with the risk. The motivation behind this suggestion is that some teens might select an option by accident, impulsively, without reading, or sufficient consideration. Therefore, a supplementary nudge asking them to confirm their actions or possible doubts and reminding them of the risks involved was recommended.

> *"I guess that's kind of a downside to it. because if they tap it [dismiss] and it sends [without any other safeguards]. Then like, it doesn't really do much."*- P1. 14, F.

In conclusion, teens want online safety nudges to not only warn them of the risks, but also want proactive actions through these interventions that can help them cope with the risk, provided at the right and optimal time, and for the right audience, in a way that creates the least inconvenience.

4.3.3 Addressing the Adverse Effects of Nudging through Risk Prevention and Perpetrator Accountability. The participants mentioned that **nudges can have adverse effects** through inadvertently worsening the risk being addressed. Most teens (N=15, 71%) were concerned about censorship-induced curiosity with the Sensitivity Filter nudges whereby censorship might make the risk more appealing or attractive to the teen leading to them paying more attention to it than if not censored. As mentioned in 4.2.1, we asked the participants how they would respond to the Sensitivity Filter as a victim, the results show that in multiple cases (Information Breach, Cyberbullying), a teen victim is more likely to uncensor ((N=13, 62%), (N=6, 29%)) a detected risk than take the safer approach by leaving it censored ((N=5, 24%), (N=5, 24%)). P3 elaborated on their rationale for uncensoring the risk,

> *"I don't feel like teens are going to ignore this and be like, 'Oh, harmful, harmful, oh I'm not going to look at this.'… I'm being totally honest with you, like I'm a teen, if I see that, I'm clicking show like, I want to see."*- P3. 16, M. (Information Breach, Sensitivity filter)

Another challenge was raised with the Guided Actions nudge for both the victim and sender. For the victim, many teens (N=12, 57%) felt responding to an aggressor (which is suggested by the nudge) might escalate the risk or prolong it, either by the risk sender forming a harmful bond with the victim, or the victim suggesting they are interested in continuing the risky conversation. While they felt this way for the automated message responses, they still had a positive response about the automated action responses, which provided proactive and permanent choices for addressing the risk. For the risk sender, the teens felt (especially for an adversarial risk sender) that the Guided Actions nudge could worsen the scenario by letting the risk sender know what what is wrong with their messages and teach them how to more covert at perpetrating risks, making it more difficult for the victim and platform to identify their risks. Therefore, while the nudges were aimed at making the teens feel safer, they identified several ways that the nudges could be misused. For example, the risk victim might inadvertently prolong the risk by continued interaction with the perpetrator, and the risk perpetrator might misuse the warning to bypass the nudge and risk detection systems.

> *"The one where it's like highlighted in red, like what part of like, the text message was bad like, I feel like the creepy person could just like change the wording to bypass that system."*- P9. 17, M. (Information Breach, Sensitivity filter)

For addressing these challenges, the teens made several recommendations as methods of **preventing the risk from reaching the victim**. Many (N=17, 81%) teens recommended features based on restricting the risk sender, which means to prevent the perpetrator to send risky content to a possible victim. Another group of teens (N=17, 81%) recommended using reprimanding language for the risk perpetrator to emphasize the consequences of their risky decision as a way of compelling them to not engage in risky behavior. This is done in the Sensitivity Filter perpetrator nudges, and their preferences are reflected in the rankings of RQ1c. Finally, like the quote from P9 below, some teens (N=5, 24%) also recommended the implementation of this to a greater degree by acting on the reprimanding language, thereby penalizing the perpetrator for their harmful actions. They recommended temporary punishments (account flagging, suspensions) and permanent punishments (bans) based on single and multiple infractions, whereby the priority is protecting the victim, but the perpetrator's punishments are made flexible enough to be appropriate and fair to the person. For example, a repentant offender might be punished with an opportunity to learn and improve their online behavior, while an unrepentant adversarial offender who does not plan to use the platform positively would face a stricter punishment. For example, P9 recommended a strike system for incrementally penalizing the perpetrator, leading to a complete ban in the event of repeated violations.

> *"The nudge could get them to reconsider by mentioning 'if you send the risky messages, you could risk getting banned from our platform...' Maybe like a strike, and like [in] three strikes, you're banned."*- P9. 17, M. (Information Breach, Sensitivity filter)

In summary, teens are aware of the negative impacts that a nudge may have such as censorship increasing interest in the risk and suggested responses possibly prolonging or escalating a risky interaction. As a result, they recommended strict considerations to identify and deemphasize nudge features that have a possibility of worsening a risk instance, as well as implementing risk penalties and prevention as measures to promote accountability of the perpetrator.

5 DISCUSSION

In this section, we discuss the implications of our results, provide comparisons with prior literature, and direction for future work to contribute to the field of adolescent online safety nudging.

5.1 Nudging for Risk Prevention for the Perpetrator and Protection for the Victim

Our results provide interesting insights into the types of nudges teens preferred for different scenarios or perspectives. While teens preferred Sensitivity Filters overall, we uncovered differences in how teens want online safety interventions to address the risk victims and perpetrators. For example, RQ1 showed that teens preferred interventions that protect the risk victim either via censorship (Sensitivity Filter), or guide them through actionable prompts (blocking and reporting with the Guided Actions nudge). While there have been contrasting perceptions in regard to general censorship and content moderation online [51], there is a lack of understanding regarding the implications of moderation for teens' online safety. In their analysis of content policy documents across social media platforms, Pater et al. [41] found that most do not have well-defined content guidelines or consequences for violation, particularly for teens. Our results (Section 4.1) show teens consider a certain amount of censorship and protection from risks to be necessary for their online well-being, as long as it is not enforced and comes with control in the hands of the teens. Other related work from Gibson [21], who analyzed content from a public forum, found content moderation as an effective tool for creating a safer online environment for users. Our work contributes to prior literature by illustrating how teens prefer nudges that protect them by censoring the risk to give them control over the type of content and users that are moderated, along with the ability to dismiss the moderation.

In contrast, most teens wanted more aggressive and preventative nudges to enforce compliance from the risk perpetrator (i.e., reprimanding language and penalties). Prior work has mostly proposed online risk prevention by putting the responsibility on parents through parental controls and restrictions [17], or the risk victims themselves by limiting their personal disclosures online [18, 44]. Recently, Agha et al. [2] found that teens want to put the responsibility of handling online risks on the perpetrator, motivating our risk perpetrator nudge designs. Our findings in Section 4.3.3 further confirm teens' preference to move away from a victim-based approach by increasing accountability for those that perpetuate harm online through strict and corrective measures.

It is also worth mentioning that the online risks studied by Masaki et al. [35] and most other works have looked at online safety nudges from a victim-only lens. In many cases, the teen can be both victim and perpetrator simultaneously (i.e., a cyber-double role) in an online safety risk instance[57]. In our work, studying both perspectives allowed us to get a better understanding of how these nudges may be implemented in the future for the victim and perpetrator in isolation. While we were able to get insights into how teens think from the perspective of a perpetrator and victim, further investigation is needed to understand nudging a teen in a cyber-double role. To ensure adolescent online safety, we recommend that social media platforms strike a balance between protection and prevention, with greater responsibility for the risk perpetrator rather than solely the teen victims or their parents. In summary, nudges can only be effective if they significantly make a difference in teens' online experiences by offering "new" or relevant information, going beyond general warnings, protecting the victim from harm, recommending safe actions, and most importantly, preventing harm when possible by enforcing online safety guidelines for good digital citizenship.

5.2 Overcoming the Challenges of Effective Nudging for Positive Behavioral Change

We uncovered several challenges with the effective implementation of nudges, such as the presented nudges being easy to ignore, ill-timed, or possibly escalating the risk. The most prominent of these challenges among teens was that they found nudges to be easy to ignore, especially if they do not communicate the risk effectively. To address this, teens recommended stronger language, emphasis on harm via design, and penalties for perpetrators when necessary. While it may seem obvious to

make nudges stronger so they are less likely to be ignored, there is a conundrum on how strongly users should be nudged, as highlighted by Chapman et al. [15]. When nudged too far, it could lead to a negative reaction or indifference towards nudges as found in 4.3.1 with the participants' negative experiences with experiencing warning fatigue. In this regard, Petrykina et al. [43] studied the concept of "warning blindness," and found that providing a clear positive or negative incentive to users (e.g., a point-based reward system), is effective in dealing with warning fatigue without affecting convenience. While their work focused on nudges for malware risks for all users, we extend prior recommendations to be specific to risks youth face in their interpersonal interactions and identified penalties and reprimands as disincentives for effectively nudging a risk perpetrator.

The timing of nudges is another challenge, where many teens were concerned that nudges may appear too late, which aligns with findings from prior work [29, 45]. If a nudge is triggered too early or in the wrong context, it could be considered disruptive and the nudge may be ignored as a false positive. Whereas nudging too late may lead to unwanted risk exposure and harm to the victim. Purohit and Holzer [45] emphasized the importance of identifying "teaching moments" which can serve as the optimal timing for nudges to influence behavior change. Our findings contribute to prior work by illustrating that the timing of nudges should be optimized based on the type of risk and category of user that the risk is coming from, while providing flexibility and control over the triggers for nudges. For instance, teens wanted nudges to be triggered earlier for Cyberbullying risks so that the perpetrator does not get a chance to prolong the interaction with or groom the victim.

The final challenge identified with nudges is the possible escalation of the risk they were designed to address. For example, teens thought that Sensitivity Filters may increase curiosity about the risk due to censorship, or Guided Actions could prolong risky interaction with the perpetrator. Recently, OpenWeb [24] shows an implementation of a real-time nudge intervention which recommended replacing harmful words (similar to Guided Actions), and they found them to be effective in positively changing a perpetrator's behavior. In contrast, teens in our study considered the Guided Actions to be more suitable for victims and rarely for the risk perpetrators as they considered most perpetrators to not be positively influenced by a nudge if they had a strong intent to harm. While Guided Actions may work for some cases, we extend prior work by presenting critical challenges by nudging the risk perpetrators toward behavior change. Our findings highlight that not all risk perpetrators have similar intentions or purposes for perpetuating risks, where some may be intentional in their harmful behaviors, others may be unaware or make a mistake. Therefore, risk perpetration cannot be treated as a generalizable process, and we recommend nudging perpetrators in an incremental way, with guided actions, followed by warnings and penalties, and finally leading to risk prevention to ensure that nudges do not escalate the risk and users have the chance to change their behavior before being penalized. While we identify unique challenges for effective nudging, we recognize that our work is based on teens' self-reported feedback and does not provide an experimental evaluation of nudges. Therefore, in our future work, we aim to redesign nudges to address teens' feedback and evaluate them in a realistic simulated setting to understand how teens would respond to nudges in practice, rather than their self-reported behaviors.

5.3 Tailoring Nudges to Provide Autonomy and Personalization

Our results show teens prefer different nudges under different contexts, such as the type and severity of the risk (4.2.2), the age (4.2.3), awareness of the user, as well as the relationship with the perpetrator (4.2.4). In line with our findings, we emphasize the need for tailored nudges that can cater to their particular needs and preferences. Prior work from Peer et al. [42] identified similar individual differences as determinants of the online safety decision process and recommended

customizing nudges for effectiveness. Aligning with their work, we suggest that there is no perfect or "one-size-fits-all" nudging approach for all users, emphasizing the need for customized nudges.

One way of personalizing nudges is based on the type and severity of the risk. When risks are detected and nudges are triggered, platforms can record a history of their response to the nudge to develop an understanding of how harmful teens consider a risk to be and nudge more or less in the future based on the response. Such algorithmic training for online safety nudges can significantly address the issue of disruptive nudges and can help cater to the unsafe experiences that are personally relevant to a teen. Yet, past research has shown that relying solely on automated decisions for online safety may not be the most effective approach [46]. Findings from our work also suggest that teens want a combination of automated tailoring for nudges, along with manual control over the characteristics of nudges they receive (4.3.2). This is supported by insights from Karlsen and Andersen [29] who suggested nudging based on a user's past behavior, along with their stated individual requirements and community efforts. Therefore, we recommend supplementing algorithms with teens' explicitly specified preferences, through online safety settings which ask them to specify online experiences they find risky, along with assessing risk severity based on their preferences and maturity. Relatedly, our results show that age and maturity play a main role in determining the right nudges for teens (4.2.3). Prior work from Lwin et al. [34] conforms with this finding, where they identified the age of teens as a major factor that indicates behavioral differences among adolescents in the context of online safety. Staksrud and Livingstone [52] also identified the age group and gender of the younger population as factors that affect how they interact with their online safety. These differences should be reflected in the design of online safety nudges, where older teens may have more control in the nudges they receive. Similarly, awareness and education regarding online safety also impact the effectiveness of nudges, where teens in our study considered that those unaware of risks are more likely to be influenced by a nudge. Due to their developmental stage, adolescent online risks are a learning experience for teens, and as a result, the nudges should reflect that. The teens' feedback (4.3.2) indicates that being able to learn from an online safety intervention with the ability to turn it off when they have learned how to manage the risk independently may also help solve warning fatigue.

Another aspect of tailoring nudges to meet teen needs is related to the relationship with the perpetrator. Many teens wanted to restrict nudges to strangers only, as they did not want their casual interactions with their friends to be disrupted as falsely detected risks. Thus, providing teens with the autonomy to control the users with whom they can receive online safety nudges will greatly solve issue of timing and disruption. To summarize, making tailored nudges is important because teens are unique in their personalities and priorities. The differences in their nudge preferences and feedback in our sessions (4.1.2, 4.1.3, 4.2) are indicative of the teens having different levels of personal discretion, such as, picking convenience over security in their online safety. The CIA triad concept in cybersecurity explains the security of a system is traditionally inversely proportional to how convenient it is, which may lead to trade-offs in one aspect to ensure the other (i.e., a more convenient to use security system might be less secure) [11]. Tailored nudging can function as an effective middle-ground as it allows future researchers and designers to prioritize teens' online safety as a vulnerable population while balancing autonomy and convenience according to their needs. To summarize, we present a conceptualized set of guidelines for designing effective adolescent online safety nudges:

5.3.1 ***Provide actionable responses beyond warnings:*** Our findings (4.1.1) showed that nudges should not only provide information or warnings about risks, but should also provide a clear, timely, and effective action to address online risks. We expand actionable nudges from prior work [4, 35], by distinguishing that Guided Actions such as block or report should be provided for

lower and impersonal risks (e.g., Information Breaches), whereas risk prevention through Sensitivity Filters is more suitable for severe and personal risks (e.g., Cyberbullying) that can have a lasting impact on teens.

5.3.2 ***Effectively communicate the need for and severity of the nudge:*** Based on our findings (4.3.1), nudges need to be more convincing as teens wanted nudges to establish why a risk instance should be taken seriously, so nudges are not misunderstood, ignored, or considered disruptive. For example, for risk victims, harm from risk needs to be communicated clearly through colour/visuals, whereas perpetrators should be prompted with follow-up nudges and warning signs to convey how their behavior is harmful to others.

5.3.3 ***Provide incentives for positive behaviors and disincentives for negative behaviors:*** Our findings (4.3.3) imply for using reward systems for behavior change, such as positive incentives for victims and accountability for perpetrators through risk preventions, penalties, or temporary banning. Previously, such incentives have mostly been proposed either in gamified approaches for cyberbullying victims [16] or for individual privacy [43]. We extend prior work to propose incentives for both victims and perpetrators that can help address interpersonal risks.

5.3.4 ***Tailor nudges based on both risk and individual differences:*** Our findings (4.2) explain that the effectiveness of a nudge does not apply to all risks, with emerging differences based on risk medium, risk severity/type, teen maturity, and risk awareness. Therefore, we recommend tailoring nudges based on the risk - where nudges should include stronger warnings and preventative measures such as sensitivity filters for risks higher in severity (e.g., explicit images, cyberbullying). We also recommend tailoring nudges to the teens' individual differences, such as customizable control [2] and safety features for early vs. late teens. Furthermore, we recommend allowing updating preferences over time as the teens build resilience and self-regulation, which can also address warning fatigue.

5.3.5 ***Overcome warning fatigue and mistrust of nudges:*** In section 4.3.1, we found that many teens have had negative experiences with nudges feeling disruptive and intrusive, leading to "trust issues". Despite the ethical challenges that come with nudges [30], our findings showed that teens approved of nudging as a promising online safety solution as long as they provided control and personalization (4.3.1). Therefore, we recommend personalizing the timing and triggers of nudges for risk victims [45] (e.g., nudging for "stranger danger" only [9]), for nudges to be catered to teen preferences. For risk perpetrators, we recommend exploring the use of multiple variants for the same nudge to prevent warning fatigue, coupled with accountability measures to influence positive behavior change.

5.4 Limitations and Future Work

One major limitation of our study is that both online risk and proposed nudges were considered to be mutually exclusive when in reality, teens may experience a combination of multiple risk scenarios in a single risk instance and different elements from multiple nudges can also be combined into one nudge. Additionally, this project covers how nudges can be used to handle common online risks faced by teens. However, even though they are based on what was provided to us by the teens, the risky scenarios used are not a recreation of real risk. Another limitation of our work is that the participants were based in the United States and had access to remote-conferencing technology, and as a result, our results may not be generalizable to all teens across the world or in all socio-economic classes. Our enrollment process was on a first-come, first-served basis and the recruitment materials went to organizations that were inclusive of the age range specified in the eligibility criteria. However, most of our participants were either 16 years old or older teens, while

only four participants were below the age of 15 years, this indicates a greater willingness of older teens to participate in a research study but means the results of this study might be less applicable to younger teens. The group sessions of this study were also prone to social desirability bias or groupthink where the participants were reluctant to disagree with one another, which possibly induced similarities in their feedback or rankings. The groupthink concern as well as scheduling conflicts led us to prioritize single-participant over group sessions. Finally, the in-depth, qualitative method used to gather the teens' feedback was time-consuming and logistically restrictive in that it limited the number of nudges that could be evaluated, especially when compared to a broader method such as an asynchronous survey poll.

For future work, we plan to implement these nudge designs and risk scenarios in a realistic setting to have the teens evaluate them in practice. We also recommend future researchers to explore the option of using a virtual assistant or live demonstration as a method of presenting the designed-based study materials to the participants to evoke a greater degree of realism. Finally, we recommend future social moderation work also include the perspective of the risk perpetrator in online safety instances to provide wider context on the overarching issue.

6 CONCLUSION

Through a series of three focus groups and twelve interviews with 21 teens, our work makes a novel contribution to adolescent online safety by identifying key challenges and opportunities for designing nudges that can be effective in positive behavior change for youth online well-being. Overall, we recommend nudges to provide more specific warnings that are contextualized to the risk, and actionable next steps for safety with an emphasis on visual cues. Importantly, our findings highlight the need to move beyond a one-size-fits-all approach towards nudging to cater to teens' varying levels of maturity, unique needs, and experiences while empowering them to self-regulate their online safety. Additionally, we found that the teens' most preferred nudges vary based on the type of risk and the type of user being nudged. While teens wanted guidance for risk victims, they preferred censoring and prevention for the risk perpetrators to stop online risks at the root cause. Considering teens as a vulnerable population, we call for a paradigm shift from traditional soft paternalistic interventions, towards measures leading to risk prevention as part of nudges to ensure accountability and good digital citizenship.

ACKNOWLEDGMENTS

This research was supported by the William T. Grant Foundation (#187941) and National Science Foundation under grants IIS-2333207. Any opinion, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our sponsor. We would also like to appreciate every teen that participated in this study.

REFERENCES

[1] Patricia Agatston, Robin Kowalski, and Susan Limber. 2012. Youth views on cyberbullying. *Cyberbullying prevention and response: Expert perspectives* (2012), 57–71. https://vdoc.pub/download/cyberbullying-prevention-and-response-expert-perspectives-44p196lqq9m0

[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. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1 (April 2023), 1–32. https://doi.org/10.1145/3579625

[3] 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*. ACM, New Orleans LA USA, 1–8. https://doi.org/10.1145/3491101.3503563

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

[5] Zahra Ashktorab and Jessica Vitak. 2016. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 3895–3905.

[6] Karla Badillo-Urquiola, Zainab Agha, Mamtaj Akter, and Pamela Wisniewski. 2020. Towards Assets-based Approaches for Adolescent Online Safety. In *Badillo-Urquiola, Agha, Z., Akter, K., Wisniewski, P.,(2020)"Towards Assets-Based Approaches for Adolescent Online Safety" Extended Abstract presented at the ACM Conference on Computer-Supported Cooperative Work Workshop on Operationalizing an Assets-Based Design of Technology,(CSCW 2020)*.

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

[8] Karla Badillo-Urquiola, Zachary Shea, Zainab Agha, Irina Lediaeva, and Pamela Wisniewski. 2021. Conducting Risky Research with Teens: Co-designing for the Ethical Treatment and Protection of Adolescents. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (Jan. 2021), 1–46. https://doi.org/10.1145/3432930

[9] 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*. ACM, Boise ID USA, 394–406. https://doi.org/10.1145/3311927.3323133

[10] Adrien Barton and Till Grüne-Yanoff. 2015. From Libertarian Paternalism to Nudging—and Beyond. *Review of Philosophy and Psychology* 6, 3 (Sept. 2015), 341–359. https://doi.org/10.1007/s13164-015-0268-x

[11] Artur Baybulatov and Vitaly Promyslov. 2022. On the Availability Property and Its Metric for NPP IACS. In *2022 15th International Conference Management of large-scale system development (MLSD)*. 1–5. https://doi.org/10.1109/MLSD55143.2022.9934157

[12] Leanne Bowler, Cory Knobel, and Eleanor Mattern. 2015. From cyberbullying to well-being: A narrative-based participatory approach to values-oriented design for social media. *Journal of the Association for Information Science and Technology* 66, 6 (2015), 1274–1293.

[13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[14] Brian Saunders. 2023. New parental control features come to Instagram, Facebook and Messenger. https://www.phillyvoice.com/meta-facebook-instagram-messanger-parental-control-tools/

[15] Audrey R. Chapman. 2019. When Going Beyond Gentle Nudges Is Legitimate. *The American Journal of Bioethics* 19, 5 (May 2019), 68–69. https://doi.org/10.1080/15265161.2019.1588416

[16] Ann DeSmet, Katrien Van Cleemput, Sara Bastiaensens, Karolien Poels, Heidi Vandebosch, Steven Malliet, Maïté Verloigne, Griet Vanwolleghem, Lieze Mertens, Greet Cardon, et al. 2016. Bridging behavior science and gaming theory: Using the Intervention Mapping Protocol to design a serious game against cyberbullying. *Computers in Human behavior* 56 (2016), 337–351.

[17] Caitlin Elsaesser, Beth Russell, Christine McCauley Ohannessian, and Desmond Patton. 2017. Parenting in a digital age: A review of parents' role in preventing adolescent cyberbullying. *Aggression and Violent Behavior* 35 (July 2017), 62–72. https://doi.org/10.1016/j.avb.2017.06.004

[18] Yang Feng and Wenjing Xie. 2014. Teens' concern for privacy when using social networking sites: An analysis of socialization agents and relationships with privacy-protecting behaviors. *Computers in Human Behavior* 33 (April 2014), 153–162. https://doi.org/10.1016/j.chb.2014.01.009

[19] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173698

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

[21] Anna Gibson. 2019. Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces. *Social Media + Society* 5, 1 (Jan. 2019), 2056305119832588. https://doi.org/10.1177/2056305119832588

[22] Gary W. Giumetti and Robin M. Kowalski. 2022. Cyberbullying via social media and well-being. *Current Opinion in Psychology* 45 (June 2022), 101314. https://doi.org/10.1016/j.copsyc.2022.101314

[23] Rebecca E. Grinter and Leysia Palen. 2002. Instant messaging in teen life. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work (CSCW '02)*. Association for Computing Machinery, New York, NY, USA, 21–30. https://doi.org/10.1145/587078.587082

[24] Guy Simon. 2020. OpenWeb tests the impact of "nudges" in online discussions. https://www.openweb.com/blog/openweb-improves-community-health-with-real-time-feedback-powered-by-jigsaws-perspective-api

[25] Katrin Hartwig and Christian Reuter. 2021. Nudge or Restraint: How do People Assess Nudging in Cybersecurity - A Representative Study in Germany. In *European Symposium on Usable Security 2021*. ACM, Karlsruhe Germany, 141–150. https://doi.org/10.1145/3481357.3481514

[26] 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.

[27] Haiyan Jia, Pamela J. Wisniewski, Heng Xu, Mary Beth Rosson, and John M. Carroll. 2015. Risk-taking as a Learning Process for Shaping Teen's Online Information Privacy Behaviors. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 583–599. https://doi.org/10.1145/2675133.2675287

[28] Katarina Jurczyk, Christopher You, Mahsan Nourani, Manas Gupta, Lisa Anthony, and Benjamin Lok. 2021. Romadoro: Leveraging Nudge Techniques to Encourage Break-Taking. In *Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*. ACM, Virtual Event USA, 66–69. https://doi.org/10.1145/3474349.3480231

[29] Randi Karlsen and Anders Andersen. 2019. Recommendations with a Nudge. *Technologies* 7, 2 (June 2019), 45. https://doi.org/10.3390/technologies7020045 Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[30] Kaspersky. 2022. What is the "ransomware detected" pop-up? https://usa.kaspersky.com/resource-center/threats/identify-and-remove-fake-pop-ups Section: Resource Center.

[31] Matthew Katsaros, Kathy Yang, and Lauren Fratamico. 2021. Reconsidering Tweets: Intervening During Tweet Creation Decreases Offensive Content. http://arxiv.org/abs/2112.00773 arXiv:2112.00773 [cs].

[32] Minsam Ko, Seungwoo Choi, Subin Yang, Joonwon Lee, and Uichin Lee. 2015. FamiLync: facilitating participatory parental mediation of adolescents' smartphone use. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 867–878.

[33] Priya C. Kumar, Fiona O'Connell, Lucy Li, Virginia L. Byrne, Marshini Chetty, Tamara L. Clegg, and Jessica Vitak. 2023. Understanding Research Related to Designing for Children's Privacy and Security: A Document Analysis. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. ACM, Chicago IL USA, 335–354. https://doi.org/10.1145/3585088.3589375

[34] May O. Lwin, Benjamin Li, and Rebecca P. Ang. 2012. Stop bugging me: An examination of adolescents' protection behavior against online harassment. *Journal of Adolescence* 35, 1 (Feb. 2012), 31–41. https://doi.org/10.1016/j.adolescence.2011.06.007

[35] 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*. ACM, Honolulu HI USA, 1–11. https://doi.org/10.1145/3313831.3376666

[36] Bridget Christine McHugh, Pamela Wisniewski, Mary Beth Rosson, and John M Carroll. 2018. When social media traumatizes teens: The roles of online risk exposure, coping, and post-traumatic stress. *Internet Research* (2018).

[37] McKay Deveraux. 2020. The Dangers of Social Media for Teens. https://www.outbacktreatment.com/the-dangers-of-social-media-for-teens/

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

[39] Cristina Mele, Tiziana Russo Spena, Valtteri Kaartemo, and Maria Luisa Marzullo. 2021. Smart nudging: How cognitive technologies enable choice architectures for value co-creation. *Journal of Business Research* 129 (May 2021), 949–960. https://doi.org/10.1016/j.jbusres.2020.09.004

[40] James Nicholson, Lynne M Coventry, and Pam Briggs. 2017. Can we fight social engineering attacks by social means? Assessing social salience as a means to improve phish detection.. In *SOUPS*. 285–298.

[41] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th international conference on supporting group work*. 369–374.

[42] Eyal Peer, Serge Egelman, Marian Harbach, Nathan Malkin, Arunesh Mathur, and Alisa Frik. 2020. Nudge me right: Personalizing online security nudges to people's decision-making styles. *Computers in Human Behavior* 109 (Aug. 2020), 106347. https://doi.org/10.1016/j.chb.2020.106347

[43] Yelena Petrykina, Hadas Schwartz-Chassidim, and Eran Toch. 2021. Nudging users towards online safety using gamified environments. *Computers & Security* 108 (Sept. 2021), 102270. https://doi.org/10.1016/j.cose.2021.102270

[44] 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 (IDC '17)*. Association for Computing Machinery, New York, NY, USA, 352–357. https://doi.org/10.1145/3078072.3079722

[45] Aditya Kumar Purohit and Adrian Holzer. 2019. Functional Digital Nudges: Identifying Optimal Timing for Effective Behavior Change. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312876

[46] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J. Wisniewski. 2021. A Human-Centered Systematic Literature Review of the Computational Approaches for Online Sexual Risk Detection. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 1–38. https://doi.org/10.1145/3479609

[47] Shawna Malvini Redden and Amy K. Way. 2016. 'Adults don't understand': exploring how teens use dialectical frameworks to navigate webs of tensions in online life. *Journal of Applied Communication Research* (Nov. 2016). https://www.tandfonline.com/doi/full/10.1080/00909882.2016.1248465 Publisher: Routledge.

[48] Maheswaran S, Sathesh S, Ajith Kumar P, Hariharan R S, Chandra Sekar P, Ridhish R, and Gomathi R D. 2022. YOLO based Efficient Vigorous Scene Detection And Blurring for Harmful Content Management to Avoid Children's Destruction. In *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*. 1063–1073. https://doi.org/10.1109/ICESC54411.2022.9885640

[49] Sam Cook. 2022. Cyberbullying Statistics and Facts for 2022. https://www.comparitech.com/internet-providers/cyberbullying-statistics/

[50] Katherine Schaeffer. 2019. Most U.S. teens who use cellphones do it to pass time, connect with others, learn new things. https://www.pewresearch.org/fact-tank/2019/08/23/most-u-s-teens-who-use-cellphones-do-it-to-pass-time-connect-with-others-learn-new-things/

[51] Qinlan Shen and Carolyn Rose. 2019. The Discourse of Online Content Moderation: Investigating Polarized User Responses to Changes in Reddit's Quarantine Policy. 58–69. https://doi.org/10.18653/v1/W19-3507

[52] Elisabeth Staksrud and Sonia Livingstone. 2009. Children and Online Risk. *Information, Communication & Society* 12, 3 (April 2009), 364–387. https://doi.org/10.1080/13691180802635455 Publisher: Routledge _eprint: https://doi.org/10.1080/13691180802635455.

[53] Team Snap. 2022. Introducing Family Center on Snapchat. https://values.snap.com/news/introducing-family-center-on-snapchat

[54] Richard H. Thaler and Cass R. Sunstein. 2009. *Nudge: improving decisions about health, wealth, and happiness* (rev. and expanded ed ed.). Penguin Books, New York.

[55] The POOP Master. 2023. Poop Sewer Puzzle. https://soundcloud.com/xxpooplordxx/poop-sewer-puzzle

[56] Twitter Help. 2022. Twitter's sensitive media policy. https://help.twitter.com/en/rules-and-policies/media-policy

[57] Arminda Vale, Filipa Pereira, Mariana Gonçalves, and Marlene Matos. 2018. Cyber-aggression in adolescence and internet parenting styles: A study with victims, perpetrators and victim-perpetrators. *Children and Youth Services Review* 93 (Oct. 2018), 88–99. https://doi.org/10.1016/j.childyouth.2018.06.021

[58] Greg Walsh. 2018. Towards equity and equality in American co-design: a case study. In *Proceedings of the 17th ACM conference on interaction design and children*. 434–440.

[59] Ding Wang, Xuan Shan, Qiying Dong, Yaosheng Shen, and Chunfu Jia. 2021. No Single Silver Bullet: Measuring the Accuracy of Password Strength Meters. (2021). https://www.usenix.org/conference/usenixsecurity23/presentation/wangding-0

[60] Zhiyou Wang and Shan Jiang. 2022. Influence of parental neglect on cyberbullying perpetration: Moderated mediation model of smartphone addiction and self-regulation. *Health & Social Care in the Community* 30, 6 (2022), 2372–2382.

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

[62] 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*. ACM, Portland Oregon USA, 51–69. https://doi.org/10.1145/2998181.2998352

[63] 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.

[64] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, and John M. Carroll. 2017. Parents Just Don't Understand: Why Teens Don't Talk to Parents about Their Online Risk Experiences. In *Proceedings of the 2017 ACM Conference on*

Computer Supported Cooperative Work and Social Computing. ACM, Portland Oregon USA, 523–540. https://doi.org/10.1145/2998181.2998236

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

[66] Michele L. Ybarra, David Finkelhor, Kimberly J. Mitchell, and Janis Wolak. 2009. Associations between blocking, monitoring, and filtering software on the home computer and youth-reported unwanted exposure to sexual material online. *Child Abuse & Neglect* 33, 12 (2009), 857–869. https://doi.org/10.1016/j.chiabu.2008.09.015

A COMPLETE NUDGE SOLUTIONS

A.1 Victim Nudges for Information Breach Risky Scenario

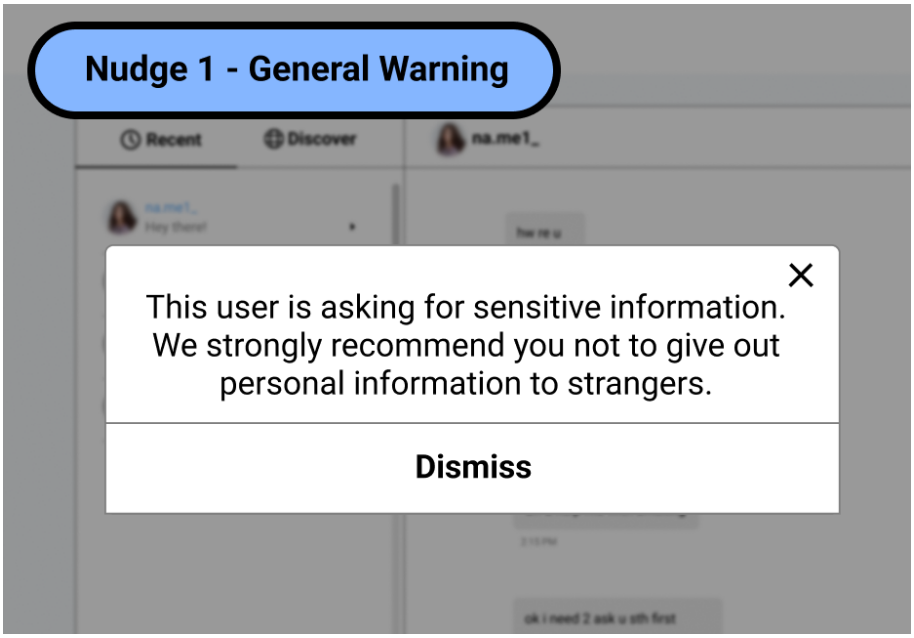


Fig. 8. General Warning Nudge

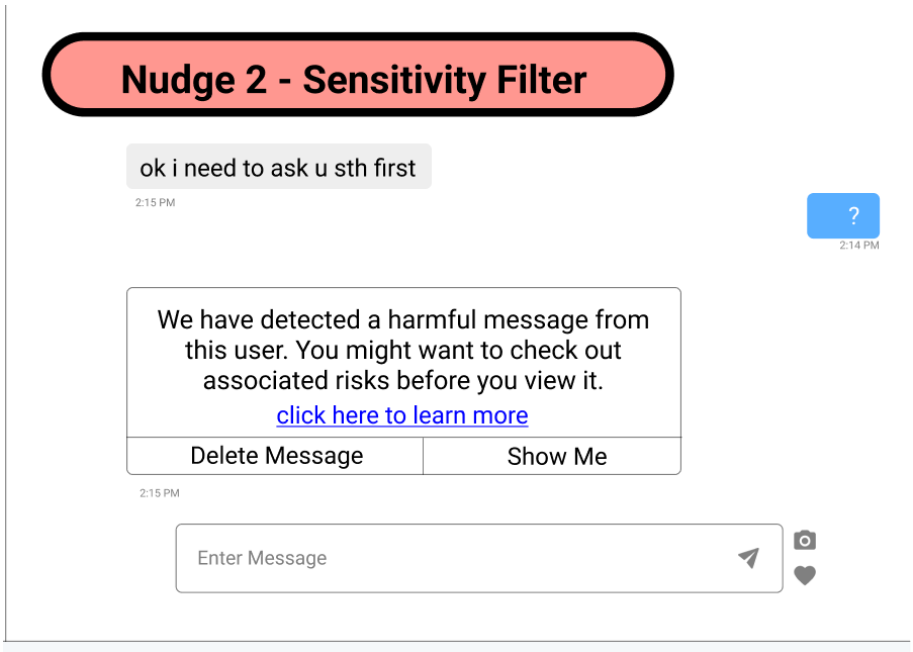


Fig. 9. Sensitivity Filter Nudge

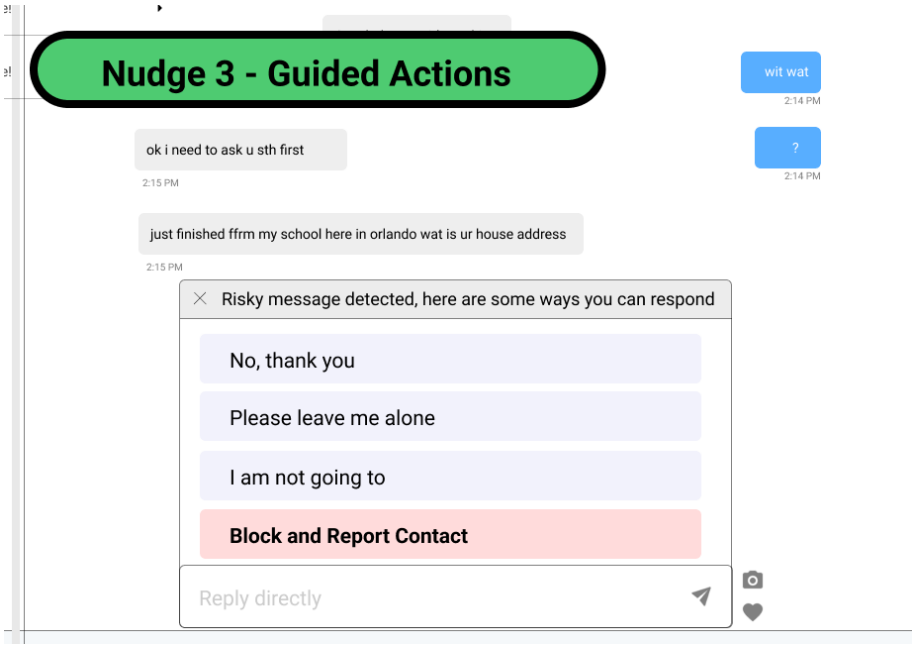


Fig. 10. Guided Action Nudge

A.2 Victim Nudges for Cyberbullying Risky Scenario

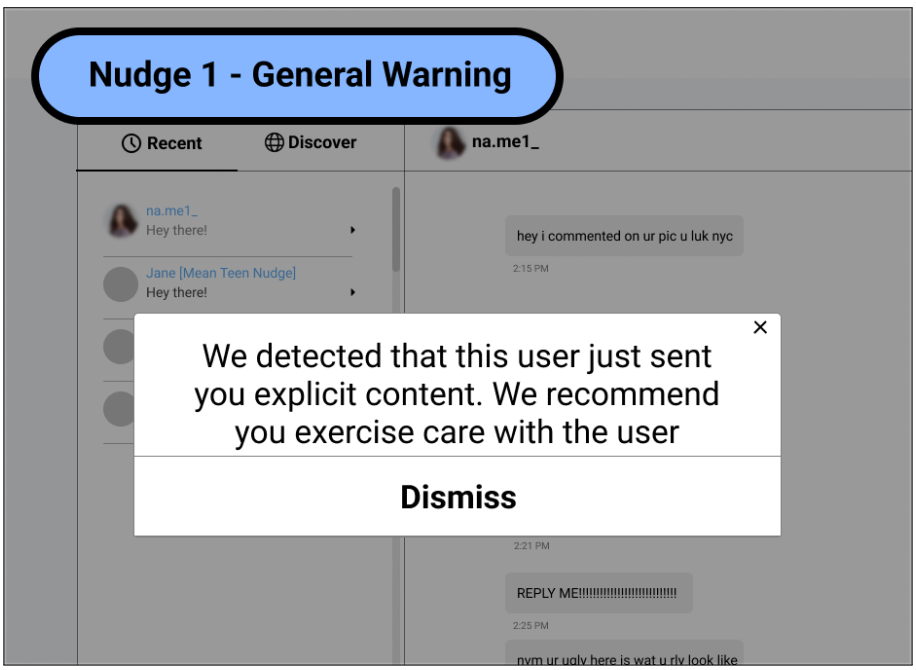


Fig. 11. General Warning Nudge

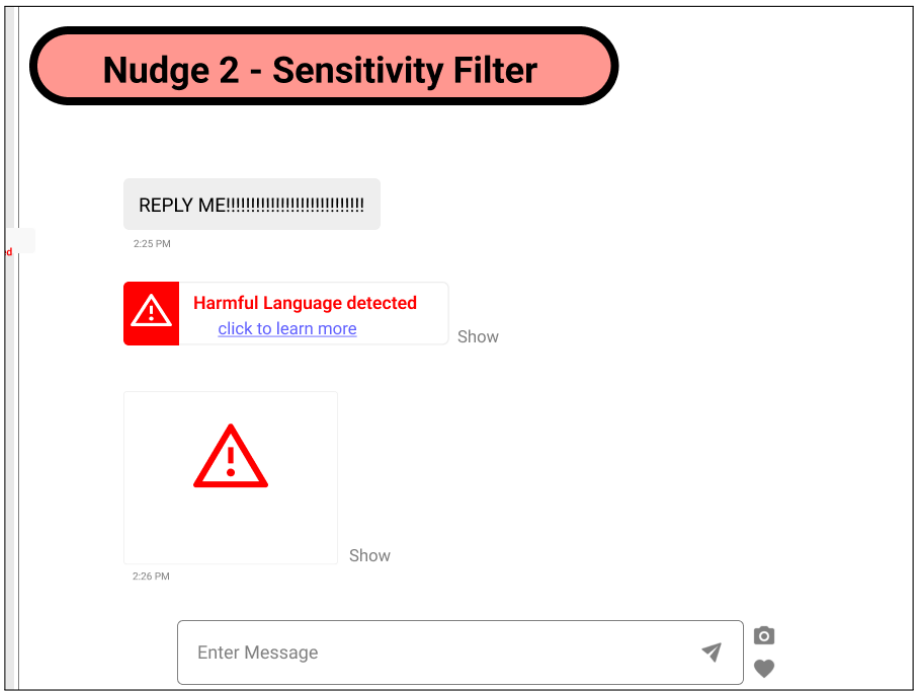


Fig. 12. Sensitivity Filter Nudge

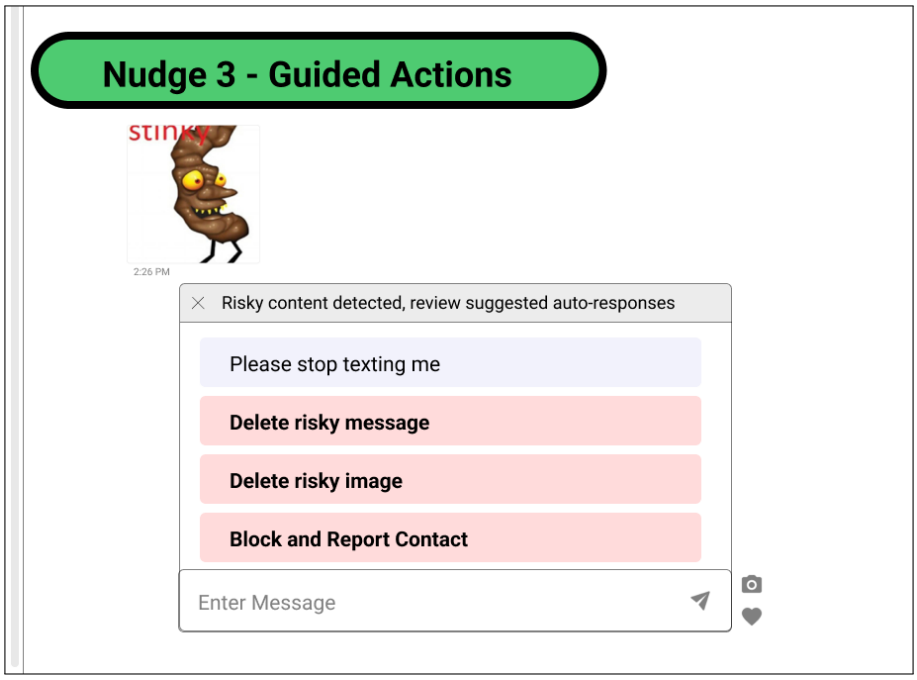


Fig. 13. Guided Action Nudge

A.3 Perpetrator Nudges for Information Breach Risky Scenario

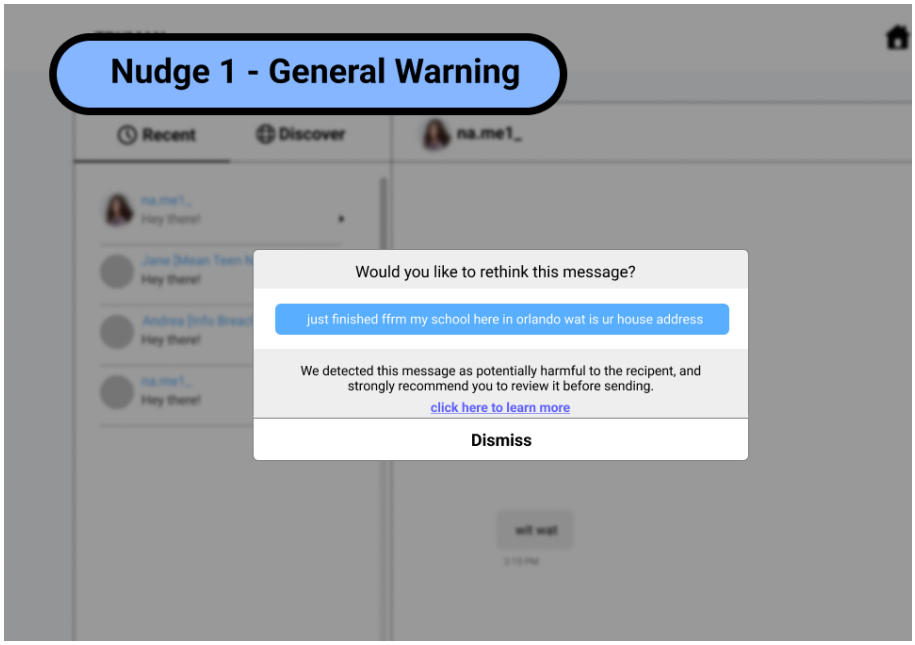


Fig. 14. General Warning Nudge

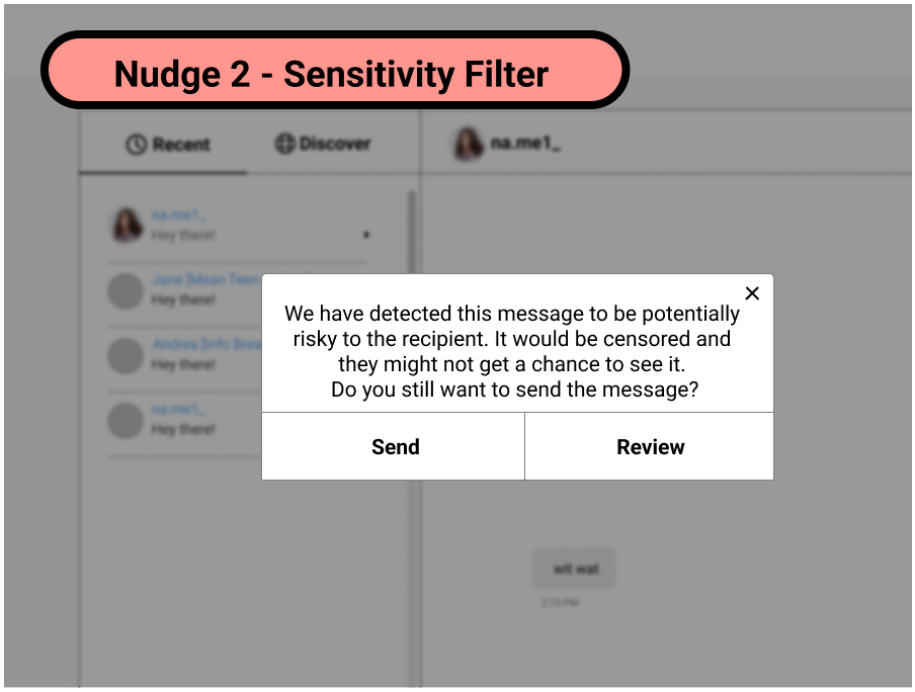


Fig. 15. Sensitivity Filter Nudge

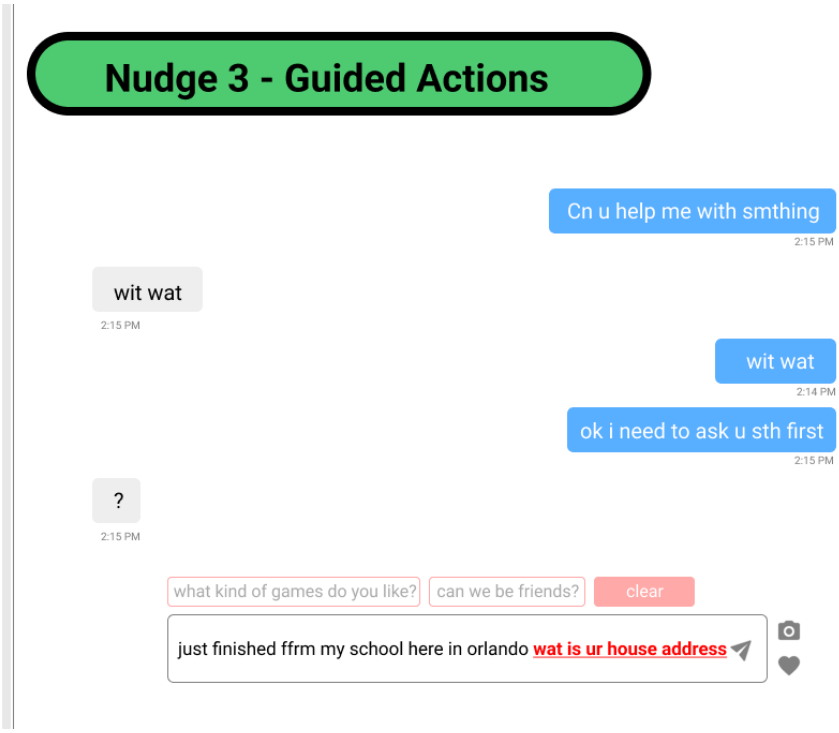


Fig. 16. Guided Action Nudge

A.4 Perpetrator Nudges for Cyberbullying Risky Scenario

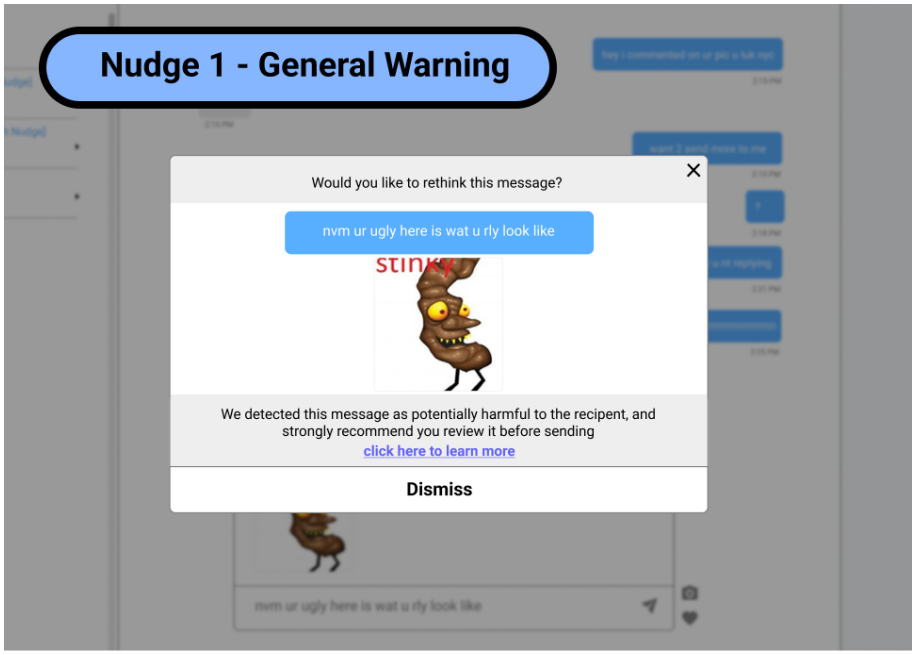


Fig. 17. General Warning Nudge

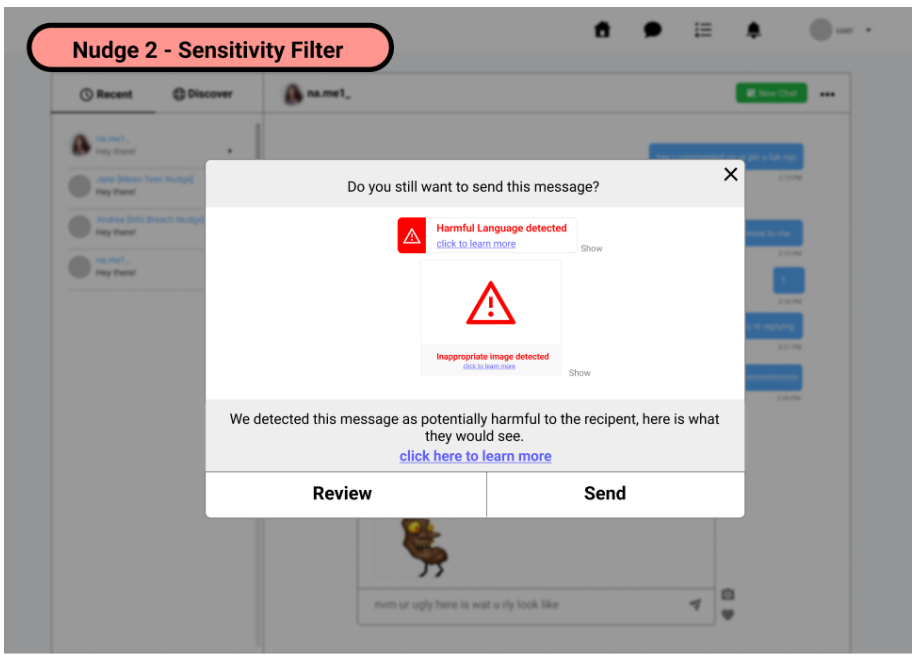


Fig. 18. Sensitivity Filter Nudge

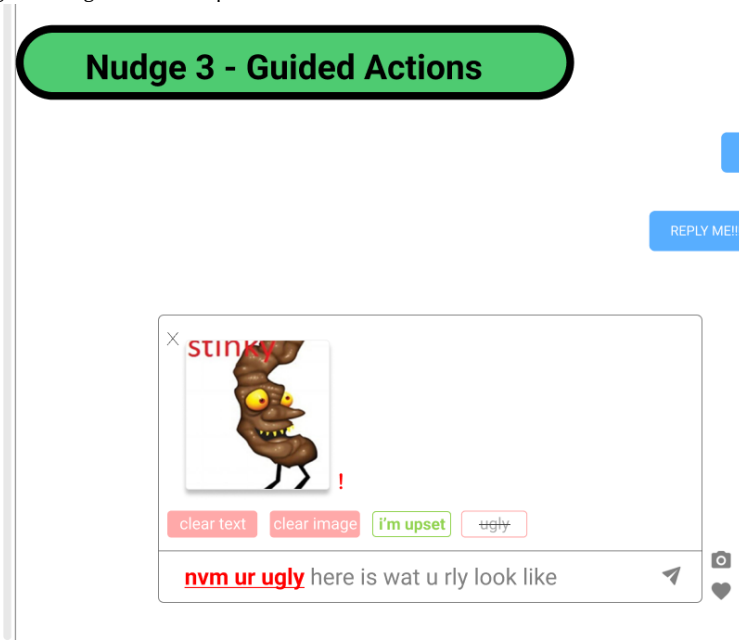


Fig. 19. Guided Action Nudge

Received January 2023; revised July 2023; accepted November 2023