

Training Wheels, Not Guardrails: Supporting Teen Agency Through Real-Time Online Safety Nudges for Social Media

Zainab Agha

Department of Computer Science
San Francisco State University
San Francisco, California, USA
zainabagha@sfsu.edu

Naima Samreen Ali

School of Information
University of Michigan
Ann Arbor, Michigan, USA
naimaali@umich.edu

Prakriti Dumar

Socio-Technical Interaction Research Lab
International Computer Science Institute
Berkeley, California, USA
pdumar@icsi.berkeley.edu

Pamela J. Wisniewski

Socio-Technical Interaction Research Lab
International Computer Science Institute
Berkeley, California, USA
pwisniewski@icsi.berkeley.edu

Abstract

We investigate whether nudges can serve as in-the-moment support for youth online safety through a qualitative Wizard-of-Oz study with 19 adolescents (ages 13–19) using a realistic social media simulation (“SocialSim”). Even before nudges were introduced, participants demonstrated safety competence by interpreting social cues and using protective strategies. When nudges appeared, teens responded differently depending on risk severity and context; they relied on nudge guidance in high-risk, ambiguous situations (e.g., exposure to explicit content), while preferring to exercise their own judgement in lower-risk scenarios (e.g., information requests). We found that nudges were most effective when framed as developmental teaching moments that increased risk awareness and offered actionable advice while preserving autonomy. Participants emphasized proportionality, personalization, and trustworthy risk detection, and cautioned against overly restrictive interventions. Using a Social Information Processing (SIP) framework, this work informs the design of in-the-moment supports that foster adolescents’ decision-making, resilience, and self-regulation online.

CCS Concepts

• Human-centered computing → Empirical studies in HCI.

Keywords

Online Safety; Adolescents; Teens; Nudges; Real-Time; Simulation; Ecological Validity; Social Media

ACM Reference Format:

Zainab Agha, Prakriti Dumar, Naima Samreen Ali, and Pamela J. Wisniewski. 2026. Training Wheels, Not Guardrails: Supporting Teen Agency Through Real-Time Online Safety Nudges for Social Media. In *Proceedings of the 25th Interaction Design and Children Conference (IDC '26)*, June 22–25, 2026, Brighton, United Kingdom. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3773077.3806123>



This work is licensed under a Creative Commons Attribution 4.0 International License. IDC '26, Brighton, United Kingdom

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2283-7/26/06

<https://doi.org/10.1145/3773077.3806123>

1 Introduction

In recent years, the relationship between social media use and adolescent well-being has become the focus of intense public [33], scientific [41], and policy debate [94]. While social media platforms offer adolescents opportunities for connection, self-expression, and identity development [34, 64], they have also been linked to different risks including cyberbullying [10], privacy violations [60, 101], exposure to explicit content [68], and negative mental health outcomes [33]. Consequently, due to the growing concern about youth online safety, policymakers and regulators globally are becoming inclined toward restricting [63] or banning youth access to social media [47, 59]. Several countries have imposed restrictive measures, such as recent age-based social media bans in Australia [59] and requiring parental consent and age verification mechanisms in multiple states within the U.S. [11, 47]. While these efforts are motivated by desire to reduce exposure to harm, they often fail to account for adolescents’ developmental need to learn how to navigate risks and may disproportionately affect marginalized youth [18]. Hence, there is growing interest in alternative resilience-based approaches [69, 96, 102] that move away from restriction and instead emphasize strengthening adolescent autonomy [13, 84]. As such, our work focuses on supporting teens’ agency, which we define as their ability to actively interpret social situations online, develop skills for navigating online risks, and make informed safety decisions for themselves, rather than having parents or platforms make those decisions for them.

One such approach is “Safety by Design” [32, 95], which emphasizes shifting the responsibility for safety from individual users to the sociotechnical systems in which they engage by proactively embedding safety mechanisms into platforms to help young users anticipate and mitigate risks. Within this paradigm, nudges [88] are lightweight interventions that influence behavior without compromising autonomy in decision-making. Prior work showed nudges as a promising avenue for supporting adolescent online safety [21, 60] as they can be designed as in-the-moment support to help them recognize risks, reflect on possible consequences, and practice safer behaviors [6, 83]. Yet, despite the potential promise of nudges, research to date has primarily focused on early stages of nudge design [3, 15],

survey-based feedback using hypothetical scenarios [60, 89] or computational approaches to detect risks for understanding “when” to introduce a nudge [51, 79], with limited inquiry as to how adolescents would respond to such interventions in real-time social media contexts. To address this gap, we present a qualitative, ecologically grounded evaluation of real-time online safety nudges using a Wizard-of-Oz (“WoZ”) [20] social media simulation with teens. Focusing on three common online risks (i.e., cyberbullying, information breaches, and exposure to explicit content), we investigate how adolescents interpret social cues, how nudges shape decision-making in-situ, and how safety can be supported without undermining agency. Through this study, we address the following high-level research questions:

- **RQ1:** *How do teens process social cues and employ safety strategies for building connections with others, while protecting themselves online?*
- **RQ2:** *How do teens respond to real-time safety nudges, and how do these nudges relate to their online safety decisions when interacting with others via social media?*
- **RQ3:** *What preferences, concerns, and design considerations do teens voice regarding the use of safety nudges on social media?*

To address these research questions, we conducted a qualitative WoZ evaluation (i.e., participants interacted with what appeared to be other social media users, but in actuality were engaging with hidden “wizard” researchers) of real-time safety nudges with 19 teens (ages 13-19) over Zoom using a web-based simulated social media platform (i.e., “SocialSim”). Grounded in Social Information Processing (SIP) theory and developmental perspectives on adolescent autonomy and learning [22, 91], this study captured teens’ in-the-moment interpretations of social cues, relational goals, and safety decisions during semi-realistic online interactions. Participants created profiles, engaged with content, and exchanged direct messages with other “users,” making decisions to friend, block, or ignore them, enabling analysis of impression formation, connection building, and self-protective strategies in the absence of intervention (**RQ1**). Risk scenarios then triggered context-sensitive safety nudges, allowing us to examine how teens responded to nudges during risky encounters, including when nudges were acted upon, bypassed, or deferred (**RQ2**). Finally, post-task reflections captured teens’ perspectives on nudge usefulness, trust, and appropriateness, revealing how prior experience, perceived risk severity, and autonomy shaped their evaluations (**RQ3**).

Through our approach and findings, this paper makes three primary contributions: First, we present a novel, first-of-its-kind Wizard-of-Oz evaluation of real-time safety nudges with adolescents in a realistic social media simulation, enabling the study of in-the-moment safety decisions beyond hypothetical or retrospective methods. Second, we provide empirical evidence that teens actively employ self-protective strategies even without intervention (**RQ1**), and that their engagement with nudges varies by risk type and perceived severity (**RQ2**). This highlights the need for proportional, context-sensitive support. Third, we offer design insights grounded in teens’ perspectives that position nudges as scaffolding mechanisms that support safety and learning without undermining autonomy (**RQ3**). Overall, these contributions shift the conversation from whether teens should engage on social media to how

platforms can function as “training wheels,” offering supportive guidance that helps younger teens build agency and judgment as they encounter inevitable online risks, that can be radually reduced as competence develops.

2 Theoretical Framing and Background

Below, we first introduce the theoretical foundations of this work and then synthesize prior empirical research on nudge-based online safety interventions for adolescents.

2.1 Social Information Processing as a Developmentally Appropriate Theoretical Lens

Our study is grounded in Social Information Processing (SIP) theory [24, 91], which explains how people interpret social cues, infer others’ intent, and decide how to respond during social interactions, particularly in computer-mediated communication (CMC) [61, 99], where cues are often limited or ambiguous. SIP frames interaction as an ongoing process with three stages: *impression formation*, where individuals interpret available cues and engage in risk appraisal; *social information exchange*, where they make decisions about selective self-disclosure and privacy management; and *relational development over time*, where interactions evolve through continued communication and boundary management over time. SIP has been applied in relevant IDC contexts, such as in the study of aggressive behavior among children in social virtual reality [37] and more generally related to the use and effects of social media by adolescents [42, 97]. Yet, SIP has rarely been used to examine how adolescents navigate online risk during real-time interactions. This gap is notable because adolescence is marked by increasing autonomy and learning through experience [25, 65], where self-regulation often develops through engaging with uncertain social situations rather than avoiding risk entirely [72]. Educational research on “teachable moments” [46] further suggests that learning is most effective when support is provided during moments of ambiguity or heightened attention. SIP therefore offers a well-suited lens for understanding online safety as an in-the-moment process of interpreting risk cues and making relational decisions, directly informing our Safety by Design approach to online safety nudges. In this regard, nudges function as context-specific support that shape how adolescents decode social cues, reassess intent and evaluate safe responses as part of their ongoing relational decision-making [1, 4, 6]. Grounding our qualitative analyses in this developmental framing of SIP enables us to examine how interventions shaped teens’ decision-making while interacting with new people via social media, i.e., how teens process risk cues, while trying to make new connections (**RQ1**), how they respond to the nudges during risky interactions (**RQ2**) and how in-the-moment support may (or may not) scaffold online safety learning without undermining teen autonomy (**RQ3**).

2.2 The Promising Potential and Existing Gaps of Nudge-based Interventions for Adolescent Online Safety

Recent work on adolescent online safety has shifted away from restriction and surveillance-based approaches toward “Safety by

Design” [26, 29, 32, 95], which emphasizes design patterns that support teen resilience, self-regulation, and decision making. In line with this shift, researchers have begun to explore nudges, or behavioral interventions that provide cues to influence behavior without undermining autonomy [88], as a promising mechanism for supporting adolescents’ online safety (c.f., [3, 5, 6, 21, 27, 28, 60, 66, 86]). For instance, Dumarú et al. [26] worked with parents to re-envision parental controls to emphasize open communication, self-regulation, and granular nudges, rather than restrictive monitoring. Masaki et al. [60] conducted an online survey to examine teens’ preferences for different nudge framings, and found that negative framing was more effective at reducing risky choices than affirmative framing, while social comparison nudges could backfire by inadvertently normalizing risk-taking behavior. Similarly, Obajemu et al. [66] used interviews and focus groups with teens to surface more nuanced perceptions of nudges, showing that teens prefer nudges that move beyond generic warnings toward situational guidance and highlighting the need for personalization based on factors such as age, risk medium, risk awareness, and perceived risk severity. To this end, Agha et al. [3] involved teens in UX design bootcamps to conceptualize risks they encounter on social media and co-design nudge-based interventions to deal with those risks. Subsequently, Agha et al. [6] examined best practices for ecologically and ethically evaluating nudge-based online safety interventions in the context of social media use [5], suggesting that simulation-based approaches would balance realism with safety.

This prior research on adolescent safety nudges has made important contributions by exploring nudge design, eliciting teens’ perspectives through survey-based and hypothetical scenarios, engaging teens in the co-design of such systems, and developing platforms and methods to evaluate the effectiveness of nudge-based interventions for youth. While this foundational work has established the promise of nudges for supporting online safety and informed their design, it has not yet empirically evaluated youth-facing safety nudges in real-time, leaving open questions about how teens engage with these interventions in situ. As suggested by Agha et al. [2], simulation-based methods offer a promising approach for studying in-the-moment online safety behavior in controlled yet socially rich environments and have been successfully used in privacy and security research, including mock social media platforms [103], simulated phishing systems [55], and with children [37]. Building on this prior work, we use a realistic social media simulation (“SocialSim”) to examine how teens respond to real-time safety nudges during authentic interactions.

3 SocialSim: A Social Media Platform for Evaluating Real-Time Safety Nudges with Teens

Studying teens’ online safety behaviors in live social media platforms presents significant ethical and methodological challenges, particularly when interventions must be delivered during ongoing interactions and involve risk [6]. To address this challenge, we utilized “SocialSim” [5], which is an open-source, web-based social media simulation platform designed to support ecologically valid and ethically responsible evaluation of behavioral interventions

with adolescents. Built on and extending Truman [23], an open-source research platform for social media experimentation [87], SocialSim provides a controlled environment that closely mirrors core social media functionality while allowing researchers to study in-the-moment interaction and decision-making without exposing participants to real-world harm.

In our study, we integrated online safety risk scenarios and nudges directly into the SocialSim platform, enabling the delivery and evaluation of real-time interventions during live social media interactions. The key features of the SocialSim platform (as summarized in Figure 1) include:

- **User Profiles and Feed:** Participants create profiles and interact with a prepopulated feed through posting, liking, commenting, and flagging content.
- **Direct Messaging:** A real-time messaging interface that supports private text, image, and reaction-based interactions with online users.
- **Real-Time Safety Interventions:** Nudge interventions are triggered during risky interaction to allow the research to assess in-situ responses.
- **Friending, Blocking, and Reporting:** Participants can manage relationships by friending, blocking, or reporting other users based on perceived safety.
- **Hybrid Automation and Researcher Oversight:** Automated scripts and researcher-mediated interaction ensure realism, consistency, and participant safety.

4 Methods

This section describes the methods used to study adolescents’ responses to real-time safety nudges using “SocialSim”, including the study design, WoZ procedures, data collection and analysis, and participant demographics.

4.1 Study Overview and Tasks

We conducted a qualitative WoZ study with 19 adolescents (ages 13–19) based in the United States via synchronous think-aloud sessions over Zoom. Sessions were conducted individually (i.e., a single teen participant accompanied by multiple researchers) and focused on understanding how each teen navigated online interactions, encountered safety nudges, and made safety decisions about other users in the chat (personas played by the researchers) via SocialSim. Each session lasted approximately two hours. Participants first completed a pre-survey about their online experiences, followed by the researcher moderated Zoom session. At the start of the session, researchers explained the study procedure and introduced the features of SocialSim. Participants then accessed SocialSim and shared their screen throughout the session while thinking aloud as they completed following sequence of tasks designed to mirror common social media behaviors:

- (1) Complete your user profile so other teens can learn more about you.
- (2) Engage with posts from other users (e.g., like, comment) to connect with them.
- (3) Create a post on your profile to introduce yourself to the community.

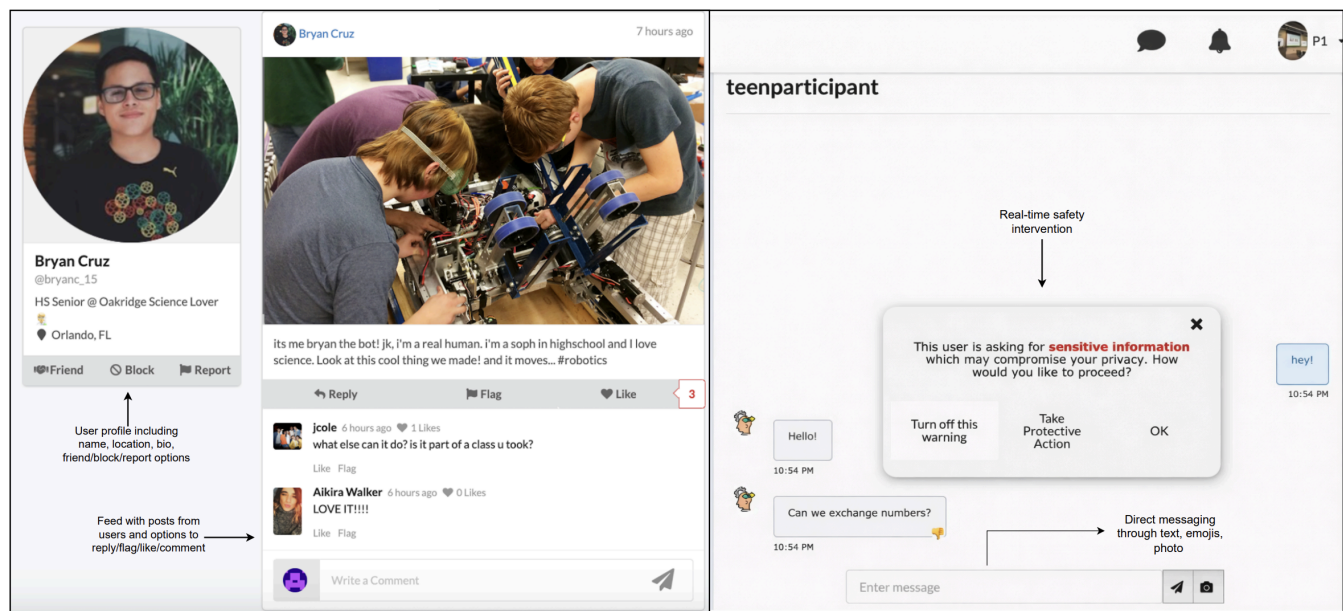


Figure 1: SocialSim Main Features: (a) User profile and feed, (b) Private chat and real-time safety intervention. Here, SocialSim is populated with a risky persona “Bryan,” for our study, along with the safety nudge, as described in more detail in our Methods.

- (4) Interact with all other active users on the chat platform using Direct Messaging. Continue to interact with them until you can decide on their safety.
- (5) After interacting with each user, choose one of the following options for that user: friend, block, or ignore.

After completing all interactions, participants filled out a post-survey and participated in a semi-structured post-task reflection on their experiences and compare safety nudges. Before concluding the session, participants were debriefed about the WoZ nature of the study and the role of the researchers as simulated users. Next, we describe our WoZ risk simulation, including the risk scenarios that teens experienced during the study, along with our safety nudges.

4.2 Wizard-of-Oz Simulation of Online Risks

We employed a WoZ approach using SocialSim to simulate realistic social media interactions while maintaining experimental control and participant safety [40, 81]. In this setup, trained researchers acted as a subset of users on the platform, allowing for responsive, context-aware interactions that would be difficult and unsafe to fully automate. Participants came in with no prior familiarity with the SocialSim platform. They were informed that they were interacting within a simulated environment, but were not told that interactions were researcher-controlled, supporting more naturalistic engagement. Across sessions, participants interacted with a fixed set of six personas: three associated with risk scenarios and three non-risky users that were included to balance the social environment. In addition, the feed contained scripted posts from other consistent, non-interactive users to simulate a broader social context. Researchers followed structured interaction scripts (Table 10 in Appendix) and persona profiles (Figure 2), while allowing limited improvisation in response to participants’ messages for realism.

Each session was conducted by a team of four to five researchers, each logged into a designated user account. To ensure consistency and ecological validity, we conducted pilot sessions with teens to refine risk personas, scenarios, and nudges, standardized interaction sequences where possible, and exposed all participants to the same initial feed content, user personas, and risk triggers. To further mitigate the potential WoZ cofounds related to researcher’s performance, we standardized persona enactment using teen-appropriate consistent language guidelines (e.g., tone, abbreviations, emoji use, etc.), and structured turn-taking rules for response timing (e.g., brief delays and coordinated turns between personas for participant replies), with limited improvisation to adapt responses within the context of the chat to maintain realism while preserving persona intent, risk trajectory and tone.

4.2.1 Risk Types, Risky Personas, and Scenarios. The study focused on three categories of online risk commonly encountered by adolescents on social media: public cyberbullying [90], private information requests [8], and exposure to explicit content [39, 58]. These risks were embedded into the simulation to reflect how they typically emerge in everyday interactions. Informed by pilot studies with teens and by prior research (c.f., [6, 60, 66]), all scenarios were designed to be realistic yet low in severity, enabling meaningful engagement without exposing participants to overt harm. Each risk was enacted through a distinct user persona with a defined profile, posting behavior, and conversational style. Personas were designed to resemble plausible peers or near-peers and to introduce risk incrementally rather than through overtly malicious behavior. Figure 2 illustrates the three risky personas and corresponding risk scenarios. Cyberbullying appeared as a dismissive or backhanded comment in the public-facing feed (not an online user in direct messages), request for sensitive information developed gradually during

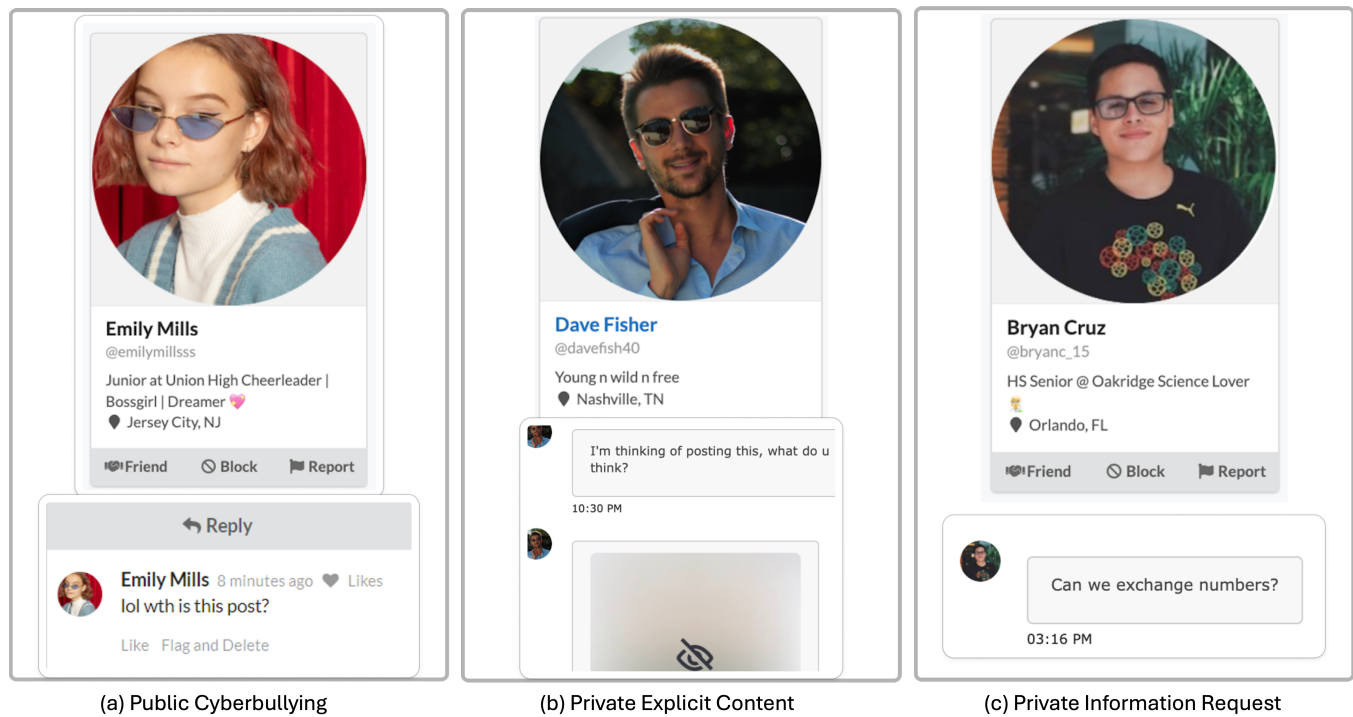


Figure 2: Shown above, we present the three risky personas at the top with their corresponding risk scenario enacted below. Risky Personas and Scenarios by Risk Type: (a) Public Cyberbullying Comment, (b) Private Explicit Content with Sensitivity Filter via Direct Message, and (c) Private Information Request via Direct Message

friendly private exchanges, and explicit content was introduced through direct messages using a blurred image. To maintain a balanced and realistic social environment, three additional non-risky users (Table 9 in Appendix) with varied interests and communication styles were included; these personas did not trigger safety nudges.

4.2.2 Safety Nudges. For the three risky scenarios, safety nudges were embedded directly into the platform and triggered in real-time during interaction. While nudges were triggered by predefined platform rules based on risk scenarios introduced by researchers, most teens interpreted them as automated (e.g., AI-driven) rather than human-mediated or scripted. Nudges were iteratively designed with direct input from teens through pilot co-design studies and informed by prior work (c.f., [6, 60, 66]). Teens’ feedback emphasized the need to clearly signal potential risk, manage exposure to flagged content while preserving autonomy, and provide multiple protective options rather than enforcing a single response. Each nudge alerted participants to the relevant risk and provided a range of options, allowing teens to decide how to proceed. Depending on the scenario, nudges supported teen agency by allowing participants to view or dismiss flagged content, take protective actions (e.g., blocking, reporting, deleting, hiding similar risks, or notifying other users), or continue the interaction without intervention. In the case of information breaching, nudges also provided optional suggested messages to support teens in navigating the situation, based on findings from our pilot work. Figure 3 illustrates how

safety nudges were embedded within the SocialSim interface and surfaced during ongoing interaction. The designs reflect a deliberate balance between *safety* (i.e., by signaling potential risk and offering protective actions) and *autonomy* (i.e., allowed to dismiss, override, or independently manage warnings as part of their own decision-making) [93].

4.3 Data Analysis Approach

We conducted a qualitative analysis of teens’ nudge responses and think-aloud feedback using interview transcripts, Zoom screen recordings, and survey data. For RQ1, we used a hybrid approach informed by Social Information Processing (SIP) theory [24, 91], combining top-down template coding [36] aligned with SIP stages (i.e., *impression formation*, *social information exchange* and *relational development over time*) with open coding of teens’ interactions and interview data to capture emergent strategies through which teens process social cues, self-disclosure, and safety in online interactions for connection and protection, derived from analyzing teens’ end-to-end interactions (e.g., sharing profile information, cross-checking users) alongside their interview reflections. Table 1 summarizes the resulting strategies, where percentages indicate the proportion of participants (N=19) who demonstrated each strategy behaviorally and/or reported it in think-aloud reflections. For RQ2, we conducted a qualitative analysis of teens’ situated behavioral responses to real-time safety nudges, examining how nudges shaped action pathways and relational outcomes during interaction, triangulated

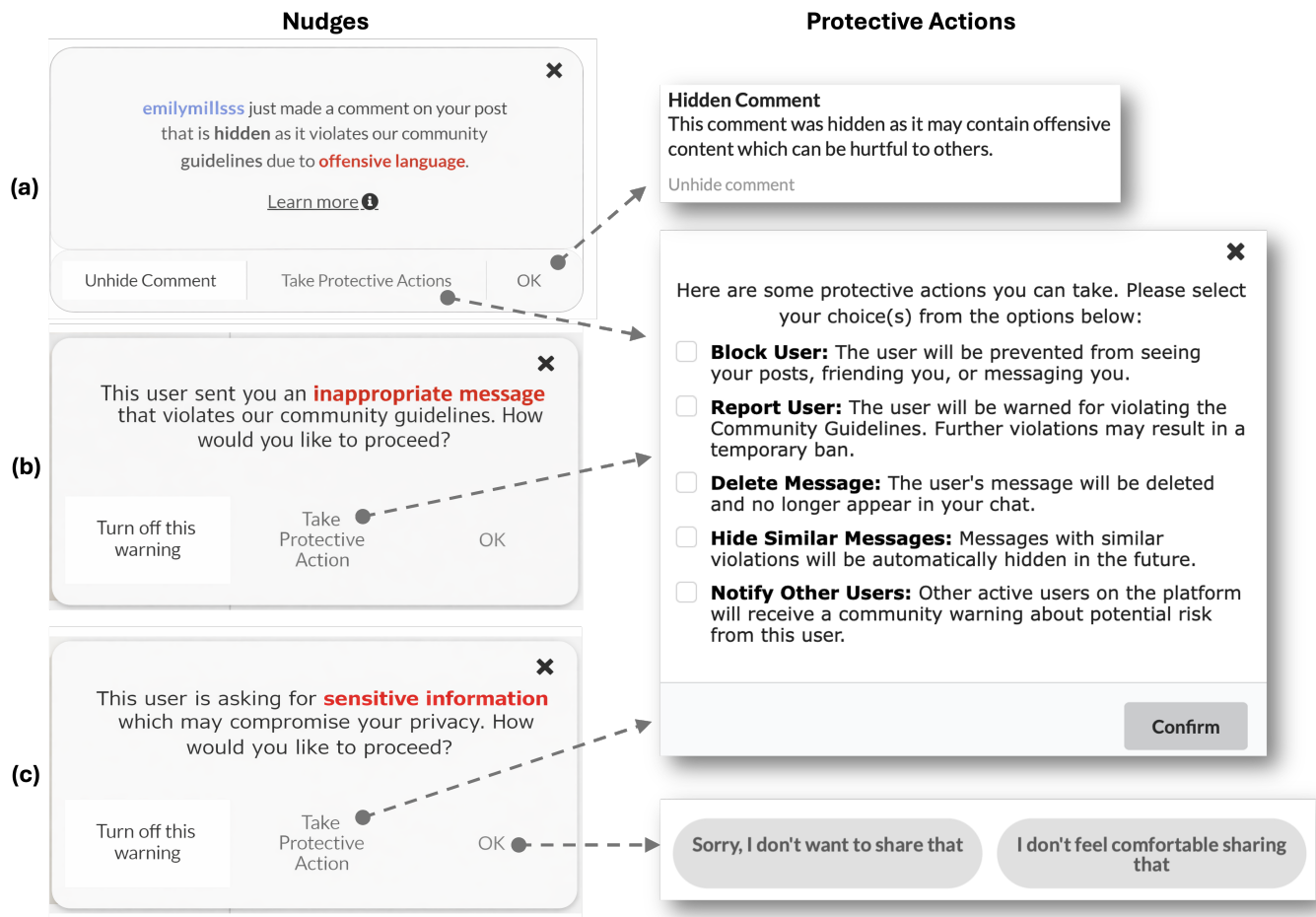


Figure 3: Overview of real-time safety nudges implemented in SocialSim, including (a) Public Cyberbullying, (b) Private Explicit Content and, (c) Private Information Requests. All nudges flagged potential risk during interaction and offer multiple safety choices, with additional options to unhide for (a) and suggested responses for (b).

with think-aloud data [85] to examine how nudges supported risk interpretation (Table 2, Figure 4). For RQ3, we conducted thematic analysis [16] of post-task reflection to identify nudge design preferences. Coding was led by the first and third authors. First, we did open-coding of a subset of the participants' think-aloud and post-interview responses to generate initial codes (e.g., preferences for multiple options, desire for transparency, concerns about AI accuracy) and applied these codes across the dataset, with iterative generation and application of new codes. Codes were refined through regular team discussions and any conflicts were resolved through iterative consensus-building before coding the rest of the data. Finally, the codes were conceptually grouped and categorized into broader themes (e.g., proactive safety and learning) to form our final codebook (Table 8). Given our collaborative, iterative coding process involving multiple rounds of discussion, comparison, and refinement to surface emergent themes rather than enforce agreement, we did not compute inter-rater reliability [62].

The unit of analysis differed across research questions. For RQ1 and RQ3, the unit was the participant, with each code counted once per participant; thus, percentages are out of $N=19$ and sum to at most 100%. For RQ2, the unit is the nudge interaction within each scenario. Participants could take multiple sequential actions within a single nudge flow, so multiple codes could be assigned per participant. Percentages may sum to 100% within the first nudge choice, but not across the full interaction. Table 2 reports aggregated actions, with detailed breakdowns in Tables 5–7 in the Appendix.

4.4 Participant Recruitment and Demographics

Prior to engaging teens, all researchers completed youth protection training, and study procedures were approved by the university's Institutional Review Board (IRB). To ensure the inclusion of perspectives from under-represented and marginalized youth, we intentionally prioritized recruitment from communities that are often underrepresented in youth safety research. Participants were recruited from several sources including existing connections

with minority-serving educational institutions across the U.S., from existing contacts with youth-serving organizations (e.g., boys' and girls' clubs) and our research team's connections (consisting primarily of minority women of color). Recruitment through these channels resulted in a sample in which all participants identified as people of color through self-selection into the study. We contacted organizations via email, phone, social media, and flyer distribution. After pre-screening for eligibility, participants completed consent forms, as well as parental consent forms for those under 18. For completing the study, a \$30 Amazon gift card was given to the youth. Most of the participants were between the ages 15–19 (79%, $N=15$), with a mean age of 16.4. Based on our demographic survey, several participants self-identified as marginalized by responding to an open-ended question asking whether they considered themselves marginalized. Additional survey questions about household structure, family educational background, sexual identity, and race further contextualized participants' experiences and informed our sample (Table 4 in Appendix).

5 Results

In this section, we summarize the main emerging themes for answering our research questions. Table 1, 2, and 8 present the distribution of themes corresponding to RQ1, RQ2 and RQ3 respectively and structure the reporting of our results.

5.1 Decoding Social Cues and Strategies for Online Connection and Protection (RQ1)

Drawing on Social Information Processing (SIP) theory [91], which explains how relationships form in computer-mediated contexts through the gradual interpretation of social cues, we organize our RQ1 findings around SIP's stages of *impression formation*, *social information exchange*, and *relational development over time*. Table 1 summarizes how teens navigated these stages, highlighting the connection and protection strategies they used to leverage social cues and share selectively while managing risk. We present these stages as themes below.

5.1.1 Teens formed impressions of users by seeking common ground, selectively shared information to build trust, and built on reciprocal interactions. SIP [91] emphasizes that, in the absence of rich face-to-face cues, users rely on available textual, visual, and behavioral signals to infer others' intentions, trustworthiness, and relational potential. Consistent with SIP's *impression formation* stage, all participants **sought common ground** with other users (100%, $n=19$), forming impressions by assessing others' content for similarities in age, interests, and *"aesthetic"* as summarized in Table 1. When scrolling through the public feed, many teens (74%, $n=14$) engaged based on shared interests by liking or commenting on posts with relatable humor, positivity, or friendliness, which SIP theory characterizes as substitutes for offline interpersonal cues. When asked to chat with other users, many teens first checked their profiles for assessing safety and relatability. Even when teens could not explicitly articulate why they felt drawn to a user, profile content enabled them to form tentative impressions and assess trustworthiness prior to direct interaction, demonstrating how SIP operates through cue accumulation rather than immediate certainty. For

instance, P5 friended Sarah (safe persona) after seeing similarities that made her feel safe to connect:

"I would just start off with going through their profile. And see if there's anything similar to mine, or something that I like so it felt pretty safe talking to Sarah." - P2 (13-year-old, Female)

Once teens felt an initial connection, they often built upon these interests by inquiring further on hobbies or interest-based recommendations (e.g., *"what soccer position do you play?"*) or (*"I saw that you love reading and I've been looking for some new books"*). Participants also found location-based information shared by others helpful to connect, asking them for recommendations (42%, $n=8$) or about their favorite local parks or attractions, e.g. (*"Do you go to Disney often?"*). Some teens (47%, $n=9$) asked about school or education as low-risk probes to assess commonalities for connection. This was a SIP-consistent strategy for reducing uncertainty before deepening engagement [91].

As interactions progressed, teens engaged in selective self-presentation to **share personal information for relatability**, reflecting SIP's emphasis on *controlled information sharing* for trust-building. Rather than disclosing indiscriminately, teens monitored reciprocity and mirrored the level and type of information others shared. For instance, most participants (95%, $n=18$) included city or state-level location information on their profiles or shared it during conversations only after others had done so, framing disclosure as socially normative and low-risk. Some justified this based on how they share on other platforms or noting that city-level location is easily discoverable online. P2 shared,

"I added my location, because, like on my Instagram, you can easily find my location through my school, and I used to have it on there." - P2 (17-year-old, Male)

Several teens shared personalized bios (58%, $n=11$) or academic information (58%, $n=11$), such as their interests, hobbies, school name or graduation year, as identity signals and conversation starters, e.g. *"I love going out, watching movies and reading in my free time. I want to become a millionaire and be my own boss"*, *"Class of #ABC2026"*. When asked to upload their profile pictures or create post, most teens (63%, $n=12$) shared photos related to personal interest or experience, as icebreakers, ranging from fan-based content about their favorite music band or sports team, food, pet photos, or photos with friends. A few teens (21%, $n=4$) redirected the conversation with the risky persona, Bryan to share a safer way for staying connected (e.g., email instead of the requested phone number). Overall, teens used reciprocity and selective self-presentation to build trust and express themselves.

Consistent with SIP's *relational development* stage, all participants moved beyond surface-level interaction to offer **advice, encouragement, or informational support**. For instance, several participants offered (79%, $n=15$) Dave (risky persona) advice on what to post (e.g., *"Maybe you should post about your last vacation"*) and encouragement (e.g., *"I believe in you"*). Several teens interested in robotics inquired Bryan about his robotics post to get inspiration and information. When Bryan sought academic advice, many of the older teens provided online resources, advice, and encouragement for college applications (47%, $n=9$). P17 summarized her academic connection with Bryan as,

Table 1: Teens' Strategies for Connection and Protection Using SocialSim through the SIP Framework

SIP Stage / Principle	Connection (Strategy & Actions)	Protection (Strategy & Actions)
Stage 1: Impression Formation Users interpret social cues & appraise potential risks.	Sought Common Ground <i>Formed initial impressions from age, interests, "aesthetic".</i> - Engaged around shared interests (100%) - Engaged with similar content (74%) - Asked for location/interest recommendations (42%) - Asked about school/academic stage (47%)	Investigated Safety/Identity <i>Decoded demographic traits, content, and behavior.</i> - Assessed context cues (42%) - Checked profiles for mismatched information/scams (32%) - Asked fact-checking questions (21%) - Preferred personalized profiles (21%) - Considered gender cues (42%)
Stage 2: Social Information Exchange Users engage in selective self-disclosure & privacy management.	Shared Personal Info for Relatability <i>Reciprocated disclosure via school/location/bios.</i> - Shared some level of location (95%) - Shared school/batch year (58%) - Shared bios/intros (58%) - Shared photos (63%) - Shared other ways to connect (21%)	Limited Personal Information <i>Controlled disclosure to reduce risk.</i> - Shared only broad location (79%) - Used default/non-identifiable profile pictures (68%) - Shared obscured/partial post photos (26%) - Used pseudonyms/nicknames (47%)
Stage 3: Relational Development Users adjust relational expectations & interaction strategies.	Gave/Received Advice and Support <i>Sympathized with struggles and similar goals.</i> - Shared advice for posting (79%) - Shared academic goals/college advice (47%) - Shared struggles/rants (26%)	Avoided Unsafe Interaction <i>Judged tone, enthusiasm, and language.</i> - Engaged in reciprocal efforts only (53%) - Avoided initiating interaction (37%) - Replied briefly to suspicious users (37%) - Delayed/avoided friending (26%)

"Since they're also a sophomore in high school, and I'm about to be a sophomore, so it'd be good to see if I wanted to be friends with them, or be inspired by something they did." - P17 (15-year-old, Female)

Teens' interactions closely followed SIP, progressing from cue-based impression formation to selective, reciprocal self-disclosure and deeper relational engagement. By interpreting social cues and adjusting their sharing over time, teens balanced social connection with safety in online interactions.

5.1.2 Teens used several strategies and actions to protect themselves, such as investigating users' identity for safety, limiting identifiable information, and avoiding interaction with suspicious users. Alongside connection building, teens simultaneously engaged in protective behaviors, aligned with SIP's staged process of *risk appraisal and relational management*, as summarized in Table 1. During *impression formation*, many participants (42%, n=8) **investigated the safety and identity of other users** and assessed their trustworthiness early in interactions through evaluating contextual information, such as the type of content they posted, bio, apparent age, typing style, and *"energy"*, labeling users as *"safe"* or *"suspicious"*. They were quick to notice if any user (e.g., Dave) did not use acronyms or common *"Gen-Z"* language, considering them to be older and potentially unsafe, showing how social cues were used to identify risks. Several participants were discouraged by Dave's post or because his bio felt *"too millennial"* as P12 explained,

"I don't think I like this one (Dave), cause I'm not a big fan of men trying to boast [on post]." - P12 (19-year-old, Female)

Teens also cross-checked profile information and chat messages to identify any mismatches. For example, P3 asked all users which school they were from or specific questions about their location to identify if they were a scammer or a bot. Alternatively, some teens preferred talking to those who had shared some personal

information and posted personalized introductions or photos (21%, n=4), as it helped verify that the user was a real person. Gender bias (42%, n=8) played a role as well, as some of the female participants preferred connecting with other female users due to prior unsafe experiences with male users, especially stranger accounts. P19 explained,

"I feel bad, but gender kind of played a role. You guys probably know, online when it's the opposite gender, it can be a little sus, and you have this hunch that they might not be texting you for the best purpose." - P19 (17-year-old, Female)

Beyond initial appraisals, all teens (100%, n=19) engaged in selective self-presentation by sharing **limited personal information** to conceal their identity, aligning with the SIP stage on *selective sharing for privacy management* as relationships develop [91]. Specifically, many teens either did not provide their location or provided broad location-based information (79%, n=15). For instance, some teens only shared their country, whereas others gave their state and/or city, or vaguely described their location, e.g., *"I live close to the area,"* without disclosing the area. A few teens thought that giving school information was safe as long as they were not identified. P8 justified,

"I don't give like I'm in X building. But I'll give state or city, that's usually as far as I'll go or, I think college is also fine because there's a lot of people there." - P8 (19-year-old, Male)

Many teens also prioritized privacy when posting photos on the platform. For instance, several teens did not upload a profile picture or used the default picture (68%, n=13), to further protect their identity. Similarly, many teens found online photos that represented them or their interests, personal photos that hid their faces (26%, n=5), or photos of pets, food, or nostalgic places. Additionally, some teens (47%, n=9) used pseudonyms or fake user names to hide their

real names, reflecting SIP's notion of strategic identity management in mediated settings.

Finally, as interactions unfolded, teens relied on SIP's *relational change* stage to determine whether to deepen engagement or disengage. Many **avoided unsafe interaction** (89%, n=17) with suspicious users, as they learned and processed more information about them. For instance, some teens avoided commenting on posts (e.g., likes only) or initiating chats (i.e., waiting for the chat actors to message them) as they generally did not talk to strangers online (37%, n=7). While chatting, many participants used social cues to remain safe by judging the other user's enthusiasm, tone, and language before engaging further. Some teens continued with those who reciprocated their effort during the conversation (53%, n=10). For instance, many users "ghosted" the non-risky user, Frank who appear uninterested due to his reserved personality. Alternatively, P5 was quick to trust the non-risky user, Sarah, as she felt that they reciprocated efforts in equally sharing about their lives:

"I really felt like this chat was very good. Cause she shared information from her side. So like we're equal, and we trust each other a bit. Felt like I'm talking to my friend." - P5 (13-year-old, Female)

At times, teens were particularly brief in their replies to suspicious users (37%, n=7), or delayed friending strangers (26%, n=5) as they wanted to get to know them or wanted mutual friends before friending them. Across the SIP stages, teens used several protective strategies to keep themselves safe: using social cues to guide impression formation, selectively managing disclosure, and disengaging when interactions felt unsafe. These behaviors highlight SIP's relevance for understanding how teens balance connection and protection within the progressive nature of relational development online.

5.2 Teens Behavioral Responses to Risks, Nudges, and Safety Actions (RQ2)

Building on RQ1, which showed that teens actively engage in SIP through *impression formation*, *selective disclosure*, and *relational management* even without intervention, in RQ2, we examine how teens interact with real-time safety nudges in these processes during risky interactions. We structure RQ2 around teens' action pathways: protective actions, self-managed responses, and final relational outcomes, highlighting how strategies manifested differently depending on perceived risk severity.

5.2.1 Protective Actions as Relational Interruptions. Across risk scenarios, **protective actions** (e.g., blocking, reporting, deleting, flagging, hiding similar messages) marked moments where teens chose to interrupt or terminate relational progression, consistent with SIP's *relational change* stage. However, teens' thresholds for invoking protective actions varied substantially by perceived risk severity. Protective actions were most common in response to potentially **explicit content**. When Dave shared a blurred image marked with a sensitivity filter, most teens (42%, n=8) selected the *Protective Actions* option provided by the nudge (Table 2), most often choosing to block and report the user (see Table 7 and Figure 5 for the flow of protective actions). Teens described these actions as necessary to immediately cut off contact with a user they perceived

as unsafe, even when the explicit nature of the content could not be fully verified. As P2 explained:

"If there's something that the platform flags. I really don't want to see it like it could be something explicit, like a gun or revealing photo." - P2 (17-year-old, Male)

This was also the only scenario in which teens frequently used *Hide Similar Messages* or *Notify Other Users*, reflecting a desire to prevent future exposure and protect others. Some teens framed these actions as a responsibility to warn peers, particularly given the perceived severity of the risk. P16 explained the need for notifying other users:

There are people who don't really want to see that stuff so you can just notify them ahead of time. So they have a warning, at least. - P16 (17-year-old, Female)

In contrast, the **cyberbullying comment** elicited mixed reactions, likely due to the low-level perceived risk. After viewing the hidden cyberbullying comment from Emily, which said "lol wth is this post?", teens often shrugged it off as normal social media behavior. While some participants questioned the comment, others tried to justify it; for example, believing that Emily may have disagreed with any opinion expressed on the participants' post (e.g., favorite sports team). However, most teens (95%, n=18) chose the option to "Unhide" the comment (see Table 5), after which they chose to "Flag and Delete" the comment (53%, n=10) as a final protective action, as they found the comment to be unnecessarily mean. P18 explained:

"I was like, well, what's the comment? who's commenting? Maybe they have an issue with the post. And it was an unnecessary comment that wasn't helpful at all, so I decided to flag." - P18 (17-year-old, Female)

For the **information breach**, protective actions were least common. When Bryan asked, "Can we exchange numbers?," only a small subset of teens (16%, n=3) escalated to blocking or reporting Bryan (see Table 6), typically when the request felt uncomfortable or suspicious. Some teens (21%, n=4) relied on suggested messages (Table 2, Figure 3) to respond back to Bryan as a protective measure, noting that it can be difficult to think of a reply on the spot, so the suggested options made it easier to respond. Overall, teens reserved protective actions for situations they perceived as posing immediate or uncontrollable harm, illustrating SIP's emphasis on proportional relational responses to risk cues.

5.2.2 Self-Managed Responses and Nudge Bypassing. In many interactions, teens treated nudges as **informational cues rather than directives**, continuing interactions while managing risk themselves, which was a SIP-consistent pattern of ongoing cue interpretation and selective self-presentation (Figure 5 details the flow of self-managed responses). Across all risk types, teens frequently sought to **verify risk before acting**, reflecting SIP's emphasis on uncertainty reduction during impression formation. As mentioned above, nearly all teens (95%, n=18) chose to unhide the **cyberbullying comment** to assess its severity, while many attempted to unblur the **explicit image** or asked whether it could be viewed (68%, n=13). Teens explained these behaviors as driven by curiosity, distrust of automated filtering, and the need to calibrate appropriate responses. As P14 noted:

Nudge Dimension	Nudge Choice	Cyberbullying	Info Breach	Explicit Content
Protective Action(s)	Block, Report, Hide Similar Risks, Notify Others	0%, n=0	16%, n=3	42%, n=8
	OK → Suggested Message	N/A	21%, n=4	N/A
	Unhide/OK → Flag & Delete	53%, n=10	N/A	N/A
Self-Managed Responses	Unhide/OK → Self-Response	11%, n=2	42%, n=8	26%, n=5
	Unhide/OK → Ignore	37%, n=7	21%, n=4	32%, n=6

Table 2: RQ2: Summary of Teens' Responses to Nudges. *The percentages may not add up to 100% as the categories were grouped thematically and represent actions available at different levels of the nudge flow (e.g., suggested responses appeared only after clicking "OK"). The most common action for each nudge is bolded.

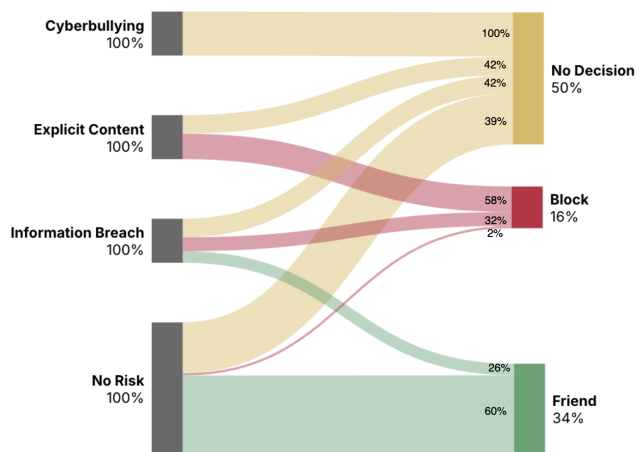


Figure 4: Sankey diagram: Teens' final relational decisions (friend, block, or no decisions) across risky and non-risky users.

"Yeah, obviously, I'm curious, cause instead of only censoring bots, [other platforms] censor a lot of things. Like on Tiktok, lot of the times these filters are a little flawed and censor some unnecessary things that you know they're not hurtful. They're just comments." - P14 (15-year-old, Male)

Some teens ignored the comment (37%, n=7) or responded with a comeback (11%, n=2), as they thought that users who make unnecessarily mean comments often back off when they receive a response. For instance, P9 responded to Emily with a reply *"Haters are my motivators :)"*. One participant described negative comments as beneficial for engagement and visibility, illustrating how prior experience shaped relational interpretation. Rather than indicating recklessness, these actions reflect deliberate risk assessment prior to relational decisions, highlighting how teens actively processed cues even when nudges were present.

For the **information breach**, most teens (48%, n=8) acknowledged the nudge by clicking "OK" and responded to Bryan in their own words (Table 2). Teens often rejected the request directly (e.g., *"I don't really know you"*), offered alternative ways to connect (e.g., *email*), or changed the topic. One of the participants offered to help Bryan with college applications on the platform, while also saying no to giving her phone number *"I don't want to [share my*

number] but I think there's lots of resources on Reddit and if you have questions, I can respond here." Teens justified responding to Bryan as they did not consider this scenario to be high-risk, as the requested information was not considered too sensitive. As P7 explained:

"I didn't know if this person was being genuine or had malicious intent. If they were asking for my address, or more personal info, like where my parents work or who they are then I'd block." - P7 (14-year-old, Male)

A few teens felt comfortable sharing limited information (e.g., email or phone number), emphasizing their ability to disengage later if needed. As P8 explained:

"I've given my number that's not going to do anything, and if he turns out to be someone that shouldn't have my number. I'll probably just block them." - P8 (19-year-old, Male).

For **explicit content**, some teens acknowledged the nudge ("OK") and chose a different action. Some participants (26%, n=5) responded back to Dave inquiring about the photo (e.g., *"What did you send?"*). Others responded by educating him that the photo violated community guidelines, or replying that he should not post that at all (e.g., *"Hell no"*). Other participants were indifferent to the risk or wanted to continue chatting to assess risk before any drastic action. In a few cases, teens tried to guess what the photo was as explained by P1:

"You can kind of guess what the image is, even though it's blur. But from what I can see now, I don't even know what it is, but it doesn't look that bad." - P1 (17-year-old, Female)

These behaviors demonstrate that nudges did not substitute for teens' own SIP-based risk appraisals, but instead operated alongside ongoing processes of cue interpretation and judgment.

5.2.3 Final Relational Outcomes and the Role of Nudges. Final decisions to *friend*, *block*, or make *no decision* reflect the cumulative outcome of SIP processes across impression formation, selective disclosure, and relational evaluation. These outcomes are synthesized in Figure 4 and Figure 5 that visualize teens' final relational decisions across risky and non risky scenarios. Most teens blocked Dave (58%, n=11), even when uncertain about the content, indicating that severe or ambiguous risk cues stalled relational progression. Notably, none of the participants friended Dave. In contrast, teens were most ambivalent towards Bryan (42%, n=8), with some blocking him (32%, n=6) and others friending him (26%, n=5) after verifying

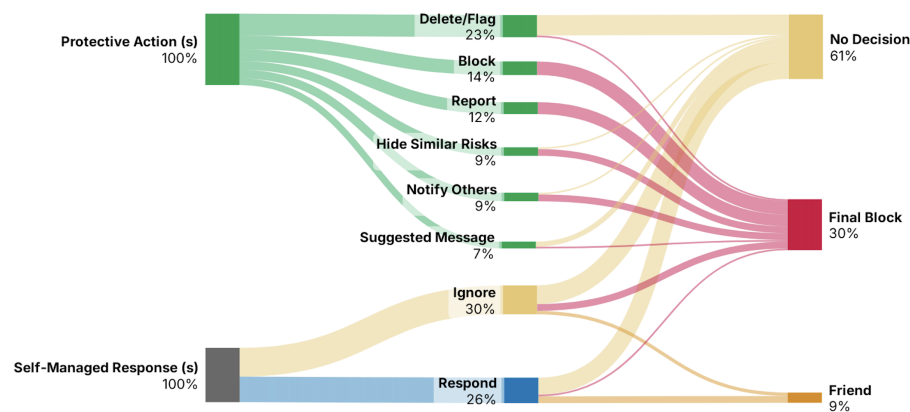


Figure 5: Sankey diagram: Flow from teens' nudge responses (protective actions, self-managed responses to final relational decisions for risky users. Through the SIP lens, the diagram illustrates how teens' interpretation of risk cues and nudge guidance shaped whether relationships progressed, stalled, or terminated.

his information or shared interests. This distribution highlights how selective self-presentation and reciprocity shaped relational outcomes under ambiguity. As P6 summarized:

"It's either they're [Bryan] like a weird scammer or they're a genuine high school student. I guess I'm not entirely sure which one they are so I wouldn't take the option to block them yet." - P6 (19-year-old, Female)

For Emily, teens largely contained the risk by deleting the comment without taking any final decision Figure 5, as the comment-based interaction often did not evolve to relational development. A few participants (10%, $n=2$) explicitly preferred the cyberbullying nudge, viewing public risks as more socially damaging than private ones. One of the teens explained their preference for addressing the public cyberbullying comment over the information breach:

"The comment was a little more forward and discriminating, so that notification would really help, so I could delete it. For the phone number one, I could probably tell he's asking for my information, I don't need a notification to tell me that." - P19 (17-year-old, Female)

Overall, nudges most strongly influenced outcomes in high-risk or uncertain situations, particularly explicit content, where teens favored decisive protective actions. Also, when teens took protective actions, it never resulted in a friending decision. Teens took no decision most commonly across all risks as in lower-risk contexts, they relied more heavily on their own judgment, treating nudges as contextual cues that could be bypassed or adapted when they felt capable of managing situations independently. These findings describe how nudges operate as supportive scaffolding within adolescents' ongoing decision-making processes without displacing autonomy.

5.3 Teen Nudge Perceptions and Preferences (RQ3)

In this section, we elaborate on teens' preferences for adopting and improving nudges, focusing on proactive safety, learning, trust,

explainability, and personalization, as summarized in Table 8 of Appendix.

5.3.1 Valuing Nudges for Proactive Safety and Learning. A majority of teens (95%, $n=18$) **valued the immediate protection** (Table 8) offered by nudges, with many considering the simulated platform as safer than typical social media (58%, $n=11$), with a preference to keep similar nudges enabled on other social media (68%, $n=13$). Teens found nudges particularly helpful (84%, $n=16$) when they felt confused about whether an encounter was risky or not. They appreciated how the nudges alerted them about the risks using attention-grabbing cues so that they can assess the risk and take multiple safety options (37%, $n=7$). With regards to cyberbullying comment risk, P2 mentioned,

"Like the comment at the beginning - I felt safe that the platform took steps and actions, saying, 'Hey! This comment is blurred for so and so reason'." - P2 (17-year-old, Male)

Beyond immediate protection, many teens (89%, $n=17$) perceived **educational value** in nudges. Many believed that nudges could support long term behavior change (68%, $n=13$) by repeatedly training users to recognize and respond to online risks. Teens viewed nudges as especially beneficial for younger or less experienced users (58%, $n=11$), describing them as "training wheels" for online safety on social media. Some teens further added that that nudges could be turned on as a default when teens first sign up on the app, after which they can gradually take them off once they become resilient. P12 explained,

"When I had first entered the Internet and gotten these warnings, it would've taught me, oh, this is bad. Cause I never got those when I was growing up. I think it would be pretty helpful, especially for younger teens, it's almost like training wheels with repeat practice of learning what's good and bad online" - P12 (19-year-old, Female)

Teens also wanted nudges to adapt based on prior risk experience (26%, $n=5$). For instance, the information breaching nudge was

often considered less useful as they had already encountered such risks before and knew how to handle it, suggesting the need to account for prior exposure. Others wanted educational nudges to have **broader impacts** that target safety training for not only victims but also potential offenders (21%, n=4), thereby shifting the responsibility from victims to preventing the risk at the root cause.

5.3.2 Improving Trust and Explainability for Nudges. While teens valued nudges for proactive safety and learning, many teens also expressed **distrust in AI accuracy** for detecting risk (74%, n=14), based on prior experiences with false flags on other platforms. At times, this skepticism increased the need to verify risks themselves before using protective action offered by nudges. As a result, some teens (42%, n=8) wanted hybrid approaches to build trust through verifying risks with third-party moderators or the community review. P13 summarized,

"If you flag someone and want to report a user, I think having a moderator, or some sort of mechanism, to review each case why that report is being filed. I think that'd be helpful." - P13 (18-year-old, Male)

Teens had similar concerns for the "notify others" option, as they were worried it could spread misinformation or fuel "cancel culture" (26%, n=5), if used unethically. To overcome this, some teens (74%, n=14) wanted **hybrid and transparent systems** that could help verify risks accurately and assess their prior risk knowledge. For instance, some teens preferred questions about safety knowledge on signing up to an app, while others wanted automated ways for the algorithm to track risk handling, which would inform future need for nudge. P10 explained,

"The algorithm should take into account and record what you want to block. I think the first post where it was like, oh, be careful, some harmful words. How I decided to ignore that, and bypass the warning, maybe it should take into account for what nudge I wanna see next." - P10 (19-year-old, Female)

Teens also emphasized **tailoring nudges by risk severity** (74%, n=14). For high-risk cases (e.g., explicit content), they wanted clearer explainability about safety choices (21%, n=4), as well as specific detail on *where* and *how* the community guidelines were violated, possibly through a "See More" option. A few teens (16%, n=3) also wanted familiar terminologies to make protective actions consistent with other platforms. In contrast, for lower-risk situations, several teens (32%, n=6) preferred subtle, non-disruptive guidance that supported their decision making without heavy-handed blocking as P11 shared,

"I feel like a different type of pop-up might be more appropriate for this [information breach], like 'this text might be asking for sensitive information. Here are some tips or ways to get out of the situation' - instead of like explicitly blocking them." - P11 (19-year-old, Female)

Overall, teens raised important concerns regarding the trust, explainability and tailoring of underlying algorithms and AI-based risk detection, which directly influenced their nudge perception and use.

5.3.3 Personalization and Collective Safety. Many teens acknowledged that risk is often subjective, therefore, they wanted nudges that offered **personalization and collective safety** according to their preferences (42%, n=8). For instance, they desired control over specific keywords for triggering nudges, disabling nudges for trusted users and options for long-term privacy preferences rather than a one-time block or report. To this end, P8 wanted algorithms for custom safety preferences that shape the user feed on social media, as they are currently primarily based on interests:

"What if you make your own version, like define as the user what you don't want to see on your feed cause maybe I find something that's fine with me. But [researcher] doesn't find something that's cool with her so we can customize that, like whatever I would like to have removed completely." - P8 (19-year-old, Male)

Teens also wanted personalization of nudges based on frequency of violations. They were willing to give others the benefit of doubt and preferred stronger protective actions only after repeated or escalating offenses (32%, n=6) (e.g., if the risky persona insisted on asking for a phone number). In contrast, a few teens (10%, n=2) wanted to customize the timing of the nudge to appear, either earlier, before or as soon as the risk occurred.

Importantly, many teens valued **collective safety** (63%, n=12) offered by "notify others" feature (57%, n=11), seeing it as a way to preemptively warn peers about risk, provided it includes verification and third-party reviews. However, they also wanted control over who is notified and how, such as anonymous or selectively sharing (37%, n=7) with close friends, or only with others connected to the risky user. This felt safer to teens than publicly notifying others, to protect themselves from judgement and retaliation. P18 explained,

"They might judge me for getting a message I didn't even want. So if it could be done anonymously instead of identifying me - I would not want to be personally put on blast like that." - P18 (17-year-old, Female)

Our findings from RQ2 and RQ3 show that teens could envision a path toward more effective nudging when interventions are personalized to their preferences and safety expertise, supporting not only individual safety but also collective safety and learning. Teens viewed nudges as a useful but limited form of support, most helpful in ambiguous or higher-risk situations, while relying on their own judgment in lower-risk contexts. They treated nudges as optional cues rather than directives and preferred designs that were proportionate, trustworthy, and responsive to experience.

6 Discussion

In this section, we discuss the implications of our findings and provide recommendations for nudges and behavior change interventions based on our evaluation.

6.1 SIP as a Lens for Showing How Teens Are Already Active Risk Managers (RQ1)

Our findings challenge dominant narratives that frame teens as risk unaware and in need of paternalistic online protections, such

as parental controls [31, 38, 48], cellphone restrictions [70], or social media bans [14, 56]. Instead, teens in our study consistently demonstrated existing strategies for balancing connection with self-protection. Drawing on the Social Information Processing (SIP) framework [92], we observed teens actively forming impressions through implicit social cues (e.g., language, profiles, assumed age), selectively exchanging information to protect their identity, and disengaging when early signals raised concern. By examining teens' in-situ decision-making, we found that many relied on nuanced social judgments, such as identifying "red flags," concealing personal information, and assessing relatability, often in ways that automated systems may fail to capture, building on prior insights on teens' safety practices [30]. Early cue interpretation shaped later boundary-setting and safety actions, including how teens responded to platform nudges, suggesting a developmental pathway from impression formation to protective behavior. As interactions progressed, teens' relational decisions further reflected their ability to assert social agency, befriending strangers when reciprocity and trust were established, while disengaging or blocking others even in the absence of overt risk. These patterns align with adolescents' developmental drive for connection and identity exploration [35, 75], as well as prior work showing teens' capacity for controlled self-disclosure and safe relationship-building online [44, 67].

Importantly, our results position teens as **active risk managers**, rather than passive recipients of platform protection, whose cue-based risk processing reflects adaptive privacy norms rather than dysfunctional fear [49]. In contrast to approaches that emphasize constant monitoring, restriction [63], and youth vulnerability online [43, 53], our findings point to the importance of "Safety by Design" approaches that augment teens' existing cue-processing and resilience [71, 78]. Importantly, our work highlights an information asymmetry: platforms often know more about other users (e.g., prior violations or interaction patterns) than teens can infer from visible cues. This points to a role for contextual risk cues drawn from platform-level signals to support teens' fine-grained social reasoning and contextual risk assessments [8, 50]. By applying the lens of SIP, we demonstrate how teens' early cue-based judgments scaffold later boundary-setting and safety decisions, suggesting that interventions are most effective when they are embedded within designs to augment teens' in-the-moment decisions, without making prescriptive choices on their behalf.

6.2 Contextual Support, Not Control: How Teens Used and Perceived Safety Nudges (RQ2 & RQ3)

Table 3 synthesizes findings from RQ2 and RQ3, showing close alignment between teens' real-time behaviors and their perceptions of safety nudges. Teens treated nudges as contextual cues weighed alongside their own social judgments, relying on them most in high-risk or ambiguous situations, particularly explicit content, to validate risk and be immediately protected. In contrast, for lower-level risks such as information breaches or cyberbullying, teens either deleted the risk with no safety decision or primarily relied on self-managed strategies, consistent with prior work on skepticism toward platform safeguards [17]. These differences observed across risk scenarios should be interpreted in light of

both severity and interaction modality. For instance, the cyberbullying scenario unfolded in a public comment context, whereas the information request and explicit scenarios occurred via private messaging, which changed the social stakes, visibility and available response strategies. As a result, variations in perceived risks and nudge engagement may reflect not only differences in risk type, but also these underlying structural factors. A key contribution of our work is that by capturing both in-situ behavior and post-hoc reflection [77], we show that teens' preferences for agency-preserving, risk proportionate nudges were reflected in their actions, challenging traditional accounts of the privacy paradox [12], where stated privacy concerns often do not translate into safe choices. Nudges were valued not as directives, but as confirmation in "gray-area" encounters [6, 82], building on prior findings on proportional and context-aware nudging [52, 73]. Teens emphasized the educational value of nudges, particularly for less experienced younger users, describing them as early "training wheels" that support risk awareness through repeated practice, supporting developmental trajectories toward independence [19]. Yet, teens' selective bypassing of nudges highlights that their effectiveness is bounded by teens' existing competencies and confidence in the nudge. Building on Kim et al.'s trust-enabled privacy framework [50] on how platform design shapes teens' trust and boundary regulation, our findings show how trust is negotiated in real-time and how nudges influence boundary decisions only when perceived as credible, trustworthy and verified by humans. For instance, teens envisioned trusted personalization of feed algorithms beyond nudges, that reflect their safety needs (e.g., downranking explicit content while preserving lightweight banter). Together, this motivates moving beyond nudges for risk detection alone toward "Safety by Design" supports that surface platform-held context teens otherwise lack (e.g., repeated terms-of-service violations).

Yet, our findings also surface a methodological challenge in evaluating online safety nudges for youth. Due to ethical constraints, our study focused on simulating relatively low-risk scenarios, aligning with prior work showing compromise in realism to ensure control and safety [7]. As a result, many teens perceived information breaches and cyberbullying as manageable, reducing the apparent need for nudges in these contexts. This limitation complicates efforts to conduct evidence-based research on interventions intended for higher-risk encounters, identified as moments where nudges may matter most. To address this challenge, future research should follow the approach advocated by Ali et al. [9], who argue for engaging teens as long-term partners in shaping research methods and protocols, enabling the study of more realistic and escalating risk encounters while balancing ethical protections. This challenge is further compounded by teens' distrust of algorithmic risk detection [76], as they often questioned the consistency and accuracy of automated flagging and instead sought verification or relied on their own judgment. Even collective safety features, such as notifying other users, were viewed cautiously without verification, reinforcing prior calls for human-in-the-loop [57] and community-based [100] moderation approaches.

Table 3: Alignment Between Teens' Behaviors (RQ2) and Perceptions of Nudges (RQ3)

Shared Theme	RQ2: Observed Behaviors	RQ3: Teens' Perceptions of Nudges
Contextual Use of Nudges	Teens acted on nudges primarily in high-risk or ambiguous situations and bypassed them in lower-risk contexts.	Teens viewed nudges as most useful when tailored to risk severity or ambiguity, as they could handle low-risks.
Preservation of Agency	Teens selectively followed, adapted, or ignored nudges while continuing to self-manage interactions.	Teens emphasized that nudges should support, not override, their agency and safety skills.
Validation of Risk Appraisal	Teens used nudges to confirm or unhide the risk before escalating to protective actions or self-managing responses.	Teens described nudges as helpful for confirming “gray-area” risks and validating risks for collective safety.
Distrust in Risk Detection	Teens sought verification or delayed action when they distrusted automated risk detection and sensitivity filters.	Teens expressed skepticism toward algorithmic detection and preferred hybrid approaches with human verification.
Learning and Skill Development	Teens referenced nudges when reflecting on how they recognized and assessed the most suitable responses to the risk.	Teens viewed nudges as early-stage learning support for long-term safety training, adapting to their experience.

6.3 Design Implications for Online Safety Nudges for Teens

We provide design guidelines for effective safety nudges, based on our findings that nudges should function as contextual learning supports building on teens' risk experiences, rather than rigid guardrails.

6.3.1 Design Nudges to Be Context-Sensitive and Risk Proportionate. Nudges should be calibrated to risk severity and ambiguity, providing stronger intervention in high-risk situations (e.g., explicit content) while remaining lightweight or optional for lower-level risks that teens can manage independently. This supports proportional, context-aware nudging that complements teens' existing strategies rather than disruptive, one-size-fits-all nudging [73, 74].

6.3.2 Support Learning and Teen Agency Through Developmentally Calibrated Nudges. Nudges should function as learning supports rather than directives, helping teens reflect on risk in ambiguous situations while remaining optional and easy to bypass. For less experienced users, nudges may serve as early “training wheels” that scaffold risk awareness and can be gradually reduced as confidence and autonomy develop, aligning with developmental trajectories toward independent decision-making [19].

6.3.3 Address Trust Through Transparency and Human Verification. Nudges should be grounded in transparent, explainable, and human-verified risk detection, as teens' trust and reliance on nudges depends on accuracy of risk detection and clarity on why a nudge appeared. Incorporating youth involvement and human-in-the-loop moderation can strengthen nudge credibility and accountability, particularly for collective safety actions such as notifying others [57, 80].

6.4 Limitations and Future Work

This work captures teens' real-time responses to safety nudges, revealing how their effectiveness varies across risk contexts. However, this study has several limitations that inform directions for future

work. While the simulated environment using WoZ approaches has been established to enable controlled and ethically responsible evaluation [37, 55, 103], it may not have fully elicited genuine friendships or long-term relational dynamics. Although designed to approximate in-situ use, sessions were conducted over video conferencing with researchers able to observe participants' screens, which may have influenced their behavior. While our participants were not familiar with SocialSim platform before the session, we also acknowledge that the familiarity with similar social media apps and initial learning effects from the first-time use of our app may also have influenced how they engaged with the system. As such, our findings primarily reflect how teens form initial impressions and navigate early-stage interactions with unfamiliar others online. Further longitudinal research is required to determine whether these nudges remain effective as risks escalate and as users become accustomed to system alerts. Real-world deployment also introduces additional challenges, including imperfect risk detection, how teens respond to false alarms, and when and how such “training-wheels” supports should be introduced, adapted, or withdrawn. In practice, frequent false alarms could lead to fatigue, raising important questions about the threshold at which AI interventions shift from helpful scaffolding to intrusive surveillance, and who, i.e., whether youth, platforms, or parents should control these features.

Alternative approaches, such as LLM-based agents simulating peer interactions, could improve automation and consistency. However, because of growing concerns and legal scrutiny around youth interacting with AI chatbots [45, 54, 98], we used a human-mediated WoZ setup to preserve oversight and reduce potential harm. We also standardized and monitored persona enactment, though some variation is unavoidable in WoZ methods. As a result, teens' judgments of linguistic authenticity may partly reflect researcher performance, not just the intended risk manipulation, and this should be considered when interpreting their suspicious impressions of personas.

Further, given our ethical responsibility to avoid exposing teens to more than minimal risk, some participants perceived the simulated risk scenarios as relatively low in severity. Therefore, the

research community faces an ethical challenge in modeling realistic teen risk scenarios, particularly those beyond minimal risk. Additionally, due to the qualitative WoZ design, we did not include a baseline or no-nudge condition, limiting our ability to isolate the causal effects of nudges. Future work should build on simulation-based methods to directly compare outcomes with and without nudges and more precisely benchmark their impact.

Finally, while our sample intentionally prioritized youth from underrepresented backgrounds, the sample size was small, limiting generalizability across broader populations. Future work should replicate and extend these findings with larger and more diverse samples, including youth from different geographic regions, socioeconomic backgrounds, and cultural contexts, to examine how experiences with risk and nudges may vary across populations. Despite these limitations, this work represents an important first step by providing an ecologically grounded, near real-time examination of how teens engage with safety nudges embedded within a realistic, simulated social media context.

7 Conclusion

This work advances the adolescent online safety literature by showing that many teens already possess meaningful competencies for navigating online risk, and that real-time safety nudges are most effective when they complement rather than override these capabilities. Rather than functioning as guardrails that constrain behavior, nudges should act as training wheels, offering situational support in moments of ambiguity or heightened risk. Their effectiveness depends on how well they align with teens' developmental needs, prior experience, and evolving sense of agency. Our findings challenge deficit-based narratives that frame teens as uniformly vulnerable online, instead highlighting their active role in interpreting cues, managing relationships, and making proportional safety decisions. Looking ahead, this work points to the need for teen-centered research that examines how support should adapt over time, including when it is necessary, when it can be removed, and how systems can recognize growing competence. Grounding intervention design in teens' lived practices offers a path toward online safety systems that foster learning, resilience, and autonomy by empowering adolescents as capable social actors rather than restricting their participation.

8 Selection and Participation of Children

This study involved adolescents aged 13–19 and was conducted in accordance with institutional ethical approval. Participants were recruited through youth-serving organizations and community networks to include perspectives from underrepresented backgrounds. Written informed consent was obtained from participants aged 18 and older, and both parental consent and youth assent were obtained for participants under 18. Participation was voluntary, and teens could withdraw at any time without penalty. The study was designed to minimize risk by using a simulated social media environment and age-appropriate scenarios, ensuring that participants were not exposed to real-world harm.

Acknowledgments

We thank the Associate Chairs and anonymous reviewers for their constructive feedback and we thank all participants for their time and contributions. Dr. Wisniewski's research is supported by the U.S. National Science Foundation under grants #TI-2550746, #CNS-2550834, and #IIS-2550812, and by the William T. Grant Foundation grant #187941. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsors.

References

- [1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. 2017. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)* 50, 3 (2017), 1–41.
- [2] Zainab Agha, Naima Samreen Ali, Jinkyung Park, and Pamela J Wisniewski. 2025. A systematic review on design-based nudges for adolescent online safety. *International Journal of Child-Computer Interaction* 43 (2025), 100702.
- [3] Zainab Agha, Karla Badillo-Urquiola, and Pamela J Wisniewski. 2023. "Strike at the Root": Co-designing Real-Time Social Media Interventions for Adolescent Online Risk Prevention. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.
- [4] Zainab Agha, Neeraj Chatlani, Afsaneh Razi, and Pamela Wisniewski. 2020. Towards conducting responsible research with teens and parents regarding online risks. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.
- [5] Zainab Agha, Matthew Martinez, Naima Samreen Ali, Dominic DiFranzo, and Pamela J Wisniewski. 2025. SocialSim: An Open Source Platform for Conducting Behavioral Intervention Research on Social Media. In *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing*. 528–531.
- [6] Zainab Agha, Jinkyung Park, Ruyuan Wan, Naima Samreen Ali, Yiwei Wang, Dominic DiFranzo, Karla Badillo-Urquiola, and Pamela J Wisniewski. 2024. Tricky vs. Transparent: Towards an Ecologically Valid and Safe Approach for Evaluating Online Safety Nudges for Teens. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.
- [7] Zhila Aghajari, Eric PS Baumer, Jess Hohenstein, Malte F Jung, and Dominic DiFranzo. 2023. Methodological Middle Spaces: Addressing the Need for Methodological Innovation to Achieve Simultaneous Realism, Control, and Scalability in Experimental Studies of AI-Mediated Communication. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–28.
- [8] Mamtaj Akter, Jinkyung Katie Park, Campbell Robinson Headrick, Xinru Page, and Pamela J Wisniewski. 2025. Calculating Connection vs. Risk: Understanding How Youth Negotiate Digital Privacy and Security with Peers Online. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–26.
- [9] Naima Samreen Ali, Renkai Ma, Zainab Agha, Jinkyung Katie Park, and Pamela J Wisniewski. 2025. Teens as Co-Researchers: Advocating for Disruptive Change to Engage Youth Meaningfully in Online Safety Research and the Design of Social Media. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–31.
- [10] 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.
- [11] AVPA. [n.d.]. US state age assurance laws for social media. <https://avpassociation.com/us-state-age-assurance-laws-for-social-media/#:~:text=1st%20July%2C%202023,1st%20July%202025>
- [12] Susanne Barth and Menno D.T. de Jong. 2017. The privacy paradox: Investigating discrepancies between expressed privacy concerns and actual online behavior. *Telematics and Informatics* 34, 7 (2017), 1038–1058. <https://doi.org/10.1016/j.tele.2017.04.013>
- [13] Diana Baumrind. 2005. Patterns of parental authority and adolescent autonomy. *New directions for child and adolescent development* 2005, 108 (2005), 61–69.
- [14] BBC News. [n.d.]. Australia is banning social media for teens. Others could follow. <https://www.bbc.com/news/articles/cwyp9d3ddqyo>. Accessed Month Day, Year.
- [15] Leanne Bowler, Eleanor Mattern, and Cory Knobel. 2014. Developing design interventions for cyberbullying: A narrative-based participatory approach. *iConference 2014 Proceedings* (2014).
- [16] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis*. American Psychological Association.

- [17] Victoria ME Bridgland, Benjamin W Bellet, and Melanie KT Takarangi. 2023. Curiosity disturbed the cat: Instagram's sensitive-content screens do not deter vulnerable users from viewing distressing content. *Clinical psychological science* 11, 2 (2023), 290–307.
- [18] Linda Charmaraman, J Maya Hernandez, and Rachel Hodes. 2022. Marginalized and understudied populations using digital media. *Handbook of adolescent digital media use and mental health* (2022), 188–214.
- [19] Deborah Christie and Russell Viner. 2005. Adolescent development. *Bmj* 330, 7486 (2005), 301–304.
- [20] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*. 193–200.
- [21] Katie Davis, Petr Slovak, Rotem Landesman, Caroline Pitt, Abdullatif Ghajar, Jessica Lee Schleider, Saba Kawas, Andrea Guadalupe Perez Portillo, and Nicole S Kuhn. 2023. Supporting Teens' Intentional Social Media Use Through Interaction Design: An exploratory proof-of-concept study. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 322–334.
- [22] Edward L Deci and Richard M Ryan. 2012. Self-determination theory. *Handbook of theories of social psychology* 1, 20 (2012), 416–436.
- [23] Dominic DiFranzo. [n. d.]. The Truman Platform. <https://truman.gitbook.io/the-truman-platform>
- [24] Kenneth A Dodge and Nicki R Crick. 1990. Social information-processing bases of aggressive behavior in children. *Personality and social psychology bulletin* 16, 1 (1990), 8–22.
- [25] Prakriti Dumar, Audrey Flood, and Mahdi Nasrullah Al-Ameen. 2025. One Size Doesn't Fit All: Towards Design and Evaluation of Developmentally Appropriate Parental Control Tool. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.
- [26] Prakriti Dumar, Hanieh Atashpanjeh, and Mahdi Nasrullah Al-Ameen. 2024. "It's hard for him to make choices sometimes and he needs guidance": Reorienting Parental Control for Children. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–51.
- [27] Prakriti Dumar, Audrey Flood, and Mahdi Nasrullah Al-Ameen. 2025. "Synchronized parenting is like mixing oil and water": Reimagining Parental Control for Co-parenting in the Divorced Households. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–35.
- [28] Prakriti Dumar, Bryson D Hackler, Audrey Flood, and Mahdi Nasrullah Al-Ameen. 2024. "I feel like he's looking in the computer world to be social, but I can't trust his judgement": Reimagining Parental Control for Children with ASD. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–25.
- [29] Prakriti Dumar, Jared Perry, and Mahdi Nasrullah Al-Ameen. 2025. Authoritative by Design: Towards Evaluation of Reimagined Parental Control Tool Through Field Study. *International Journal of Human-Computer Interaction* (2025), 1–25.
- [30] Lee B Erickson, Pamela Wisniewski, Heng Xu, John M Carroll, Mary Beth Rosson, and Daniel F Perkins. 2016. The boundaries between: Parental involvement in a teen's online world. *Journal of the Association for Information Science and Technology* 67, 6 (2016), 1384–1403.
- [31] eSafety Commissioner. [n. d.]. Parental controls. <https://www.esafety.gov.au/parents/issues-and-advice/parental-controls> Accessed Month Day, Year.
- [32] eSafety Commissioner. [n. d.]. Safety by Design. <https://www.esafety.gov.au/industry/safety-by-design>
- [33] Anderson Monica Faverio, Michelle and Eugenie Park. [n. d.]. Teens, Social Media and Mental Health. <https://www.pewresearch.org/internet/2025/04/22/teens-social-media-and-mental-health/>
- [34] Michelle Faverio. 2024. Teens, social media and technology 2024. <https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/>
- [35] Álvaro Feal, Paolo Calciati, Narseo Vallina-Rodriguez, Carmela Troncoso, and Alessandra Gorla. 2020. Angel or devil? a privacy study of mobile parental control apps. *Proceedings on Privacy Enhancing Technologies* (2020).
- [36] Jennifer Freeday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods* 5, 1 (2006), 80–92.
- [37] Cristina Fiani, Robin Bretin, Mark McGill, and Mohamed Khamis. 2023. Big buddy: Exploring child reactions and parental perceptions towards a simulated embodied moderating system for social virtual reality. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 1–13.
- [38] Cory Gasaway. [n. d.]. Announcing PlayStation Family App for parental controls and family management. <https://blog.playstation.com/2025/09/10/announcing-playstation-family-app-for-parental-controls-and-family-management/>
- [39] Risa Gelles-Watnick. 2022. Explicit content, time-wasting are key social media worries for parents of US teens. *Pew Research Center* 15 (2022).
- [40] Thomas Grill and Manfred Tscheligi. 2013. The ConWIZ protocol: a generic protocol for wizard of Oz simulations. In *International Conference on Computer Aided Systems Theory*. Springer, 434–441.
- [41] Sijia Guo and Chau-Kiu Cheung. 2023. Social media and adolescents' well-being. In *Healthcare*, Vol. 11. 2297.
- [42] Alexa Guy, Kirsty Lee, and Dieter Wolke. 2017. Differences in the early stages of social information processing for adolescents involved in bullying. *Aggressive behavior* 43, 6 (2017), 578–587.
- [43] Jonathan Haidt. 2024. *The anxious generation: How the great rewiring of childhood is causing an epidemic of mental illness*. Penguin.
- [44] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. Safe sexting: The advice and support adolescents receive from peers regarding online sexual risks. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–31.
- [45] Shabna Ummer Hashim. 2025-10-27. AI Chatbot Lawsuits and Teen Mental Health. https://www.americanbar.org/groups/health_law/news/2025/ai-chatbot-lawsuits-teen-mental-health/
- [46] Eunsook Hyun and J Dan Marshall. 2003. Teachable-moment-oriented curriculum practice in early childhood education. *Journal of Curriculum Studies* 35, 1 (2003), 111–127.
- [47] Andrea Jimenez. [n. d.]. Senators revive bill to ban kids under 13 from social media. <https://www.washingtonpost.com/politics/2025/01/28/senators-revive-bill-ban-kids-under-13-social-media>
- [48] Matt Kaufman. [n. d.]. New Tools for Parents to Personalize Their Child's Experience on Roblox. <https://about.roblox.com/newsroom/2025/04/new-parental-controls-on-roblox>
- [49] JaeWon Kim, Soobin Cho, Robert Wolfe, Jishnu Hari Nair, and Alexis Hiniker. 2025. Privacy as Social Norm: Systematically Reducing Dysfunctional Privacy Concerns on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–39.
- [50] JaeWon Kim, Robert Wolfe, Ramya Bhagirthi Subramanian, Mei-Hsuan Lee, Jessica Colnago, and Alexis Hiniker. 2025. Trust-Enabled Privacy: Social Media Designs to Support Adolescent User Boundary Regulation. *arXiv preprint arXiv:2502.19082* (2025).
- [51] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J Wisniewski, and Munmun De Choudhury. 2021. A human-centered systematic literature review of cyberbullying detection algorithms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–34.
- [52] Loukas Konstantinou and Evangelos Karapanos. 2025. How Do Users Perceive Nudges Against Online Misinformation? A Repertory-Grid Study. *International Journal of Human-Computer Interaction* (2025), 1–19.
- [53] Brunhild Kring. 2018. iGen: Why Today's Super-connected Kids Are Growing Up Less Rebellious, More Tolerant, Less Happy-and Completely Unprepared for Adulthood-and What That Means for the Rest of Us by Jean Twenge. *group* 42, 4 (2018), 363–365.
- [54] Laura Kuenssberg. 2025-11-08. 'A predator in your home': Mothers say chatbots encouraged their sons to kill themselves. <https://www.bbc.com/news/articles/ce3xgwywye4o>
- [55] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. 2010. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)* 10, 2 (2010), 1–31.
- [56] Daniel B Kurz and Spencer Jahng. 2022. Texas' Teens Face a Social Media Ban: A New Start or a Recipe for Destructive Isolation?. In *RAIS Conference Proceedings 2022-2023*. Research Association for Interdisciplinary Studies.
- [57] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [58] Michaela Lebedíková, Vojtěch Mýlek, Kaveri Subrahmanyam, and David Šmahel. 2023. Exposure to sexually explicit materials and feelings after exposure among adolescents in nine European countries: The role of individual factors and social characteristics. *Archives of Sexual Behavior* 52, 1 (2023), 333–351.
- [59] Helen Livingston. [n. d.]. Australia has banned social media for kids under 16. How will it work? <https://www.bbc.com/news/articles/cwyp9d3ddqyo>
- [60] Hiroaki Masaki, Kengo Shibata, Shui Hoshino, Takahiro Ishihama, Nagayuki Saito, and Koji Yatani. 2020. Exploring nudge designs to help adolescent sns users avoid privacy and safety threats. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.
- [61] Teale W Masrani, Jack Jamieson, Naomi Yamashita, and Helen Ai He. 2023. Slow-ing it down: Towards facilitating interpersonal mindfulness in online polarizing conversations over social media. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–27.
- [62] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.
- [63] Adam Mosseri. [n. d.]. Introducing Family Center and Parental Supervision Tools on Instagram and in VR. <https://about.instagram.com/blog/announcements/introducing-family-center-and-supervision-tools>

- [64] Jacqueline Nesi, Sophia Choukas-Bradley, and Mitchell J Prinstein. 2018. Transformation of adolescent peer relations in the social media context: Part 1—A theoretical framework and application to dyadic peer relationships. *Clinical child and family psychology review* 21, 3 (2018), 267–294.
- [65] Jacqueline Nesi, Eva H Telzer, and Mitchell J Prinstein. 2022. *Handbook of adolescent digital media use and mental health*. Cambridge University Press.
- [66] Oluwatomisin Obajemu, Zainab Agha, Farzana A Chowdhury, and Pamela J Wisniewski. 2024. Towards Enforcing Good Digital Citizenship: Identifying Opportunities for Adolescent Online Safety Nudges. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–37.
- [67] Ayla Oden and Lance Porter. 2023. The kids are online: teen social media use, civic engagement, and affective polarization. *Social Media+ Society* 9, 3 (2023), 20563051231186364.
- [68] Amy Orben. 2020. Teenagers, screens and social media: a narrative review of reviews and key studies. *Social psychiatry and psychiatric epidemiology* 55, 4 (2020), 407–414.
- [69] Lisa H Papatraianou, Diane Levine, and Dean West. 2014. Resilience in the face of cyberbullying: An ecological perspective on young people's experiences of online adversity. *Pastoral Care in Education* 32, 4 (2014), 264–283.
- [70] Eugenie Park and Monica Anderson. [n.d.]. About 4 in 10 teens support cellphone bans in classrooms; fewer back all-day restrictions. <https://www.pewresearch.org/short-reads/2026/01/13/about-4-in-10-teens-support-cellphone-bans-in-classrooms-fewer-back-all-day-restrictions/>
- [71] Jinkyung Park, Mamtaj Akter, Naima Samreen Ali, Zainab Agha, Ashwaq Al-soubai, and Pamela Wisniewski. 2025. Towards resilience and autonomy-based approaches for adolescents online safety. *arXiv preprint arXiv:2504.15533* (2025).
- [72] David J Paulsen, Michael L Platt, Scott A Huettel, and Elizabeth M Brannon. 2012. From risk-seeking to risk-averse: the development of economic risk preference from childhood to adulthood. *Frontiers in psychology* 3 (2012), 313.
- [73] 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 (2020), 106347.
- [74] Eyal Peer and Stuart Mills. [n.d.]. From one, many: How can nudges be personalized? *Behavioral Science & Policy* ([n.d.]), 23794607251403327.
- [75] María del Carmen Pérez-Fuentes, María del Mar Molero Jurado, José Jesús Gázquez Linares, Nieves Fátima Oropesa Ruiz, María del Mar Simón Márquez, and Mahia Saracostti. 2019. Self-expressive creativity in the adolescent digital domain: Personality, self-esteem, and emotions. *International Journal of Environmental Research and Public Health* 16, 22 (2019), 4527.
- [76] Elvira Perez Vallejos, Liz Dowthwaite, Helen Creswich, Virginia Portillo, Ansgar Koene, Marina Jirotko, Amy McCarthy, and Derek McAuley. 2021. The impact of algorithmic decision-making processes on young people's well-being. *Health Informatics Journal* 27, 1 (2021), 1460458220972750.
- [77] Anthony T Pinter, Pamela J Wisniewski, Heng Xu, Mary Beth Rosson, and Jack M Carroll. 2017. Adolescent online safety: Moving beyond formative evaluations to designing solutions for the future. In *Proceedings of the 2017 Conference on Interaction Design and Children*. 352–357.
- [78] Rezki Suci Qamaria, Dedi Kuswandi, Ninik Setiyowati, and Akhmedova Mehriyor Bahodirova. 2025. Digital resilience in adolescence: A systematic review of models, methods and theoretical perspectives. *Multidisciplinary Reviews* 8, 9 (2025), 2025287–2025287.
- [79] Afsaneh Razi, Ashwaq AlSoubai, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J Wisniewski. 2023. Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–29.
- [80] Afsaneh Razi, Ashwaq AlSoubai, Seunghyun Kim, Nurun Naher, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J Wisniewski. 2022. Instagram Data Donation: A Case Study on Collecting Ecologically Valid Social Media Data for the Purpose of Adolescent Online Risk Detection. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–9.
- [81] Stephan Schlögl, Gavin Doherty, and Saturnino Luz. 2015. Wizard of Oz experimentation for language technology applications: Challenges and tools. *Interacting with Computers* 27, 6 (2015), 592–615.
- [82] Emily Setty. 2025. Young people and sexual consent: Contextualising 'miscommunication' amid 'grey areas' of ambiguity and ambivalence. *Sex Education* 25, 1 (2025), 140–155.
- [83] John Snakenborg, Richard Van Acker, and Robert A Gable. 2011. Cyberbullying: Prevention and intervention to protect our children and youth. *Preventing School Failure: Alternative Education for Children and Youth* 55, 2 (2011), 88–95.
- [84] Laurence Steinberg and Susan B Silverberg. 1986. The vicissitudes of autonomy in early adolescence. *Child development* (1986), 841–851.
- [85] Lucille Alice Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.
- [86] Elizabeth A Sweigart, Aahil Valliani, and Pamela J Wisniewski. 2025. Pause, Reflect, and Redirect: An Approach to Empowering Youth to Be Safer Online by Helping Them Make Better Decisions. *Social Sciences* 14, 5 (2025), 302.
- [87] Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N Bazarova. 2019. Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.
- [88] Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
- [89] Kathleen Van Royen, Karolien Poels, Heidi Vandebosch, and Philippe Adam. 2017. "Thinking before posting?" Reducing cyber harassment on social networking sites through a reflective message. *Computers in human behavior* 66 (2017), 345–352.
- [90] Emily A Vogels. 2022. Teens and cyberbullying. *Pew Research Center, December* 15 (2022).
- [91] Joseph B Walther. 1992. Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication research* 19, 1 (1992), 52–90.
- [92] Joseph B Walther. 2008. Social information processing theory. *Engaging theories in interpersonal communication: Multiple perspectives* 391 (2008).
- [93] Ge Wang, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2023. 12 Ways to empower: Designing for children's digital autonomy. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–27.
- [94] Jasper Ward. [n.d.]. New York to require social media platforms to display mental health warnings. <https://www.reuters.com/legal/litigation/new-york-require-social-media-platforms-display-mental-health-warnings-2025-12-26/>
- [95] Pamela Wisniewski. 2018. CAREER: Safety by Design: Protecting Adolescents from Online Risks. Available at SSRN 5421934 (2018).
- [96] 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.
- [97] Pai-Lu Wu and Wen-Bin Chiou. 2009. More options lead to more searching and worse choices in finding partners for romantic relationships online: An experimental study. *CyberPsychology & Behavior* 12, 3 (2009), 315–318.
- [98] Angela Yang. 2024-10-23. Lawsuit claims Character.AI is responsible for teen's suicide. <https://www.nbcnews.com/tech/characterai-lawsuit-florida-teen-death-rcna176791>
- [99] Seraphina Yong, Ashlee Milton, Evan Suma Rosenberg, Stevie Chancellor, and Svetlana Yarosh. 2025. 'I'm Petting the Laptop, Which Has You Inside It': Reflecting on Lived Experiences of Online Friendship. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–33.
- [100] Jina Yoon, Amy X Zhang, and Joseph Seering. 2025. "It's Great Because It's Ran By Us": Empowering Teen Volunteer Discord Moderators to Design Healthy and Engaging Youth-Led Online Communities. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–30.
- [101] Dorothy Zhao, Mikako Inaba, and Andrés Monroy-Hernández. 2022. Understanding Teenage Perceptions and Configurations of Privacy on Instagram. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.
- [102] Marc A Zimmerman. 2013. Resiliency theory: A strengths-based approach to research and practice for adolescent health. , 381–383 pages.
- [103] Maximilian Zinkus, Oliver Curry, Marina Moore, Zachary Peterson, and Zoë J Wood. 2019. Fakesbook: A social networking platform for teaching security and privacy concepts to secondary school students. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. 892–898.

Appendix

A Participant Demographic Information

Table 4: Participant Demographic Information

Demographics		%
Age	16-19	63%, N = 12
	13-15	37%, N = 7
Gender	Female	58%, N = 11
	Male	42%, N = 8
Social Media Usage	Instagram	68%, N = 13
	TikTok	21%, N = 4
	Youtube	16%, N = 3
	Snapchat	11%, N = 2
Race	South Asians	47%, N = 9
	East Asians	32%, N = 6
	Black/African Americans	11%, N = 2
	Hispanic/Latino	11%, N = 2
Sexual Identity	Heterosexual	68%, N = 13
	LGBTQ+	32%, N = 6
Self-identified as Marginalized	No	68%, N = 13
	Yes	32%, (N = 5 Race, N = 1 Sexual Identity)
Living Situation	Dual-Parent Household	79%, N = 15
	Single-Parent Household	16%, N = 3
	On their own	5%, N = 1
Mother's Education Background	Bachelor's	58%, N = 11
	Graduate Degree	32%, N = 6
	High School or less	5%, N = 1
	Prefer not to disclose	5%, N = 1
Fathers's Education Background	Graduate Degree	74%, N = 14
	Bachelor's	16%, N = 3
	High School or less	5%, N = 1
	Prefer not to disclose	5%, N = 1

B Teens Nudge Responses by Risk Type (RQ2)

Table 5: Teens' Behavioral Responses for the Cyberbullying Nudge

Nudge Choice	Confirm Unhide/ Unhide	Final Choice		
		Flag and Delete	Ignore	Self Response
Unhide (42%, n=8)	42%, n=8	21%, n=4	16%, n=3	5%, n=1
OK (26% , n=5)	26%, n=5	11%, n=2	11%, n=2	5%, n=1
Learn More (16%, n=3)	11%, n=2	5%, n=1	11%, n=2	N/A
Close (11%, n=2)	11%, n=2	11%, n=2	N/A	N/A
Protective Action (5%, n=1)	n=1	5%, n=1	N/A	N/A
Total	95%, N=18	53%, n=10	37%, n=7	11%, n=2

Table 6: Teens' Behavioral Responses for the Info Breach Nudge

Nudge Choice	Secondary Choice
Protective Action (16%, n=3)	Block (10%, n=2)
	Report (10%, n=2)
	Notify (10%, n=2)
OK (68%, n=13)	Self-Response (32%, n=6)
	No Response (16%, n=3)
	Suggested Message (21%, n=4)
Close (10%, n=2)	Self-Response (10%, n=2)
Turn off Warning (n=1)	N/A

Table 7: Teens' Behavioral Responses for the Explicit Content Nudge

Nudge Choice	Secondary Choice
Protective Action (42%, n=8)	Block (32%, n=6)
	Report (26%, n=5)
	Hide Similar Messages (26%, n=5)
	Delete (16%, n=3)
	Notify Others (16%, n=3)
OK (36%, n=7)	No Response (21%, n=4)
	Respond Back (16%, n=3)
Close (10%, n=2)	Self-Response (10%, n=2)
Turn off Warning (10%, n=2)	N/A

C Teens Overarching Nudge Preferences (RQ3)

Table 8: Codebook for Teens' Overarching Preferences for Nudges

Themes	Preferences and Descriptions
Valuing Nudges for Proactive Safety and Learning	• Immediate Protection (95%, n=18): Teens felt protected by nudges/filters (84%, n=16), valued having multiple options (37%, n=7), felt simulated platform as safer than typical social media (58%, n=11), and anticipated using nudges on other platforms (68%, n=13). • Educational Value (89%, n=17): Nudges promoted long-term behavior change (68%, n=13), offered greater benefits for younger teens (58%, n=11), and needed to account for prior safety knowledge (26%, n=5). • Broader Impact (21%, n=4): Some preferred nudges that also support offender training (21%, n=4).
Improving Trust and Explainability for Nudges	• Distrust in AI Accuracy (74%, n=14): Teens distrusted AI-based detection (47%, n=9) and worried about misuse of notify-others (26%, n=5). • Hybrid and Transparent Systems (74%, n=14): They recommended human/AI hybrid review (42%, n=8), wanted clearer explanations (37%, n=7), and preferred specific violation details (21%, n=4). • Tailoring AI Risk Detection (74%, n=4): Teens wanted more details in high-risk cases (21%, n=4), preferred non-disruptive guidance in low-risk cases (32%, n=6), and wanted familiar safety terms (16%, n=3).
Personalization and Collective Safety	• Personal Control (42%, n=8): Teens wanted customized nudges that appear earlier (10%, n=2), and provide accountability for repeated offenses (32%, n=6). • Collective Safety (63%, n=12): They valued community notifications (57%, n=11) but preferred selective or anonymous sharing (37%, n=7).

D Chat Scripts and Personas

Table 9: Non-risky personas and their description

Non-Risky Personas	Description
Artsy Introverted Teen	Sarah is an introverted girl involved in arts and music. She limits giving out a lot of personal information and is usually shy with new people but keeps a conversation going by mostly talking about her hobbies without asking too many personal details.
Sporty Introverted Teen	Frank is a reserved sporty guy and does not post much on social media. He rarely starts a conversation with people and is good at making small talk with people online.
Friendly Extroverted Teen	Alice is a smart extroverted teen, interested in reading and writing. She has a likable personality, is supportive, and makes aesthetic posts about nature and poetry.

The following chats served as guidelines for researchers who acted as risky users. Minimal improvisations were applied based on participant responses.

Table 10: Sample private conversation guidelines by persona with nudge triggers in bold

Risky Persona	Script
Bryan Cruz	• <i>Bryan</i> : Hey hey! how's it going? • [wait for response or go ahead] • <i>Bryan</i> : How are u liking the app so far • [wait for response] • <i>Bryan</i> : Yea I'm still figuring it out • <i>Bryan</i> : Nice to see a few ppl online though • <i>Bryan</i> : Oh I saw ur post it looks cool! • [wait for response or go ahead] • <i>Bryan</i> : Where are u joining from? • [wait for response] • <i>Bryan</i> : Nice, I have a few cousins there • <i>Bryan</i> : I'm at Oakridge High • <i>Bryan</i> : Which school do you go to? • <i>Bryan</i> : Btw do u wanna stay connected after the study? • <i>Bryan</i> : I was thinking.. • <i>Bryan</i> : Can we exchange numbers?
Dave Fisher	• <i>Dave</i> : Hi there • <i>Dave</i> : I'm Dave, nice to connect with you • [wait for response] • <i>Dave</i> : Yea I'm still figuring it out • <i>Dave</i> : I'm having a bit of a hard time on here, how abt u? • [wait for response] • <i>Dave</i> : Just struggling a bit with the tasks, can u guide me? • [wait for response, ask help according to response] • [wait for response] • <i>Dave</i> : Thanks so much!! You're awesome • <i>Dave</i> : Ok finally I'm gettin the hang of it haha • <i>Dave</i> : All thanks to u! • <i>Dave</i> : Btw loved your post • <i>Dave</i> : I'm thinking of posting this, what do u think • <i>Dave</i> : [sends image]