

Analyzing Social Media Claims regarding Youth Online Safety Features to Identify Problem Areas and Communication Gaps

RENKAI MA*, University of Cincinnati, USA

DOMINIQUE GEISSLER*, LMU Munich, Germany and Munich Center for Machine Learning (MCML), Germany

STEFAN FEUERRIEGEL, LMU Munich, Germany and Munich Center for Machine Learning (MCML), Germany

TOBIAS LAUINGER, New York University, USA

DAMON MCCOY, New York University, USA

PAMELA J. WISNIEWSKI, International Computer Science Institute, USA

Social media platforms have faced increasing scrutiny over whether and how they protect youth online. While online risks to children have been well-documented by prior research, how social media platforms communicate about these risks and their efforts to improve youth safety have not been holistically examined. To fill this gap, we analyzed $N = 352$ press releases and safety-related blogs published between 2019 and 2024 by four platforms popular among youth: YouTube, TikTok, Meta (Facebook and Instagram), and Snapchat. Leveraging both inductive and deductive qualitative approaches, we developed a comprehensive framework of seven *problem areas* where risks arise, and social media platforms claim to address these risks through various online safety features. Our analysis revealed uneven emphasis across problem areas, with most communications focused on Content Exposure and Interpersonal Communication, whereas less emphasis was placed on Content Creation, Data Access, and Platform Access. Additionally, we identified three problematic communication practices related to their described safety features, including discrepancies between feature implementation and availability, unclear or inconsistent explanations of safety feature operation, and a lack of evidence regarding the effectiveness of safety features in mitigating risks once implemented. Based on these findings, we discuss the communication gaps between risks and the described safety features, as well as the tensions in achieving transparency in platform communication. Our analysis of platform communication informs guidelines for responsibly communicating about youth safety features.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**.

Additional Key Words and Phrases: youth online safety, social media, safety feature, communication practice

ACM Reference Format:

Renkai Ma, Dominique Geissler, Stefan Feuerriegel, Tobias Lauinger, Damon McCoy, and Pamela J. Wisniewski. 2026. Analyzing Social Media Claims regarding Youth Online Safety Features to Identify Problem Areas and Communication Gaps. *Proc. ACM Hum.-Comput. Interact.* 10, 2, Article CSCW005 (April 2026), 31 pages. <https://doi.org/10.1145/3788041>

*Both authors contributed equally to this research.

Authors' Contact Information: Renkai Ma, University of Cincinnati, Cincy, OH, USA, mark@ucmail.uc.edu; Dominique Geissler, LMU Munich, München, Germany and Munich Center for Machine Learning (MCML), München, Germany, d.geissler@lmu.de; Stefan Feuerriegel, LMU Munich, München, Germany and Munich Center for Machine Learning (MCML), München, Germany, feuerriegel@lmu.de; Tobias Lauinger, New York University, New York City, NY, USA, lauinger@nyu.edu; Damon McCoy, New York University, New York City, NY, USA, mccoy@nyu.edu; Pamela J. Wisniewski, International Computer Science Institute, Berkeley, CA, USA, pwisniewski@icsi.berkeley.edu.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

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

ACM 2573-0142/2026/4-ARTCSCW005

<https://doi.org/10.1145/3788041>

1 Introduction

In recent years, protecting youth¹ online has become a major focus of regulatory efforts in the United States (U.S.) [39, 40, 42, 53] and platform governance [82, 97, 133]. For example, in 2019, the U.S. Federal Trade Commission (FTC) fined YouTube and Musical.ly (now TikTok) for violating the Children's Online Privacy Protection Act (COPPA), which prohibits collecting personal information from children under 13 without parental consent [40, 41]. Since then, U.S. lawmakers have introduced new bills, such as the Kids Online Safety Act (KOSA), that seek to hold platforms accountable for protecting minors from online risks such as cyberbullying, data misuse, and the negative impacts of algorithmic content recommendations on social media [107].

Indeed, protecting youth online has been an ongoing research effort for over a decade. The foundational 4Cs risk framework, developed by Livingstone et al. [80], has shaped the field of youth online safety (e.g., [102, 109]) by categorizing online risks to children into four over-arching categories: *content* (exposure to harmful or inappropriate material), *contact* (unwanted or risky interactions with adults), *conduct* (youths' own risky behavior online), and *contract* (risks of data and commercial practices). The 4Cs framework, which has been updated over the years, has provided a shared taxonomy for understanding the risks that youth face online and has been widely adopted by HCI and CSCW researchers to understand online risks and (re)design safety solutions for youth (e.g., [2, 13]), such as evaluating parental control apps [126] and co-designing new safety nudges with teens to address privacy breaches on social media [91].

Besides these *external* safety mechanisms, HCI and CSCW researchers have recently begun examining safety features built *internally* into social media platforms (i.e., built-in features). Much of this emerging work investigates user experiences with specific built-in features on individual platforms, such as parental content filters [33, 34] and privacy measures [70, 82] or envisioning new built-in safety features [2, 3]. To the best of our knowledge, our work is the first to investigate social media platforms' communication about built-in youth safety measures across different types of features and across multiple platforms. While the 4Cs framework [78, 80] categorizes online risks that youth encounter, an analogous taxonomy for categorizing safety measures does not currently exist, which is a notable gap given that children use multiple social media platforms every day [9, 30]. Therefore, understanding the contexts where the known online risks arise to youth and where platforms claim they apply safety features requires a better structure to supplement the current risk-centric 4Cs framework, and we conceptualize these contexts as *problem areas*.

Industries (e.g., the tobacco industry [11, 92, 93]) that are confronted with public concerns about risks in their products tend to employ various communication practices to minimize these risks and thus avoid losing customers or lawsuits. Similarly, social media platforms often describe and market their safety features through public communication channels, such as press releases and help blogs (e.g., [61, 115, 134]) to address online risks for youth. Given that end-users (youth, parents, or educators) rely on these communications, essentially as corporate communication [19, 51], to understand and utilize safety features, their clarity, consistency, and actionability are important. From this standpoint, we aim to evaluate the extent to which platform communications especially *misalign* with well-documented online risks (such as those categorized by the 4Cs framework) or the described reality of their features, including feature implementation and effectiveness. When these communications misalign with risks or present inconsistent information, they might hinder user understanding, create a false sense of security among users, including youth, parents, policymakers, or researchers, and impede the effective adoption of safety features in certain problem areas. Thus, we ask two questions to characterize problem areas and uncover issues in communication practices:

¹In this paper, we use "youth" to refer to children in different age groups, including under 13, teens, and tweens, and we use the terms children and youth interchangeably.

- **RQ1:** *What problem areas characterize the contexts where the known risks arise to youth on social media platforms, and where platforms claim to apply safety features?*
- **RQ2:** *What are the problematic communication practices that exist within platform communications regarding youth safety features?*

To answer these two research questions, we first selected the four most popular social media platforms among youth in the U.S.: Meta (Instagram and Facebook²), YouTube, Snapchat, and TikTok, and then adopted both inductive and deductive qualitative approaches to collect and analyze $N = 352$ press releases and safety-related articles published between 2019 and 2024, which are public channels through which social media platforms formally communicate about their safety features with users.³ We inductively coded the articles and then deductively mapped out the described safety features to the known risks to develop a taxonomy of safety features around the identified problem areas. Here, we extended the established 4Cs risk framework to 5Cs and added circulation risk as a new type of risk to cover emerging cases where children's content is inappropriately amplified or propagated on social media platforms.

We uncovered an uneven distribution of platform communications across problem areas, where most of the communications focus on the Content Exposure and Interpersonal Communication, while focusing less on Platform Access, Monetization, and Content Creation areas (RQ1). Through reflexive thematic analysis [24] on our collected articles, we identified three problematic practices within a recognizable lifecycle of platforms' communication about youth online safety, including (1) discrepancies between platform implementation claims and feature availability, (2) lack of clarity and consistency in explaining feature operation, and (3) lack of clarity and evidence in demonstrating effectiveness of risk mitigation (RQ2). These findings motivate us to discuss the gaps between risks and the described safety features, as well as the tensions in achieving transparency in platform communication practices. We further discuss the communication implications of the alignment and contrast between the design philosophies suggested by current platform communication practices and those supported by prior research on youth online safety and regulatory considerations. We then lay out guidelines for responsible platform communications regarding youth safety features.

Our paper makes four primary contributions. First, we extend the established 4Cs risks framework to 5Cs to cover emerging cases of *circulation risk*. Second, we introduce the concept and framework of *problem areas* to characterize contexts where youth might encounter known risks and where platforms claim to apply safety features to address those risks. This framework supplements the current understanding of online risks by addressing the relative lack of work in taxonomizing and categorizing online safety features for youth, especially those built internally into social media platforms. Third, to the best of our knowledge, our study is the first to both identify what safety features the most popular social media platforms claim to implement and to examine the communication practices by social media platforms around those features as part of their youth online safety efforts. By identifying problematic practices in this communication, our work aligns with the current trend in HCI and CSCW toward evaluating platforms' safety features (e.g., [2, 13, 25, 91]). Lastly, based on our analysis of platform communication, we outline both communication implications for youth online safety designs and communication guidelines for responsibly communicating about safety features.

²While we recognize that Facebook and Instagram might have different features and user bases, for the purposes of this initial selection, we grouped them under the parent company, Meta, due to shared ownership. Also, in later data analysis, we found that there were not many articles specifically discussing features exclusive to Facebook, as most situations are relevant to both platforms, further supporting this grouping.

³The dataset can be found in a GitHub repository: https://github.com/DominiqueGeissler/youth_online_safety_communication

2 Background

We review how prior work examines online risks to children and safety features that address those risks on social media platforms. Like many industries facing public scrutiny of risks in their products, social media platforms respond to it through different communication practices to retain users and avoid lawsuits. Given the limited work assessing these practices, our study is the first to fill this gap. Finally, we introduce the overview of safety products on social media platforms.

2.1 Youth, Social Media Use, and Online Risks

Social media platforms like YouTube, TikTok, Instagram, and Snapchat have seen a notable increase in usage among U.S. youth, according to the Pew Research Center [9, 38]. Several national surveys also indicate that using platforms such as YouTube or TikTok has been the most popular media activity among eight- to eighteen-year-olds [30, 89]. These platforms offer various benefits for children: YouTube supports children's informal learning [52] and self-expression through content creation and sharing [84, 104]. Also, TikTok immerses youth in contemporary youth culture and helps them form online communities [143].

However, youth also encounter a variety of risks on social media platforms. Teens might face harmful content (e.g., violent content [78]) and cyberbullying [10, 106], or become victims of privacy breaches [5, 83]. Researchers have found that younger children, specifically those under 13, lack technical literacy and perceive data privacy risks in simple terms, such as "stranger danger" [70]. They have limited awareness of algorithmic practices like data tracking or targeted ads, and often feel powerless to control data collection by social media platforms [127, 140].

Livingstone et al. have helped categorize these diverse risks into a 4Cs framework, comprising of *content*, *contact*, *conduct*, and *contract* risks [78, 80]. Content risks refer to exposure to harmful material, such as violent, hateful, or sexual content that is age-inappropriate. Contact risks describe harmful adult-initiated interactions, such as grooming, harassment, or sexual exploitation. Conduct risks involve negative peer interactions among children (e.g., bullying, hateful behaviors). Lastly, contract risks concern exploitation through commercial practices (e.g., scams, targeted ads).

The 4Cs framework has been widely adopted by researchers to categorize online risks to youth on social media platforms (e.g., [71, 102, 139]). For example, prior work found that contact risks emerge on Instagram, where teens report frequent sexual solicitation and harassment that increases if they ignore the solicitations [32]. Meanwhile, contract risks can emerge from social media platforms' commercial practices. For example, YouTube's opaque, ad-driven recommendation algorithms have been found to exploit younger children through misleading product promotion videos [21, 37] and attract them to watch such videos for extended periods [22, 31]. To address these different risks, CSCW research, such as Wang et al. [126], has applied the 4Cs framework to evaluate safety mechanisms like parental control apps, whereas Agha et al. [2] and Obajemu et al. [91] have co-ideated and designed online safety interventions together with teens. Our study thus aims to build on the 4Cs framework to understand and categorize the known risks to youth on social media.

2.2 Mapping Online Risks to Safety Measures to Protect Youth on Social Media

Recent HCI and CSCW work has acknowledged the role of safety measures or features in addressing online risks for youth, while much of this work has focused on (re)designing them. Researchers, for example, have engaged teens in co-designing privacy management tools to address privacy breaches from other users or the social media platforms themselves [2, 14, 128]. Another line of work has explored using in-situ social media data donations from teens to develop machine learning systems that identify various conduct and content risks [6, 97].

Different from these *external* safety measures (e.g., parental control apps typically not developed by social media companies [126, 130]), researchers have recently increasingly examined user experiences with safety features built *internally* into social media platforms (i.e., *built-in* features). For example, Dumarú et al. explored [34] and redesigned [33] parental controls on YouTube, enabling parents to limit certain content categories (e.g., war, drugs, alcohol), while further uncovering that these controls reduced exposure to inappropriate content and further promoted parent-child discussions. On Instagram, teens report dissatisfaction with built-in reporting tools, criticizing their ineffectiveness in addressing harassment and misinformation [2, 3]. Similarly, Edelson experimented with algorithmic feed recommendations on Instagram Reels and TikTok for both 13- and 23-year-old female personas, finding similarly sexually suggestive content on Reels for both personas, while TikTok showed less explicit content to the teen than the adult [35].

Despite the user experience with individual built-in safety features, a step back is needed to holistically understand what safety features are available, as claimed by the platforms. While the 4Cs framework [78, 80] helps categorize online risks, an analogous taxonomy of safety measures for assessing how platforms claim, or tend, to address youth online safety does not currently exist, particularly given that children use multiple platforms daily [9, 30]. Our study thus addresses this gap by providing a taxonomy of safety features along the *problem areas* to describe contexts, where risks arise, and where platforms claim to implement the features to address those risks.

2.3 Communication Practices of Social Media Platforms for Youth Online Safety

Industries confronted with rising public awareness of risks in their products tend to perform various communication practices to minimize or deflect these risks and thus not lose customers or lawsuits. For instance, the tobacco and e-cigarette industries have historically used terms like “light” or “safer alternatives” to downplay health risks, fostering consumer perceptions of reduced harm and increasing product appeal [11, 92, 93]. Similarly, communication strategies emphasizing healthfulness or utility for quitting cigarettes correlate with higher youth adoption of e-cigarettes [29, 95]. These parallels reflect how companies try to use their communication to align business goals with public well-being [96]. However, such communication often triggers skepticism [101] when companies’ one-way communication prioritizes publicity over dialogue [47].

As related work on online risks to children has highlighted, social media platforms now also face challenges regarding the perceived safety of their platforms. These platforms have implemented (what they claim are) safety measures, which they market or communicate to users through press releases and help blogs (e.g., [61, 115, 134]). They begin marketing features directly to users, such as Instagram promoting its Teen Accounts via TV and podcasts [16], and TikTok promoting its reporting tools to prevent harmful behavior during international football games [122]. However, discrepancies often surface between platform claims about these features and their effectiveness in protecting children. For example, a Kentucky lawsuit alleges TikTok’s screen time controls were designed primarily for public relations rather than reducing teens’ usage [12]. Instagram’s parental supervision options also face criticism for assuming all parents can navigate social media’s complexities, which is not always true [17].

To the best of our knowledge, our study is the first to examine how platforms communicate about their built-in safety features, particularly around youth online safety. Filling this gap is critical, as platforms’ communication strategies shape public perception, user adoption, and, ultimately, youth online safety. Therefore, our analysis focuses on how platforms communicate safety measures across risk categories and includes a critical examination of their communication practices.

2.4 Historical Overview of Safety Products on Social Media Platforms

Over the past decade, major social media platforms have launched various safety products to address online risks for youth. YouTube launched YouTube Kids in 2015 as a standalone, curated platform for younger audiences. In 2019, TikTok introduced TikTok for Younger Users, a separate app experience designed for children under 13 with limited functionality and stricter content controls. More recently, both Meta and Snapchat rolled out Family Center tools in 2022, providing parents with dashboards to monitor youth activity and adjust privacy or content settings.

Figure 1 provides an overview of these products over time. While these initiatives reflect efforts to respond to public and regulatory pressure, they also serve as strategic communication efforts to position platforms as proactive and responsible. Our findings build on this context by analyzing how such efforts are described at a more granular level, through the safety features they claim to address online risks, and how they communicate about their safety features.

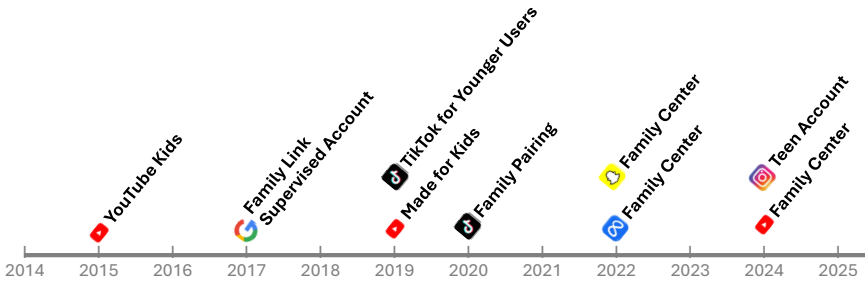


Fig. 1. Historical account of overarching youth safety products that social media platforms include in their communications. We included Google here, as YouTube claims to use Google products to safeguard youth online.

3 Methods

This section details the collection and preparation of our dataset about social media platform communications about youth online safety. We further introduce our qualitative coding (RQ1) and reflexive thematic analysis (RQ2) to answer our RQs.

3.1 Data Collection

We collected data on how social media platforms communicate about youth online safety in four steps, including Step 1: We selected four representative platforms: YouTube (including YouTube Kids), Meta (representing Instagram and Facebook), TikTok, and Snapchat. These were chosen because they are the most used platforms by youth across age groups in the U.S. [9, 113].

Step 2: We collected data directly from these platforms to identify the safety features each claims to offer, focusing on official press releases (Source 1) and Help or Safety Center blogs (Source 2) from the platforms themselves. Press releases, sourced from official newsrooms like TikTok Newsroom⁴ or Snapchat Newsroom⁵, regularly announce plans, launches, or updates for built-in safety features. We searched these newsrooms for relevant press releases (see Table 1, Source 1). Concurrently, platform Help or Safety Centers, such as Meta's Safety Center⁶, publish blogs about

⁴<https://newsroom.tiktok.com/en-us>

⁵<https://newsroom.snap.com/>

⁶<https://about.meta.com/actions/safety>

using their safety features. We collected these blogs across the centers using the platforms’ search function with keywords: youth OR child OR teen OR minor OR underage OR age (see Table 1, Source 2). Both data sources are the platforms’ official materials for their communication about safety features to end-users⁷.

To collect these articles and blogs, we used Instant Data Scraper⁸, a Chrome extension. We focused on articles published from January 1, 2019, to December 31, 2024; this period was chosen because 2019 marked the start of major legal settlements between regulators and social media companies (e.g., the U.S. FTC fined YouTube and TikTok for child privacy law violations in 2019 [40, 41]). Data collection ended on December 31, 2024. After removing pre-2019 articles, we obtained a dataset of $N = 1,952$ press release articles. While press releases (Source 1) typically included timestamps, many Safety/Help Center blogs (Source 2) did not, so we used the Internet Archive’s Wayback Machine⁹ to approximate the publication year of each blog by determining its earliest archived appearance and removing any blogs with an earliest archived appearance before 2019. This yielded $N = 376$ blogs.

Table 1. Data sources and article counts for each platform: press releases (Source 1) and Safety/Help Center blogs (Source 2). Please note that Instagram and Facebook belong to Meta overall, but we gathered their materials separately.

Platform	Source 1		Source 2		Article Count After Relevance Coding	Total Count After Fetching Article Hyperlinks
	Press Release	Article Count	Safety/Help Center	Blog/Article Count		
YouTube	https://blog.youtube/news-and-events/	379	https://support.google.com/youtube#	204	69	134
TikTok	https://newsroom.tiktok.com/en-us	886	https://www.tiktok.com/safety/en/	30	67	74
Snapchat	https://newsroom.snap.com/	295	https://values.snap.com/safety/safety-center	33	36	45
Meta	https://about.instagram.com/blog , https://about.fb.com/news/	392	https://help.instagram.com/ , https://about.meta.com/actions/safety	109	99	99
Total		1,952		376	271	352

Step 3: We performed *relevancy coding* to determine which articles and blogs are relevant to our research questions. We filter our dataset using three inclusion criteria: mentions of (1) youth (e.g., teens, children, minors), (2) online safety, and (3) technical features or implementations. These criteria are applied to each article, and we remove all articles that do not fit the criteria from the dataset. For example, we excluded articles like TikTok’s “Songs of the Summer 2024”, Instagram’s wildlife exploitation post, and YouTube’s “Get started with channel memberships”, as their topics were unrelated to youth online safety features. To apply these criteria, the dataset was divided between two coders. One coder analyzed the articles from YouTube and Meta, while the second coder analyzed those from TikTok and Snapchat. To validate coding and the consensus of identified relevant articles, we randomly selected five press releases and five Safety/Help Center blogs per platform (~20% of data) to be coded by a second reviewer (the other reviewer, respectively) and

⁷While our focus is on the direct-to-user communications found in press releases and help/safety centers, we recognize that platforms also produce regulatory documents like transparency reports or risk assessments for laws such as the Digital Services Act (DSA). We intentionally excluded these documents from our analysis to maintain a clear focus on the materials created for and consumed by the general public, including youth and their caregivers.

⁸<https://chromewebstore.google.com/detail/instant-data-scraper/ofaokhiedipichpaobibbnahnkdoiiah>

⁹<https://web.archive.org/>

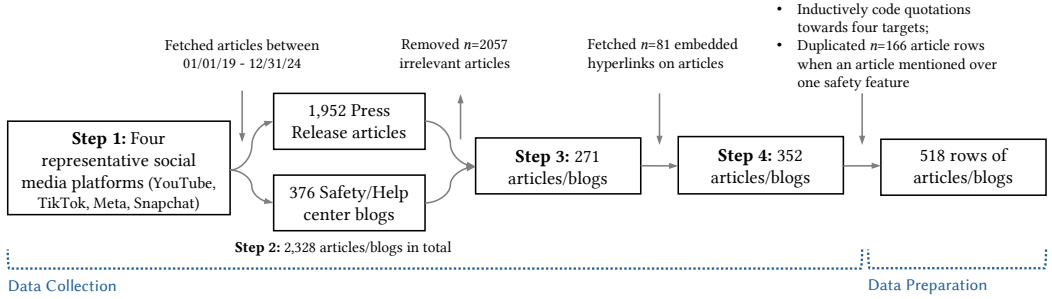


Fig. 2. Flowchart of our data collection and preparation.

calculated Cohen’s kappa. Inter-rater reliability (IRR) [28] reached substantial (0.61–0.80) or almost perfect (0.81–1) agreement. For press releases, IRR was: (1) about youth: 98% agreement rate, $\kappa = 0.953$; (2) about online safety: 98%, $\kappa = 0.958$; and (3) mention of technical implementation or features: 88%, $\kappa = 0.689$. For blogs, the values were (1) 96%, $\kappa = 0.9$; (2) 100%, $\kappa = 1$; and (3) 100%, $\kappa = 1$. The two coders resolved disagreements through weekly meetings and joint reviews, producing a final set of $N = 271$ relevant articles (see Table 1).

Step 4: We enriched our dataset through snowball sampling by following embedded hyperlinks in the relevant articles, similar to a backward snowball sampling approach in systematic literature review [132]. To do so, we went through each relevant article and manually clicked the hyperlinks in it. We included hyperlinks if they are press releases or help and safety center articles of the social media platform. We reapplied the relevancy criteria again and kept those that are about youth, online safety, and a technical feature or implementation. We repeated the snowball sampling up to three levels deep. This way, we collected 81 additional articles for our dataset. This step was done to ensure that we capture as many relevant articles from the social media platforms as possible, because relying solely on our keyword search, we might have missed relevant articles. Our final dataset contained $N = 352$ relevant articles.

3.2 Data Preparation

After collecting the set of relevant articles, we prepared an analytic dataset that could be used to answer our research questions. To do so, we applied two rounds of qualitative analysis [23] to generate codes for our data analysis (both are detailed below). In the first round, we inductively extracted quotations of safety features described in platform communications and coded these into standardized safety features. These form the unit of analysis in our dataset. In this context, “codes” refer to extracted units of meaning, such as safety features and affected age groups of the features. In the second round, we deductively mapped these features into the 5Cs risk framework. These coding rounds produced the feature-level dataset that underpins our subsequent analyses: RQ1 draws on the standardized safety features, while RQ2 draws on the quotations.

3.2.1 First-Round Coding of Safety Features (Inductive Approach). The objective of the first round of coding was to inductively extract and structure how platforms communicated safety features in their own articles. Two coders read each relevant article and identified quotations describing safety features. We added these into our dataset as a *quotation column*. Next, for each quotation, we extracted four items into new dataset columns: (1) the *described safety feature*, (2) the *affected age group* by the feature, (3) the *controlling entity* (i.e., who uses or controls the feature), (4) any *described sub-feature(s)* (if the primary feature was composite). We oriented our codes to the

platform's original quotation verbatim throughout this inductive coding. For example, from the YouTube quote, "A YouTube Kids profile is a profile within a parent's Google Account. Profiles are available only when the parent is signed in to YouTube Kids with their Google Account," we coded: (1) "individual kid profile" as the described safety feature, (2) "children under 13" as the affected age group, (3) "parents" as the controlling entity, (4) no additional sub-features. This inductive round generated an initial dataset of 124 feature codes.

Next, we standardized the wording of the described safety features to establish a uniform categorization of them across the four platforms. For example, Meta's feature involving technology to detect user age and YouTube's feature requesting additional proof of age were both consolidated under the standardized feature category "Age Verification." This standardization process consolidated an initial 124 feature codes into 30 codes across the platforms. This initial data analysis resulted in a dataset of $N = 518$ rows, with some articles or blogs repeated due to the feature-level analysis (see the final stage in Figure 2).

3.2.2 Second-Round Mapping to Risk Categories (Deductive Approach). In the second round, the same two coders deductively mapped the described sub-features into the 5Cs risk framework as five new columns in our dataset: *content*, *contact*, *conduct*, *contract*, and *circulation*. We initially used the 4Cs risk framework [75, 77, 80]. However, as we analyzed the data, particularly considering the rise of youth creating and sharing content on platforms like YouTube and TikTok [84, 104, 143], we noticed described features restricting the sharing, downloading, or propagation of children's content (i.e., content by child creators or content featuring children) beyond the intended audience and in potentially harmful ways. This led us to identify an emergent risk category as *circulation* risk and extend the 4Cs framework to 5Cs to capture these cases.

3.3 Data Analysis

3.3.1 Analysis for RQ1: Taxonomizing Safety Features Around Problem Areas. For RQ1, our goal was to understand the contexts on social media platforms where risks can arise for youth and where platforms communicate that they apply safety measures. To analyze this relationship, we developed a framework of *problem areas*. This was motivated by the need to bridge the gap between existing risk-centric frameworks (like the 4Cs or 5Cs in our study, which categorize what risks exist) and feature-specific work (e.g., experiences with individual built-in safety features [2, 34, 70]).

Our analysis for RQ1 involved two main steps. First, we grouped the 30 described safety features we coded into seven problem areas, which are defined as contexts where risks can arise for youth on social media platforms. Safety features were assigned to a problem area based on which area of user activity they affect. For example, we conceptualized the area of Interpersonal Communication, which covers user activities like messaging and commenting, as a problem area because these activities can expose youth to risks such as online bullying, harassment, or even exploitation. Platforms state that they apply features like Message Restrictions and Connection Controls to mitigate the risks within this area. In addition, we identify a cross-cutting layer of safety features that enable parents to monitor and control their children's activity across the problem areas. We call this layer *Parental Supervision* and group features that involve parents in this additional layer. While we recognize that the list of 30 safety features and the five known risks might not be exhaustive, they were instrumental in helping define the framework of problem areas on social media platforms.

Second, we analyzed the co-occurrence of described safety features and the associated 5Cs risks. For each of the five risk types, we aggregated all features coded as addressing that risk and treated this count as 100% to normalize the distribution. We then calculated the percentage of these features falling into each of the problem areas. These distributions were visualized as a heatmap through

Python 3.9 (see Figure 4), where columns represent the 5Cs risk types, rows represent problem areas, and cell values indicate the calculated percentages.

Table 2. Codebook for RQ1 used to classify platform safety features and the associated risks by problem area.

Problem Area (RQ1)	Described Safety Feature	Feature Definition
Content Exposure: Viewing, discovering, and being recommended content, including ads	Content Viewing Levels	Control the viewing of content based on topics or age-appropriateness, including ads content
	Content Removal	Remove harmful or otherwise policy-violating content
	Content Recommendation Restrictions	Limit visibility/recommendation of policy-violating content such as sensitive, mature, or violent content, but not to the extent of removing it
	Reporting	Report inappropriate content, accounts, messages, or interactions to the platform
	Audience Age Settings	Set minimum viewer ages or designate content for kids based on their age for content creators
	Content Approval	Share or approve selected videos or channels for a child, mostly by parents
	Search Controls	Control child's search access by disabling search bar and managing history, mostly by parents
Interpersonal Communication: The ways in which users engage in direct or indirect social interaction	Message Restrictions	Manage messaging permissions: restrict senders, disable DMs, control group chats
	Message Safety Filters	Detect, filter, and alert inappropriate content in messages
	Connection Controls	Manage follower requests and restrict suspicious adults from connections
	Content Engagement Restrictions	Regulate (e.g., warn, delete, block) comments, likes, tags, and mentions on user content
	Interaction Settings Nudge	Prompt users to review and potentially update user interaction-related settings, such as comments, mentions, follows, and friends
	Reporting	Report inappropriate content, accounts, messages, or interactions to the platform
Platform Usage Patterns: Concerning the frequency, duration, and nature of platform use	Time Limit Settings	Manage screen time by setting usage limits or time reminders
	Notification Settings	Pause/mute notifications via quiet modes or scheduled/automatic nighttime limits
	Autoplay Controls	Enable or disable the automatic playing of the next content
	Alternate Topic Nudge	Prompt users to switch content topics after prolonged consumption
Data Access: Managing personal data access by others or by the platforms	Data Privacy Controls	Set and control personal data privacy such as account visibility, content sharing, location, and discoverability from non-consensual entities
	Privacy Settings Nudge	Nudge users to review and update their general privacy settings
Monetization: Financial activities on platforms, such as earning or spending money	Monetization Restrictions	Limits monetization features like earning, payments, gifting based on user age
	Ads Personalization Settings	Manage ad preferences and personalized ads
	Purchase Restrictions	Limits in-app purchases and virtual gift buying based on age or content type
Content Creation: Creating, posting, and sharing public content (e.g., posts, videos, live streams)	Content Collaboration Restrictions	Restricts how others reuse or circulate children's content via Remix, Duet, Stitch, downloads
	Account Termination	Terminate accounts for violating content policies (e.g., circulating inappropriate content of youth)
	Content Creation Restrictions	Limits access to certain content creation features (e.g., Live stream) by age
Platform Access: Ability to access a platform's content and functions	Age Verification	Verifies user age via inputs, documentation, or detection tools
	Platform Access Controls	Allows caregivers to permit or block platform access
Parental Supervision: * Possibility for parents/guardians to oversee and guide a child's platform use, and we group features that mention "parent" or "parental" in a literal way here.	Parental Monitoring	Parents can monitor youth's online activity, including their connections, settings, communication metadata, and time spent
	Parental Control Settings	Manage overall parental supervision settings, including control approvals
	Individual Kid Profile	Allows parents to make distinct profiles with personalized settings for each child

*Parental Supervision is not a problem area, but it encompasses many safety features that help parents protect their children against online risks, so we include it here for completeness.

3.3.2 Analysis for RQ2: Assessing Problematic Communication Practices about Described Youth Safety Features. To examine potential misalignment in platform communication, two coders applied reflexive thematic analysis (RTA) [24] inductively to analyze the quotation column in our dataset.

RTA is a method for developing, analyzing, and interpreting patterns in qualitative data [23], incorporating researchers' pre-existing expertise and social positions. We selected RTA over other methods, such as content or discourse analysis, for two primary reasons. First, our research goal was to identify and synthesize recurring themes of problematic practices across the dataset, instead of linguistic patterns or values in platform communication that discourse analysis can help surface [131]. Second, while the communications are written for a general audience, uncovering problematic communication practices required a critical perspective that a general user's view alone might not provide. We therefore had to leverage our research team's expertise and backgrounds, making RTA the most appropriate choice as it formally incorporates this researcher positionality into the data interpretation, similar to how prior HCI work did so (e.g., [82, 141]).

The first coder reviewed all entries in the quotation column and assigned initial codes in Google Sheets. For example, the code "Vague Risk Language in describing content risks" was assigned to a TikTok quotation: "Filtering out content with complex or mature themes from teen accounts, powered by our Content Levels system." After this, the first coder consolidated similar codes into sub-themes and grouped related sub-themes into primary themes using a Miro board. In this process, the second coder regularly discussed all codes, associated quotations, sub-themes, and primary themes with the first coder to address disagreements and refine coding. This resulted in a thematic scheme (see codebook in Appendix A), answering RQ2 from Sections 4.2.1 to 4.2.3.

Researcher Positionality. Our team's diverse positionality supports the critical researcher angle adopted in our analysis, as we acknowledge that youth online safety is an often politicized and emotionally charged topic. This angle leverages our collective academic expertise and varied perspectives from computer science, HCI, and social sciences as an international team of academics. Also, our perspectives are independent of the social media technology industry, as none of us is funded by these companies. Furthermore, some members of our team are parents, adding a personal dimension to our understanding and engagement with youth online safety. This diverse academic and personal grounding, combined with our team's collective expertise in social media, platform algorithms, and youth online safety, was continuously reflected upon during regular discussions to refine our thematic schemes. These varied standpoints informed our assessment of platform communication practices and underpinned our RTA approach.

4 Findings

This section introduces the youth safety problem areas that we identified on social media platforms, and illustrates them with examples for safety features that platforms claim address risks that arise in these areas (RQ1). We then analyze how platforms *communicate* about these safety features and highlight the often vague and noncommittal nature of the communications (RQ2).

4.1 Problem Areas with Known Risks and Described Safety Features for Youth (RQ1)

In their communications, social media platforms present a large number of safety features that they claim address risks that arise on the platforms. We thus introduce a taxonomy to organize these claims and guide our discussion. Existing youth safety frameworks, such as the 4Cs [80], and our proposed extension, the 5Cs, are organized around online risks to youth. However, platform communications usually revolve around described safety features, not risks. Safety features often correspond to a "problem area" on social media platforms, which we define as the contexts where risks can emerge on the platforms, which, in turn, can be linked to more than one risk. Therefore, instead of organizing communications around risks, we organize them around *problem areas*.

4.1.1 Problem Areas on Social Media Platforms.

We identified seven problem areas: Content Exposure, Interpersonal Communication, Platform Usage Patterns, Data Access, Monetization, Content Creation, and Platform Access (see Methods Section 3.3.1 for the codebook). We also added Parental Supervision as an eighth, cross-cutting category that is not strictly a problem area. We begin by providing an overview of example safety features within each problem area that platforms claim to offer. For this, we counted the number of articles talking about a safety feature across time, source, and platforms within each problem area. These counts are descriptive aggregates of our dataset and give a trend indication of platform communication on the safety features in general. Note that the described safety features do not represent an assessment of whether these features are implemented, function as intended, or mitigate the risks claimed to address.

As shown in Figure 3, **our findings show an uneven distribution of platform communications across problem areas**, where 30.7% of communications focus on the Content Exposure area, while only 3.1% focus on Platform Access. This highlights a strong emphasis of communication on moderating what youth *see* rather than whether they can *enter* platforms. Similarly, Interpersonal Communication (27.3%) received substantial attention, while areas like Monetization (8.4%) and Content Creation (6.0%) were comparatively underrepresented. These patterns suggest that platform communication was concentrated around the interpersonal contexts, potentially overlooking the commercial or social contexts, where youth create and monetize content or perform commercial activities (e.g., purchasing products). Below, we detail each problem area.

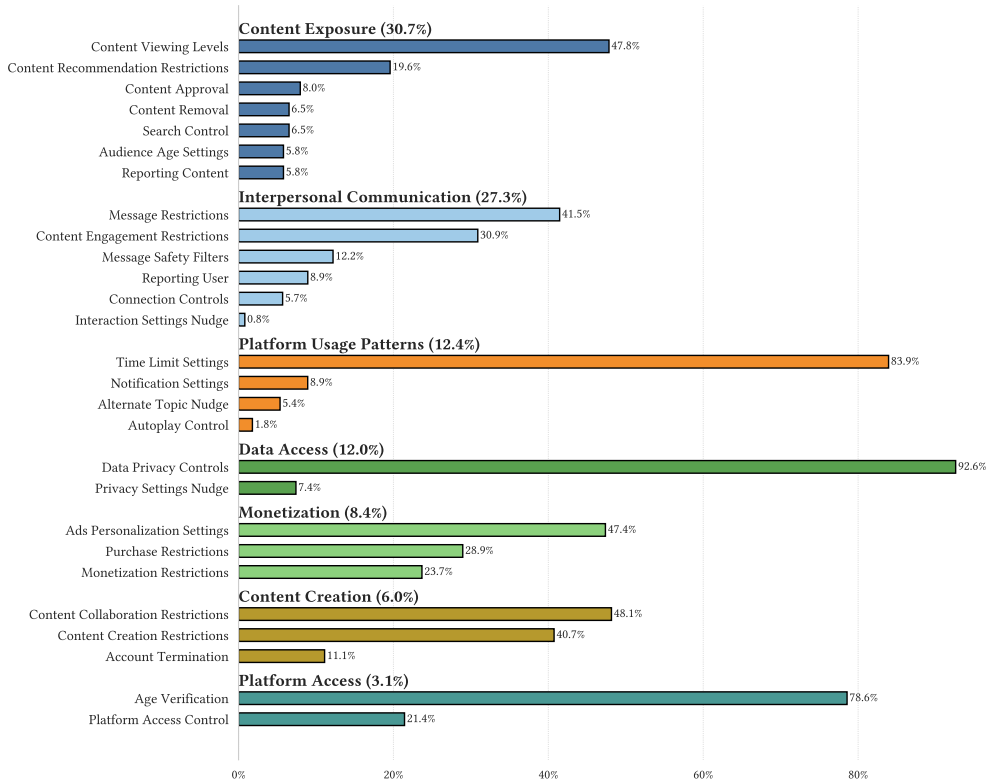


Fig. 3. Problem areas comprising the described safety features with the frequency of their mentions.

(1) Content Exposure: This is the most common problem area found in platform communications, which corresponds to the risk of youth encountering harmful, manipulative, or age-inappropriate content. The most prominent features mentioned within this area were *Content Viewing Levels* (47.8%) and *Content Recommendation Restrictions* (19.6%). These features reflect a strategy of top-down control, from either parents or platforms, in filtering what youth are allowed to see. For example, Instagram stated: “Teen Accounts are automatically set to see less sensitive content in search results and recommended content in Explore, Reels and feed” [62]. This illustrates how platforms frame safety as something that can be managed through pre-set restrictions.

(2) Interpersonal Communication: In this area, risks can emerge when communication channels such as messaging and commenting enable harassment, manipulation, or unwanted contact from peers or from adults. *Message Restrictions* (41.5%) were the most mentioned feature in the platform communications, followed by *Content Engagement Restrictions* (30.9%). The focus of these tools was primarily on *restricting access* by controlling who can message whom or who can comment or tag. Notably, while platforms mentioned the introduction of nudging features, such as Instagram’s comment warnings that prompt users to reconsider posting offensive language [54], these tools remained reactive in nature.

(3) Platform Usage Patterns: Platforms encourage continuous user engagement, which can lead to excessive or compulsive use. The most communicated feature in this area is *Time Limit Settings* (83.9 %). For instance, TikTok and Instagram mentioned automatic daily screen time limits of 60 minutes for teens [62, 67], prompting them to close the app or enter a passcode to request additional time once the limit is reached. Instagram also promoted features, such as “Quiet Mode” [58], which pauses notifications and summarizes what the user missed upon return. This feature aims to prevent excessive use through timers, reminders, or screen-time limits, as a nudging design.

(4) Data Access: This area describes how people manage access to their personal data by others or the platforms. Nearly all communication in this area centered on the feature *Data Privacy Controls* (92.6%), which allows users to manage profile visibility, discoverability, or data sharing. Data access becomes problematic when the personal data of youth is accessible to others and can be used for stalking, identity theft, targeted harassment, or grooming. For example, youth accounts were typically claimed to be set to private by default by social media platforms, which limited who could see their content or discover them.

(5) Monetization: Risks emerge in this area when youth are encouraged to spend money without understanding its value, or when youth creators generate income through their content, as this can lead to labor exploitation or performance pressure. Additionally, platforms often personalize advertising based on user behaviors, which can expose youth to age-inappropriate or persuasive ads, and youth may not be equipped to fully understand or critically evaluate them. The most frequently mentioned safety feature was *Ads Personalization Settings* (47.4%), followed by *Purchase Restrictions* (28.9%), and *Monetization Restrictions* (23.7%). These features are largely regulatory in nature (i.e., mandated by law) and aim to protect children from financial risks by reducing exposure to targeted advertising or limiting the possibility of spending money on in-app purchases.

(6) Content Creation: Social media platforms can pose risks when youth share content without fully understanding its visibility or permanence. Such behavior may result in unwanted attention, grooming, bullying, or reputational harm. Additionally, the desire for visibility and engagement can foster performance pressure and attention-seeking behaviors, such as posting risky or revealing content. *Content Collaboration Restrictions* accounted for nearly half (48.2%) of the communicated safety features in this area, followed by *Content Creation Restrictions* (40.7%). These features were designed to prevent others from remixing, downloading, or circulating youth content. The underlying logic is focused on mitigating content misuse after it is published. For example, TikTok

described a policy that disables the ability to duet or stitch videos by users under 16, and sets the default for users aged 16–17 to “Friends only” with downloading disabled [117].

(7) Platform Access: Gaining access to the platforms is the first step to experiencing social media. Risks in this area relate to the possibility that youth under the intended minimum age may access platforms without safeguards or in ways that bypass these safeguards. Platform Access had the lowest overall communication volume, but within it, *Age Verification* was communicated the most (78.6%). This feature aims to prevent underage users from entering the platform, mostly through self-reported age checks. For example, Instagram reported requiring users to input their birth date and may prompt them to verify their age if AI-based systems detect inconsistencies [55].

For completeness, we also include safety features that enable parents to mitigate risks in different problem areas. We categorized them under a cross-cutting layer we call **Parental Supervision**. The most communicated feature was *Parental Monitoring* (61.8%), followed by *Parental Control Settings* (27.9%), and *Individual Kid Profile* (10.29%). These features enable parents to observe or adjust their child’s platform use and privacy settings. The trend here is external oversight, where parents, rather than youth, are positioned as central agents of safety.

Platform safety communication may reflect underlying safety feature design priorities, for instance, a greater focus on the Content Exposure area (30.7% of mentioned safety features) and on the Interpersonal Communication area (23.8% of mentioned safety features). In the next section, we examine how known risks emerge across these problem areas and how social media platforms communicate about using safety features to address those risks.

4.1.2 Mapping Social Media Problem Areas to Known Risks Faced by Youth.

Using our 5Cs framework, this section explores how five risks, namely content, contact, conduct, contract, and circulation, emerge across the problem areas given by how platforms claim to address them (see Figure 4). Please note that the heatmap does not represent the prevalence or severity of risks, but how frequently platforms allocate their communications across the problem areas, i.e., the co-occurrence of known risks and described safety features.

Content risk most arose in the Content Exposure area, as most safety features were communicated there. This suggests that platforms primarily frame content risk as a matter of what youth are exposed to, particularly through feeds, recommendations, or searches. Other areas where content may play a harmful role, such as Content Creation or Interpersonal Communication, were rarely linked to content risk in platform communications.

Contact risk appeared most often in the Interpersonal Communication and Data Access areas. This indicated that platforms primarily addressed this risk when it came from direct messaging, commenting, or being discoverable by others. This framing thus emphasized one-on-one, private interactions and underrepresents the peer-based or public dimensions of contact risk that often occur through Content Exposure or Content Creation.

Conduct risk largely emerged in the Platform Usage Patterns area and claimed to be addressed through Parental Supervision. Here, platform communications centered on tools like Time Limit Settings and Parental Monitoring, which suggested a behavioral framing that connects youth safety to screen time and adult oversight.

Contract risk was frequently associated with Monetization, Data Access, and Platform Access areas. Platforms often cite features like Age Verification, Purchase Restrictions, and Ads Personalization Settings. These protections often relied on self-reported information, which implied that contract risks are mitigated through gatekeeping and usage policies.

Circulation risk arose across multiple problem areas, most notably Content Creation and Content Exposure. Platform communications reference various features to address this risk. For example, in Content Exposure, YouTube stated that they aimed to reduce the recommendation

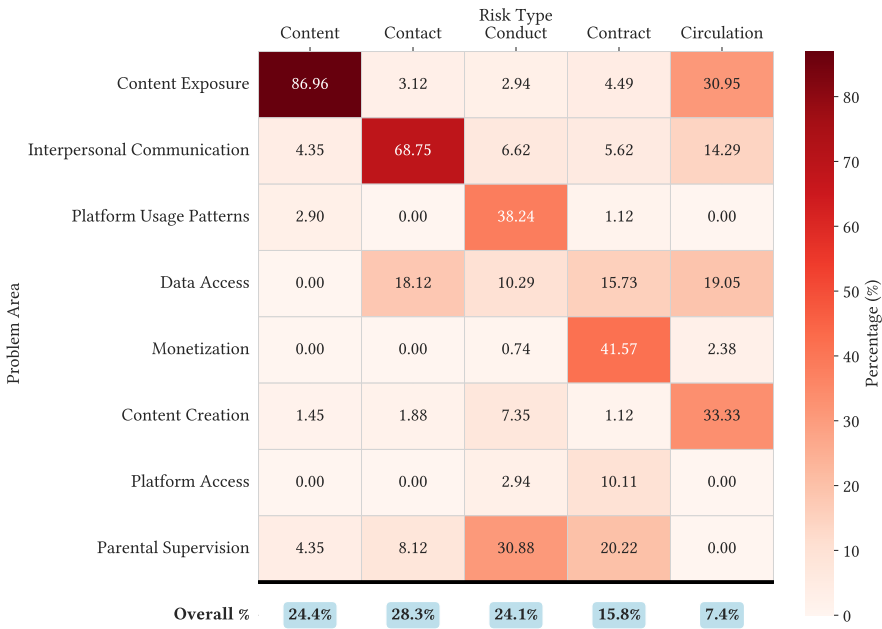


Fig. 4. Distribution of the communicated safety features for each risk type across the problem areas. Problem areas are sorted from most communicated to least communicated. The percentage shows the relative frequency of the safety features across the problem areas that aim to address a risk. Each column sums up to 100%. Overall percentages show how frequently risk types are mentioned by platforms.

of risky content from Youth: “We expanded our efforts from earlier this year around limiting recommendations of borderline content to include videos featuring minors in risky situations. While the content itself does not violate our policies, we recognize the minors could be at risk of online or offline exploitation” [60]. In Content Creation, Instagram, for example, communicated a feature that did not allow non-followers to remix the content of youth: “Teen Accounts are automatically set to only allow tags, mentions, and content remixing by people they follow” [135].

Overall, we found that platforms mostly communicated about features that aim to address content risks (24.4%), contact risks (28.3%), and conduct risks (24.1%). Specifically, features aiming to address circulation risks (7.4%) were largely undercommunicated. These patterns offer context for RQ2, where we examine how platforms communicate about the implementation or effectiveness of their youth safety features.

4.2 Problematic Communication Practices regarding Youth Safety Features (RQ2)

From an end-user standpoint, effective corporate communication should be clear, consistent, and easily actionable for a user to adopt their products [19, 51]. Thus, social media platforms ideally need to minimize users’ cognitive load by clearly explaining risks and communicating about how features function to support their safety. However, our analysis of platform communications related to youth safety found that the communications were vague and non-committal. We observed that such problematic communication practices followed a three-phase lifecycle, from (1) initial announcement to (2) describing the operational details of these features, and finally to (3) claims about their impacts. In the following, we describe these problematic practices across these three phases, where Figure 5 provides an overview.

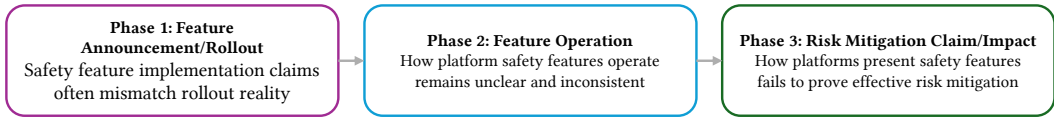


Fig. 5. Problematic practices in social media platforms' communication lifecycle about youth safety features.

4.2.1 Phase 1: Safety Feature Implementation Claims Often Mismatch Rollout Reality.

This section addresses how platform communications about the implementation of safety features often mismatch the availability of those features. This misalignment arises primarily through two problematic communication practices.

(1) Provisional and evolving language makes it unclear *whether* features are implemented. Platforms often announce features that they plan to implement or are testing, which can create uncertainty about whether or when these features are implemented. For example, in 2024, Meta repeatedly employed terms like “testing,” “rolling out,” and “planning” to market its Nudity Protection feature in the Interpersonal Communication area (e.g., [86, 87]). Such ambiguity can intensify when platforms announce feature testing without a concrete feature name, as when Meta described testing “new pop-up messages for people who may have interacted with an account we’ve removed for sextortion...” [88].

(2) Stated geographic limits make it unclear *where* features are implemented. Platforms often mention that they implement safety features only in certain regions, without stating a reason. For example, Snapchat described restricting personalized advertising in the Data Access area for teens, specifically “in the EU and UK” [112], while TikTok described similar restrictions covering the “European Economic Area, United Kingdom, and Switzerland” [119]. Neither mentioned such a restriction (or the lack of such a restriction) in the U.S. This geographically fragmented approach means that the described safety features might not be universally available, even after platforms promise a “global rollout.” For example, Meta’s narratives about the Nudity Protection feature in the Interpersonal Communication area, starting from 2024, described it as “rolling out globally” [88], while later noting it “may not be available in your location right now” [64]. Likewise, Meta stated in 2024 that its “Teen Accounts” were only active “in some locations” with a full rollout planned in early 2025 [63]. As a result, youth, parents (and researchers) must resort to testing for themselves whether a feature is available to them. This indicates an undue burden, and can be challenging for safety measures that operate in the background and have no permanently visible user interface components, such as the Nudity Protection feature.

4.2.2 Phase 2: How Platforms’ Safety Features Operate Remains Unclear and Inconsistent.

This section discusses how problematic communication practices originate from a lack of clarity and consistency in describing how safety features operate. This manifests in three primary ways.

(1) The conditions triggering safety features to activate remain unclear, especially when conditions change without explicit notice and justification. The age threshold triggering safety features to work often lacks clarity, which can affect both children and parents. For example, Snapchat requires parents to be over 25 to utilize its Family Center [111] without explanation of why there is an age threshold for parents. We observed that platforms usually do not mention a concrete rationale for age thresholds. Furthermore, on several occasions, platforms have made changes to age thresholds without highlighting that the new thresholds represent a change or what motivated them to make the change. This can lead to confusion and inconsistency, particularly as the thresholds become stricter over time. For example, TikTok’s communications regarding default Private Account settings, in the Data Access area, evolved from applying to ages 13-15 [36] in 2021

to later stating, “*all accounts for teens under 18 start out as private*” [124] in 2025. TikTok did not mention a rationale for the change, or even that there was a change. Following a similar pattern, Meta shifted from defaulting those “16 and under” in 2023 [85] to “*under 18*” in 2024 [62] without full disclosure that this represented a change, and why the change was made. Furthermore, when age thresholds are combined with other factors, like whether users follow each other, it can become difficult for youth and parents to understand what specific conditions activate safety features.

(2) How safety features intersect with each other remains unclear. Platforms mention various safety features are available, but they often fail to clarify how the features intersect with each other or function together. For example, understanding the effectiveness of content filters is complicated when platforms mention multiple features without clear differentiation. TikTok mentioned “*Restricted Mode*” [116], a “*Content Levels system*” filtering “*mature or complex themes*” [118], parental keyword filtering layered “*on top of our Content Levels system*” [120], and separate “*maturity score[s]*” [66]. These all can create confusion about how safety features function together to mitigate content risks, which might eventually hinder users from adopting suitable features.

(3) Rationales for varying restriction levels regarding similar features remain unclear. Platforms sometimes apply different levels of restrictiveness to features that may appear similar to users, but do not communicate why. For example, on TikTok, Duets and live streams are both forms of content creation, yet TikTok described different restriction levels: “*Only people over 18*” can host a livestream [121], but for Duets involving pre-recorded content, it highlighted that “*your teen has the control to decide who can make duets with or react to their content*” to imply it was teens between 13 and 18 [116]. It is likely that the platform assumes a different risk severity between live and pre-recorded content, but the communication failed to articulate this rationale well.

4.2.3 Phase 3: How Platforms Present Safety Features Fails to Exhibit Effective Risk Mitigation.

We found that platforms’ presentations of safety features fail to demonstrate the effectiveness of claimed risk mitigation through four communication practices.

(1) How platforms communicate about risks leaves the targeted risks unclear. We found that platforms often fail to clearly explain the risks that safety features aim to address. For example, platforms often do not specify what they mean by inappropriate content. Meta described its Sensitive Content Control feature in the Content Exposure area as addressing potential “*harmful or sensitive content*” [56] that “*may not be age-appropriate*” [85], without offering a concrete characterization of the content targeted by these filters. Similarly, for contact risks, explanations can be circular rather than criteria-based: Meta defined “*potentially suspicious behavior*” by adults not by describing the behavior itself, but by citing examples such as “*adults who have recently been blocked or reported by a young person*” [62]. Such vague descriptions also appear for contract risks, as when YouTube mentioned it “*won’t serve personalized ads or ads in certain categories*” [20] in the Monetization area, without specifying which ad categories are not suitable for children.

(2) The effectiveness of safety features in addressing risks lacks evidence. Platform communications about safety features’ effectiveness often lack verifiable evidence. On the one hand, platforms explicitly admit that their safety features have limitations, such as content filters that can produce false negatives, stating that “*no system is perfect and inappropriate videos can slip through*” [138]. On the other hand, when platforms do claim improvements in their safety features, the communication may lack verifiable evidence: YouTube asserted a new content classifier “*will detect and remove 2X more individual [predatory] comments*” [136] without the necessary baseline or context to validate this numerical claim. This lack of evidence-based reporting made it impossible to hold platforms accountable for their claims or to determine whether removing “*2X more*” represents meaningful progress or an insignificant change.

(3) How platforms claim safety features work mismatches the scope of the risks they claim to address. Platforms sometimes describe safety features in ways that mismatch the scope of the risks they purportedly address. For example, YouTube initially framed the “*Made for Kids*” feature through the lens of policy compliance with COPPA (Children’s Online Privacy Protection Act) [137] to address contract risks (i.e., protecting children from data collection without parental consent). However, its description stating creators are “*required to tell us that your videos are made for kids if you make kids’ content*” [137] implied a broader application that might also address content and circulation risks, which the platform did not explicitly detail, and creates ambiguity for creators. Similarly, TikTok mentioned restricting the exchange of images or videos in Direct Messages [49] but omitted mentioning other forms of content, such as “*effects, hashtags, or sounds [that can be sent] through a direct message*” [123]. This leaves it unclear whether these other forms of content are also restricted, or there is a possibility of circulation and content risks if they are exempt from the restriction.

In addition, platforms often avoid talking about risks by highlighting the opportunities their safety features offer. For example, Meta positively framed the safety feature Sensitive Content Control as a way for users to “*personalize their experience*” [57] and the Alternate Topic Nudge as designed to “*encourage teens to discover something new*” [59]. Thus, this framing around personalization and discovery can misrepresent the full spectrum of risks platforms attempt to manage.

5 Discussion

Our study introduces the framework of *problem areas* and investigates how platforms communicate about youth online safety. We uncover an unevenness of such communication: the mention of some risks, such as content and contact risks, is highly concentrated in a few problem areas, like Content Exposure and Interpersonal Communication, while other risks, especially conduct, contract, and circulation, are mentioned more sporadically across the problem areas (RQ1). In addition to such unevenness, we identified recurring problematic practices in platforms’ communication lifecycle, spanning from feature/product rollout to operation and effectiveness (RQ2). In this sense, our study highlights the need to analyze platform communication practices, not just the safety features, as an integral part of youth online safety efforts by social media platforms.

5.1 Mapping Online Risks, Problem Areas, and Safety Features to Uncover Communication Gaps

While the prior 4Cs risk framework [80] provides a strong foundation for categorizing online risks faced by children (e.g., [2, 13, 126]), understanding youth safety on social media platforms requires analyzing *where* these risks arise and *how* risks are claimed to be addressed by the platforms. Our problem areas framework contributes this missing layer by offering a structure for cross-risk and cross-platform analysis. This extends previous work that primarily focuses on the feature- or risk-specific levels (e.g., [2, 34, 82, 98]). Applying our framework to platform communications reveals a gap in how platforms communicate about the problem areas and describe safety measures.

Regarding *content* risk, we found platform communication tended to focus on addressing the problem area Content Exposure over other areas. By narrowly focusing on passive content consumption, platforms overlook the risks of how youth create and share content. Also, platforms emphasize top-down content filtering by either parents or platforms. However, this contrasts with recent HCI and CSCW research showcasing youth’s desire to shape their own content consumption experience [15, 25].

In the context of *contact* risk, we found that most communication prioritized restricting interpersonal communications with adults. While this aligns with traditional online safety narratives that focus on protecting children from online “predators” or strangers [76, 108], this communicated

focus underrepresents the contact risks that youth face in more public or peer contexts, such as Content Creation and Exposure areas. For example, peer-based harms like cyberbullying [26, 105] can occur in public view within Content Exposure (e.g., via comments or public content creation), which extends beyond interpersonal interactions.

For *conduct* risks, platform communication was concentrated on features like Time Limits and Parental Monitoring, which align with general recommendations on screen time for youth [73] and prior work suggesting that awareness of children's screen time can enhance parental mediation efforts [100]. However, risks that emerge through peer interactions such as peer aggression and cyberbullying [26, 68, 74, 108] are mentioned far less often. Through this imbalance, platforms signal a focus on individual self-regulation and parent-led oversight, mechanisms that are easy to measure and to promote. This neglects the social dynamics where mechanisms like peer support systems [50, 98] might help, but might be more complicated to communicate.

While platforms communicated about *contract* risks across multiple problem areas, platform communication fails to address the challenges uncovered by recent HCI and CSCW work, which found that youth often lack awareness of datafication risks [46, 114, 140] and that parents may be passive or underinformed about such risks [70]. As a result, platform communication related to the commercial and data-driven dimensions of contract risks can be described as superficial.

Lastly, we turn to *circulation* risk, which we argue in this paper is not well represented in the existing 4Cs framework. While youth increasingly act as content creators or influencers on social media platforms [84, 104, 143], we found that platforms claim few features to address this risk, and these claims are scattered across different problem areas. As a result, users may not understand how these elements work together to mitigate the spread of sensitive or non-consensual content.

Illustrated across the five risks, we highlight the need to move beyond categorizing risks in isolation and toward analyzing the problem areas in which they arise. Our problem area framework supports contextual evaluation of platform safety communication by exposing how risk types intersect within specific areas of user activity on the platforms. By surfacing these cross-risk implications, our framework extends prior work that tends to focus on designing safety features against singular risks like contact [2, 3] or content [33].

Beyond mapping these complex risks, the primary value of our problem areas framework is structural by providing a concrete taxonomy beneficial to future work in the auditing, comparison, and structured evaluation of safety features. For instance, future work can use our framework to conduct a direct audit of safety features within the Monetization problem area, comparing the claimed 'Ads Personalization Settings' and 'Purchase Restrictions' in Table 2 across YouTube, TikTok, and Instagram. This provides a structure to identify gaps, as our findings for RQ1 revealed (e.g., the general neglect of the Monetization area). Furthermore, our problem area taxonomy can guide the design of experiments to test the effectiveness of safety features, such as comparing how well Instagram's 'Message Restrictions' in the Interpersonal Communication area performs against those of Snapchat. Ultimately, such context awareness, at the core of our framework, can facilitate recent efforts in HCI and CSCW toward evidence-based safety solutions for youth [2, 13, 25, 91], enabling contextual evaluation rather than reliance on platforms' one-sided claims.

5.1.1 Communication Implications for Youth Safety Design.

While we did not directly study platforms' youth safety design philosophies, the way they frame their communications hints at approaches that both echo and contrast with regulatory frameworks such as age-appropriate design codes [53] and evolving research like *Safety by Design* principles [15, 45]. The latter advocates for proactive and enabling solutions that integrate safety and risk prevention directly into the design and development of a product or services for youth. We briefly

examine platforms' framing of youth safety across three dimensions: when interventions occur, who controls safety features, and how youth develop resilience:

(1) Proactive vs. reactive: When safety interventions take effect as communicated by the platforms. Prior work has uncovered that youth online safety solutions are often reactive, intervening mostly after harm occurs [99, 126, 130]. Our findings on platform communication present a nuanced picture. Platform communications were overwhelmingly dominated by proactive features, such as 'Data Privacy Controls,' 'Time Limit Settings,' and 'Content Viewing Levels;' purely reactive features like 'Reporting' were mentioned less (Figure 3). At the surface, an emphasis on proactive features appears to align with the principles of *Safety by Design* [45]. However, prior research recommends that safety features should not only be proactive, but also be proactive in an enabling manner, as discussed below.

(2) Restrictive vs. enabling control: Who has agency over safety features as depicted by the platforms. Our findings show that platforms have communicated more about safety features that centralize restrictive control of parents and platforms rather than giving agency to youth (e.g., Platform Access, Monetization, Platform Usage Patterns, Parental Supervision in Table 2 and Figure 3). This contrasts with recent research, which shows that children value personal privacy over parental controls [15] and advocates for enabling mediation that guides technology use and keeps multiple stakeholders engaged [27, 79]. **There is an opportunity for platform communication to rebalance the safety narrative from one of control to one of guided autonomy.** For example, in the context of account visibility in the Data Access area, platforms can provide age-appropriate explanations of the enabling options for youth, including benefits and risks of privacy settings. In fact, however, we found that platforms obscured these youth-controlled settings through unexplained (and imposed, rather than enabling) shifts in the default settings per age group (Section 4.2.2).

(3) Passive vs. active learning: How youth develop resilience as communicated by the platforms. Current platform communication often positions youth as passive (e.g., subject to filtering or monitoring). Platforms used vague, undefined terms like "sensitive content," which might prevent active learning (Section 4.2.3). Furthermore, they positively framed these top-down controls as "personalization" rather than explaining the risks they mitigate (Section 4.2.3). This passive framing in communication contrasts with the recognized shift in research towards strength-based approaches in online safety [45], moving beyond protecting victims [14] to empower youth with experience, knowledge, and skills in addressing risks [142]. **The implication is that platform communication itself should be an educational tool for building resilience.** For this, platforms should provide clear, age-appropriate explanations for why a feature exists and how it works. For example, an explanation for a Message Restriction feature should not only state that it blocks strangers, but also explain the types of contact risks (e.g., phishing, grooming) it helps to mitigate, thereby teaching youth risk assessment.

5.2 Platform Communication as Youth Online Safety Effort: Navigating Tensions in Enhancing Transparency

Recent HCI and CSCW research has begun to evaluate the safety efforts of social media platforms, such as assessing the utility of current built-in features [33], and co-designing new safety features with youth (e.g., [2, 14, 128]). Complementing this line of work, our study is among the first to evaluate how platforms communicate about safety features. How platforms announce, describe, and market their safety features is an integral component of their overall safety efforts for youth. Our findings reveal problematic practices across the communication lifecycle: Claims about feature implementation mismatch with rollout reality; explanations of feature operation are unclear; and descriptions fail to demonstrate the effectiveness of features in mitigating risks. These practices

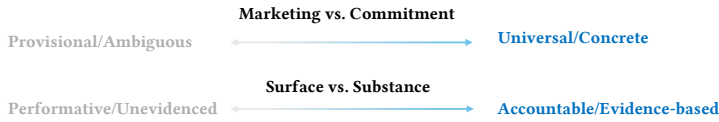


Fig. 6. Two key tensions identified in platforms' communications for youth online safety.

point to underlying tensions platforms have to navigate between demonstrating their responsibility for youth safety and managing increasing public concerns like media scrutiny [103, 125] and legal liabilities [40–42]. Examining these tensions is essential, as corporate communication can significantly affect user perceptions and adoption of products [48, 72]. Our study uncovers two primary tensions in platforms' communication: (1) marketing goals versus commitment, and (2) surface versus substance, as shown in Figure 6.

First, platforms face a tension between their marketing goals and their commitments to youth safety. Our findings show that platforms periodically announced new safety features, but their announcements over time concealed the features' actual availability, and evolving criteria made it unclear what triggers those features to activate and work (Section 4.2.1). While such communication practices aligned with corporate marketing goals like trust-building and adoption [72], they took place amid increasing public and regulatory scrutiny over harmful content circulation [1, 129] and its severe consequences on youth (e.g., suicidal incidents [103, 125]), which created a need for clear, consistent information for users [19, 51]. Prior work has shown limited platform policies concerning harmful content on social media [22, 110]. Our study adds that platform communication rarely clarified *when*, *where*, or *how* the feature enforces policies.

Second, platforms often described safety features positively by emphasizing personalization or discovery (Section 4.2.3), but they admitted limitations only vaguely or omitted evidence validating the features' effectiveness (e.g., YouTube comment classifier; Section 4.2.3). This communication approach, likely aiming to build user trust [72], mirrors risk-downplaying strategies observed in the tobacco and e-cigarette industries [92]. This reflects a tension between needing to claim effectiveness for public acceptance versus avoiding evidence-based evaluation [101] or creating legal liability (e.g., [40, 41]). This lack of substantiation in platform communications echoes how prior HCI and CSCW found that safety features were perceived as ineffective for addressing youth's real concerns [2, 7]. It suggests, therefore, that a long road remains to be taken toward safety features that are demonstrably effective.

These two tensions ultimately center on the transparency of platform communication artifacts/channels, such as the press releases and safety blogs we analyzed. Transparency would fundamentally require social media platforms, as *sociotechnical* systems, to be accountable [44] and to offer visibility into their complexity [8, 43]. Prior HCI and CSCW research shows that explanations from online platforms can enhance transparency [81] and potentially increase user trust [69]. In the context of youth safety, full operational transparency may not always be the solution. Challenges like low adoption of current safety features by parents [4] or limited awareness of those safety features among youth [91] may not be resolved through more openness alone. As prior work suggests, detailed disclosure of platform safeguard mechanisms could aid evasion by users [90] or allow bad actors to exploit system knowledge for grooming or distribution of child sexual abuse material (CSAM) [65, 94]. This highlights an important tension in platform communication practices: balancing the benefits of communication transparency for empowering users against the ramifications of transparency (e.g., feature misuse, circumvention of safeguards).

How platforms demonstrate to the public how they manage this tension, therefore, serves as the key indicator separating genuine safety efforts from reputation management.

5.2.1 *Guidelines for Communicating about Safety Feature Implementation and Effectiveness.*

Given the problematic practices in platform communications in our findings, we also make recommendations for improving the platforms' communication practices around safety features.

(1) We advocate for a mandate for transparency on feature availability. To counter the ambiguous or geographically inconsistent rollout claims as reported by our findings (Section 4.2.1), regulators such as the U.S. FTC could require platforms to clearly and accurately report *where*, *when*, and *to whom* specific safety features are operationally available. From a design perspective, this transparency must extend beyond press releases into more in-app mechanisms. For example, social media platforms could implement a personalized "safety dashboard" within the app that confirms for a user (and their parent, if linked) which safety features are actually active on their specific account, in their region, at that moment.

(2) Clear explanations of feature operation and responsibility are required. Addressing the opacity that we found in feature descriptions (Section 4.2.2), including unclear triggers, regulations should mandate consistent explanations of how safety features actually function. Clear operational details are necessary for users to utilize safety features and understand the lines of responsibility of operating those features between themselves and the platforms.

(3) Use the problem areas framework to contextually evaluate safety communication beyond features. To address the performative safety claims lacking verifiable evidence (Section 4.2.3), policy can require evidence-based substantiation through standardized metrics or independent auditing protocols. This could involve "red team" scenario testing, where independent auditors use test accounts to systematically compare platform promises against real-world feature effectiveness. Recent reports on platforms like Instagram, for example, have shown that many safety features are ineffective or easily circumvented when tested this way [18]. Moreover, recognizing that communication is an integral part of a social media platform's safety efforts, regulators should evaluate beyond the features to the communication itself. Instead of assessing communications in the abstract, regulators (e.g., FTC, state Attorneys General) can use our problem areas framework to systematically probe how clearly and accurately platforms explain risks and features within each specific context. For example, they could demand consistent explanations for features within often-overlooked areas like Monetization or Platform Access, ensuring the narratives comply with legal expectations. This would allow regulators to better demand alignment between what platforms claim about their safety features and how those features actually perform.

(4) Design communications to bridge stakeholder collaboration gaps. The communication gaps we found have practical, negative consequences for collaboration. For instance, the vague effectiveness claims (Section 4.2.3) might give parents a false sense of security, leading to conflict when a teen knows a feature is ineffective. Platforms should thus design and co-design honest communication materials for each stakeholder. This could include "one-page guides for educators" that frame the risks in under-communicated areas like Monetization or "conversation starters for families" that are based on the actual (not marketed) functionality of features. By providing a shared understanding of the problem areas and safety features, platforms can facilitate collaboration between youth, parents, educators, and policymakers.

6 Limitation & Future Work

Our study has limitations that can also inform future work. First, we analyzed how platforms communicated about their safety efforts, but we did not independently examine whether the safety features actually work the way they are described. Future work can investigate how well

platforms' actions align with their statements through empirical research. Second, we counted how often certain topics (e.g., features, risks) were mentioned in articles to gauge communication practices. While this reveals what platforms emphasized, it does not indicate whether users saw these messages or if they had any impact. Plus, these counts can change depending on how platforms publish information, such as using lots of small posts versus one large one. Examining if people actually saw and understood such youth safety-related communication would be a valuable next step. Third, for step 4 in Section 3.1, we acknowledge that such a snowballing sampling approach may over-represent some safety features that platforms themselves emphasized through their internal sources. However, our research aim is to comprehensively capture *platform-communicated* claims on safety features. Hence, we consider this data enrichment step as construct-consistent rather than a threat to our data validity, similar to a backward snowball sampling in systematic literature review [132]. Moreover, the snowballed articles constitute only about one-fourth of the final dataset, which makes them supplementary but not dominant in our dataset. Fourth, we only looked at press releases, help center entries, and safety blogs on the platforms' official websites. We did not examine other important channels through which companies communicate with their users and regulators, such as in-app user interface tutorials or help texts, safety-related advertising, videos, podcasts, and regulatory reports, including risk assessments for the Digital Services Act, which can be a valuable next step. Fifth, we focus only on problematic communication practices of social media platforms, which can introduce selection bias in data collection. However, this gave us the opportunity to identify gaps and areas for improvement. Lastly, our study focuses on the U.S. context as the social media platforms in our study are all U.S. companies, so future work can understand how platforms communicate their safety features across regions.

7 Conclusion

Social media platforms face increasing scrutiny over youth online safety, but currently, understanding how they communicate about their safety features in addressing online risks for youth remains limited. Our study is among the first to comprehensively examine how popular social media platforms, including YouTube, TikTok, Meta (Instagram and Facebook), and Snapchat, frame and present their built-in safety features for youth protection, leading to a taxonomy of safety features along problem areas that describe where risks arise and where platforms apply safety features to address the risks. By analyzing $N=352$ press releases and safety-related blogs from these platforms, we mapped the described safety features to seven problem areas. Our analysis revealed an uneven distribution of platform communication across these areas. For instance, communications are heavily focused on Content Exposure and Interpersonal Communication, while largely neglecting Content Creation, Data Access, and Platform Access areas. Furthermore, we identified three problematic practices in how platforms communicate about safety features, including discrepancies in feature rollout and operation, as well as differences between claimed and potential effectiveness of these features in mitigating known risks. We argue for holding social media platforms accountable for responsibly communicating how they protect youth from different online risks.

Acknowledgments

We thank the senior program committee (PC) member, the PC member, and all reviewers for their constructive feedback. This work was supported in part by the German Research Foundation (Grant: 543018872), joint internal funding from NYU and Ludwig Maximilian University of Munich, and the Democracy Fund. Additionally, Dr. Wisniewski's research on youth online safety is supported by the U.S. National Science Foundation under grants #IIS-2550812, #TI-2550746, and #CNS-2550834, and by the William T. Grant Foundation grant #187941. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect

the views of the research sponsors. We disclose that Dr. Damon McCoy has served as an expert witness for the plaintiffs in lawsuits against some of the social media platforms included in this study.

References

- [1] Jaclyn Abraham, Rebecca Roth, Heidi Zinzow, Kapil Chalil, and Pamela Wisniewski. 2022. Applying Behavioral Contagion Theory to Examining Young Adults' Participation in Viral Social Media Challenges. *Transactions on Social Computing* 5, 1-4 (11 2022), 1–34. doi:10.1145/3538383
- [2] Zainab Agha, Karla Badillo-Urquiola, and Pamela J. Wisniewski. 2023. "Strike at the Root": Co-designing Real-Time Social Media Interventions for Adolescent Online Risk Prevention. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (4 2023), 149. doi:10.1145/3579625
- [3] Zainab Agha, Kelsey Miu, Sophia Piper, Jinkyung Park, and Pamela J. Wisniewski. 2023. Co-Designing User Personas and Risk Scenarios for Evaluating Adolescent Online Safety Interventions. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, CSCW (10 2023), 249–253. doi:10.1145/3584931.3606964
- [4] Mamtaj Akter, Amy J. Godfrey, Jess Kropczynski, Heather R. Lipford, and Pamela J. Wisniewski. 2022. From Parental Control to Joint Family Oversight: Can Parents and Teens Manage Mobile Online Safety and Privacy as Equals? *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (4 2022). doi:10.1145/3512904
- [5] J. Alemany, E. del Val, J. Alberola, and A. García-Fornes. 2019. Enhancing the privacy risk awareness of teenagers in online social networks through soft-paternalism mechanisms. *International Journal of Human-Computer Studies* 129 (9 2019), 27–40. doi:10.1016/j.IJHCS.2019.03.008
- [6] Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Chen Ling, Munmun De Choudhury, Pamela J. Wisniewski, and Gianluca Stringhini. 2023. Getting Meta: A Multimodal Approach for Detecting Unsafe Conversations within Instagram Direct Messages of Youth. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (4 2023), 30. doi:10.1145/3579608
- [7] Abdulmalik Alluhidan, Mamtaj Akter, Ashwaq Alsoubai, Jinkyung Katie Park, and Pamela Wisniewski. 2024. Teen Talk: The Good, the Bad, and the Neutral of Adolescent Social Media Use. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (11 2024). doi:10.1145/3686961
- [8] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20, 3 (3 2018), 973–989. doi:10.1177/1461444816676645
- [9] Monica Anderson, Michelle Facerio, and Jeffrey Gottfried. 2023. Teens, Social Media and Technology 2023. <https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/>
- [10] Zahra Ashktorab and Jessica Vitak. 2016. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. *Conference on Human Factors in Computing Systems - Proceedings* (5 2016), 3895–3905. doi:10.1145/2858036.2858548
- [11] David L. Ashley and Cathy L. Backinger. 2012. The food and drug administration's regulation of tobacco: The center for tobacco products' office of science. *American Journal of Preventive Medicine* 43, 5 SUPPL. 3 (11 2012), S255–S263. doi:10.1016/j.amepre.2012.08.004
- [12] Associated Press. 2024. TikTok Knew Young User Safeguards Were Ineffective, Legal Document Shows. <https://www.inc.com/associated-press/tiktok-knew-young-user-safeguards-were-ineffective-legal-document-shows/90988443>
- [13] Karla Badillo-Urquiola, Zainab Agha, Danielle Abaquita, Scott B. Harpin, and Pamela J. Wisniewski. 2024. Towards a Social Ecological Approach to Supporting Caseworkers in Promoting the Online Safety of Youth in Foster Care. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (4 2024). doi:10.1145/3637412
- [14] Karla Badillo-Urquiola, Zachary Shea, Zainab Agha, Irina Lediaeva, and Pamela Wisniewski. 2021. Conducting Risky Research with Teens: Co-designing for the Ethical Treatment and Protection of Adolescents. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (1 2021). doi:10.1145/3432930
- [15] Karla Badillo-Urquiola, Diva Smriti, Brenna McNally, Evan Golub, Elizabeth Bonsignore, and Pamela J Wisniewski. 2019. "Stranger Danger!" Social Media App Features Co-designed with Children to Keep Them Safe Online. *Proceedings of the 18th ACM International Conference on Interaction Design and Children* (2019). doi:10.1145/3311927
- [16] Ryan Barwick. 2025. As concern about online teen safety mounts, Meta re-ups its advertising offensive. <https://www.marketingbrew.com/stories/2025/03/27/meta-instagram-advertising-campaign-teen-accounts-safety>
- [17] Anaya Baxter. 2024. New Instagram features will not keep children safe online. <https://thedailycougar.com/2024/10/09/new-instagram-features-are-not-the-answer-to-childrens-safety/>
- [18] Arturo Béjar, Cybersecurity for Democracy, Fairplay, Molly Rose Foundation, and ParentsSOS. 2025. *Teen Accounts, Broken Promises: How Instagram Is Failing to Protect Minors*. Technical Report. <https://fairplayforkids.org/wp-content/uploads/2025/09/Teen-Accounts-Broken-Promises-How-Instagram-is-failing-to-protect-minors.pdf>

- [19] C. Anthony Di Benedetto. 1999. Identifying the Key Success Factors in New Product Launch. *Journal of Product Innovation Management* 16, 6 (11 1999), 530–544. doi:10.1111/1540-5885.1660530
- [20] James Beser. 2021. A new choice for parents of tweens and teens on YouTube. <https://blog.youtube/news-and-events/supervised-experiences-for-families-on-youtube/>
- [21] Sophie C. Boerman and Eva A. van Reijmersdal. 2016. Informing Consumers about “Hidden” Advertising: A Literature Review of the Effects of Disclosing Sponsored Content. *Advertising in New Formats and Media* (12 2016), 115–146. doi:10.1108/978-1-78560-313-620151005
- [22] Sophie C. Boerman and Eva A. van Reijmersdal. 2020. Disclosing Influencer Marketing on YouTube to Children: The Moderating Role of Para-Social Relationship. *Frontiers in Psychology* 10 (1 2020), 489721. doi:10.3389/fpsyg.2019.03042
- [23] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (1 2006), 77–101. doi:10.1191/1478088706qp0630a
- [24] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (8 2019), 589–597. doi:10.1080/2159676X.2019.1628806
- [25] Xavier V. Caddle, Sarvech Qadir, Charles Hughes, Elizabeth A. Sweigart, Jinkyung Katie Park, and Pamela J. Wisniewski. 2025. Building a Village: A Multi-stakeholder Approach to Open Innovation and Shared Governance to Promote Youth Online Safety. *Proceedings of Computer Supported Cooperative Work and Social Computing (CSCW’25)* 1 (4 2025). doi:10.1145/3757459
- [26] Fong Ching Chang, Chiung Hui Chiu, Nae Fang Miao, Ping Hung Chen, Ching Mei Lee, Jeng Tung Chiang, and Ying Chun Pan. 2015. The relationship between parental mediation and Internet addiction among adolescents, and the association with cyberbullying and depression. *Comprehensive Psychiatry* 57 (2 2015), 21–28. doi:10.1016/J.COMPPSYCH.2014.11.013
- [27] Vivian Hsueh Hua Chen and Grace S. Chng. 2016. Active and restrictive parental mediation over time: Effects on youths’ self-regulatory competencies and impulsivity. *Computers & Education* 98 (7 2016), 206–212. doi:10.1016/J.COMPEDU.2016.03.012
- [28] Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70, 4 (10 1968), 213–220. doi:10.1037/H0026256
- [29] Lauren Collins, Allison M. Glasser, Haneen Abudayyeh, Jennifer L. Pearson, and Andrea C. Villanti. 2019. E-Cigarette Marketing and Communication: How E-Cigarette Companies Market E-Cigarettes and the Public Engages with E-cigarette Information. *Nicotine & Tobacco Research* 21, 1 (1 2019), 14–24. doi:10.1093/NTR/NTX284
- [30] Common Sense Media. 2022. The Common Sense Census: Media Use by Tweens and Teens, 2021. <https://www.common SenseMedia.org/research/the-common-sense-census-media-use-by-tweens-and-teens-2021>
- [31] Marijke De Veirman, Liselot Hudders, and Michelle R. Nelson. 2019. What Is Influencer Marketing and How Does It Target Children? A Review and Direction for Future Research. *Frontiers in Psychology* 10 (12 2019), 498106. doi:10.3389/fpsyg.2019.02685
- [32] Prema Dev, Jessica Medina, Zainab Agha, Munmun De Choudhury, Afsaneh Razi, and Pamela J. Wisniewski. 2022. From Ignoring Strangers’ Solicitations to Mutual Sexting with Friends: Understanding Youth’s Online Sexual Risks in Instagram Private Conversations. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (11 2022), 94–97. doi:10.1145/3500868.3559469
- [33] Prakriti Dumar and Nasrullah Al-Ameen. 2024. “It’s like educating us older people...”: Unveiling Needs and Expectations Regarding Educational Features within Parental Control Tools. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (2024). doi:10.1145/3613905
- [34] Prakriti Dumar, Hanieh Atashpanjeh, and Mahdi Nasrullah Al-Ameen. 2024. “It’s hard for him to make choices sometimes and he needs guidance”: Re-orienting Parental Control for Children. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (4 2024). doi:10.1145/3637359
- [35] Laura Edelson. 2024. *Teen Experiences on Social Media*. Technical Report. <https://cybersecurityfordemocracy.org/teen-experiences-on-social-media>
- [36] Tracy Elizabeth. 2021. Our work to design an age-appropriate experience on TikTok. <https://newsroom.tiktok.com/en-us/our-work-to-design-an-age-appropriate-experience-on-tiktok>
- [37] Nathaniel J. Evans, Mariea Grubbs Hoy, and Courtney Carpenter Childers. 2018. Parenting “YouTube Natives”: The Impact of Pre-Roll Advertising and Text Disclosures on Parental Responses to Sponsored Child Influencer Videos. *Journal of Advertising* 47, 4 (10 2018), 326–346. doi:10.1080/00913367.2018.1544952
- [38] Michelle Faverio and Olivia Sidoti. 2024. Teens, Social Media and Technology 2024. <https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/>
- [39] Federal Trade Commission. 1999. Children’s Online Privacy Protection Rule - 16 CFR Part 312. <https://www.ftc.gov/legal-library/browse/federal-register-notices/childrens-online-privacy-protection-rule-16-cfr-part-312>
- [40] Federal Trade Commission. 2019. Google and YouTube Will Pay Record \$170 Million for Alleged Violations of Children’s Privacy Law. <https://www.ftc.gov/news-events/news/press-releases/2019/09/google-youtube-will-pay->

record-170-million-alleged-violations-childrens-privacy-law

- [41] Federal Trade Commission. 2019. Musical.ly, Inc.
- [42] Federal Trade Commission. 2023. FTC Proposes Blanket Prohibition Preventing Facebook from Monetizing Youth Data. <https://www.ftc.gov/news-events/news/press-releases/2023/05/ftc-proposes-blanket-prohibition-preventing-facebook-monetizing-youth-data>
- [43] Mikkel Flyverbom. 2016. Digital Age| Transparency: Mediation and the Management of Visibilities. *International Journal of Communication* 10, 0 (1 2016), 13. <https://ijoc.org/index.php/ijoc/article/view/4490>
- [44] Archon Fung, Mary Graham, and David Weil. 2007. *Full Disclosure: The Perils and Promise of Transparency*.
- [45] Arup Kumar Ghosh, Charles E. Hughes, and Pamela J. Wisniewski. 2020. Circle of Trust: A New Approach to Mobile Online Safety for Families. *Conference on Human Factors in Computing Systems - Proceedings* (4 2020). doi:10.1145/3313831.3376747
- [46] Cami Goray and Sarita Schoenebeck. 2022. Youths' Perceptions of Data Collection in Online Advertising and Social Media. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (11 2022). doi:10.1145/3555576
- [47] James E. Grunig, Larissa A. Grunig, K. Sriramesh, Yi Hui Huang, and Anastasia Lyra. 1995. Models of Public Relations in an International Setting. *Journal of Public Relations Research* 7, 3 (7 1995), 163–186. doi:10.1207/S1532754XJPRR0703_01
- [48] Joseph P. Guiltinan. 1999. Launch Strategy, Launch Tactics, and Demand Outcomes. *Journal of Product Innovation Management* 16, 6 (11 1999), 509–529. doi:10.1111/1540-5885.1660027
- [49] Eric Han. 2020. Protecting against exploitative content. <https://newsroom.tiktok.com/en-us/protecting-against-exploitative-content>
- [50] 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 (4 2021), 1–31. doi:10.1145/3449116
- [51] E. J. Hultink, A. Griffin, S. Hart, and H. S.J. Robben. 1997. Industrial new product launch strategies and product development performance. *Journal of Product Innovation Management* 14, 4 (7 1997), 243–257. doi:10.1016/S0737-6782(97)00009-X
- [52] Ikhti Imaniah, Nurul Fitria, Kumala Dewi, and Akhmad Zakky. 2020. Youtube kids channels in developing young children's communication skills in English: parents' beliefs, attitudes, and behaviors. *IJLECR (International Journal of Language Education and Cultural Review)* 6, 1 (6 2020), 20–30. doi:10.21009/IJLECR.061.03
- [53] Information Commissioner's Office. 2021. Age appropriate design: a code of practice for online services. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services/>
- [54] Instagram. 2020. Safer Internet Day 2020. <https://about.instagram.com/blog/announcements/safer-internet-day-2020>
- [55] Instagram. 2021. Asking People for Their Birthdays. <https://about.instagram.com/blog/announcements/asking-people-for-their-birthdays>
- [56] Instagram. 2022. How to See More of What You Want on Instagram. <https://about.instagram.com/blog/tips-and-tricks/control-your-instagram-feed>
- [57] Instagram. 2022. Updates to the Sensitive Content Control. <https://about.instagram.com/blog/announcements/updates-to-the-sensitive-content-control>
- [58] Instagram. 2023. Giving People Control Over Their Time and What They See on Instagram. <https://about.instagram.com/blog/announcements/new-ways-to-control-what-you-see-on-instagram>
- [59] Instagram. 2024. About Instagram teen privacy and safety settings. <https://help.instagram.com/3237561506542117>
- [60] Instagram. 2024. Instagram Teen Accounts. <https://about.instagram.com/community/teen-accounts>
- [61] Instagram. 2024. Teen Accounts: Protections for Teens, Peace of Mind for Parents. <https://about.instagram.com/blog/announcements/instagram-teen-accounts>
- [62] Instagram. 2024. Tips for Parents. <https://help.instagram.com/154475974694511>
- [63] Instagram. 2025. Manage your teen safety settings. <https://help.instagram.com/455298907304715/>
- [64] Instagram Help Center. 2025. Nudity protection in chats on Instagram. https://www.facebook.com/help/instagram/503437025160040?cms_platform=android-app
- [65] Eden Kamar, Roberta Liggett O'Malley, C. Jordan Howell, David Maimon, and Dekel Shabat. 2025. "Cutie, click on the link". *Computers in Human Behavior* 162 (1 2025). doi:10.1016/J.CHB.2024.108454
- [66] Cormac Keenan. 2022. More ways for our community to enjoy what they love. <https://newsroom.tiktok.com/en-us/more-ways-for-our-community-to-enjoy-what-they-love>
- [67] Cormac Keenan. 2023. New features for teens and families on TikTok. <https://newsroom.tiktok.com/en-us/new-features-for-teens-and-families-on-tiktok-us>
- [68] Lucyna Kirwil. 2009. Parental Mediation Of Children's Internet Use In Different European Countries. *Journal of Children and Media* 3, 4 (11 2009), 394–409. doi:10.1080/17482790903233440

- [69] Rene F. Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2390–2395. doi:10.1145/2858036.2858402
- [70] Priya Kumar, Shalmali Milind Naik, Utkarsha Ramesh Devkar, Marshini Chetty, Tamara L. Clegg, and Jessica Vitak. 2017. 'No Telling Passcodes Out Because They're Private'. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (12 2017). doi:10.1145/3134699
- [71] Sook Jung Lee. 2013. Parental restrictive mediation of children's internet use: Effective for what and for whom? *New Media and Society* 15, 4 (6 2013), 466–481. doi:10.1177/1461444812452412
- [72] Yikuan Lee and Gina Colarelli O'Connor. 2003. The Impact of Communication Strategy on Launching New Products: The Moderating Role of Product Innovativeness. *Journal of Product Innovation Management* 20, 1 (1 2003), 4–21. doi:10.1111/1540-5885.T01-1-201002
- [73] Luke Legner. 2022. Kids' screen time: How much is too much? . <https://www.osfhealthcare.org/blog/kids-screen-time-how-much-is-too-much/>
- [74] María E. Len-Ríos, Hilary E. Hughes, Laura G. McKee, and Henry N. Young. 2016. Early adolescents as publics: A national survey of teens with social media accounts, their media use preferences, parental mediation, and perceived Internet literacy. *Public Relations Review* 42, 1 (3 2016), 101–108. doi:10.1016/J.PUBREV.2015.10.003
- [75] Sonia Livingstone and Leslie Haddon. 2008. Risky Experiences for Children Online: Charting European Research on Children and the Internet. *Children & Society* 22, 4 (7 2008), 314–323. doi:10.1111/J.1099-0860.2008.00157.X
- [76] Sonia Livingstone and Ellen Helsper. 2009. Balancing opportunities and risks in teenagers' use of the internet: the role of online skills and internet self-efficacy. *New media & society* 12, 2 (11 2009), 309–329. doi:10.1177/1461444809342697
- [77] Sonia Livingstone and Ellen J. Helsper. 2008. Parental Mediation of Children's Internet Use. *Journal of Broadcasting & Electronic Media* 52, 4 (10 2008), 581–599. doi:10.1080/08838150802437396
- [78] Sonia Livingstone, Lucyna Kirwil, Cristina Ponte, and Elisabeth Staksrud. 2014. In their own words: What bothers children online? *European Journal of Communication* 29, 3 (3 2014), 271–288. doi:10.1177/0267323114521045
- [79] Sonia Livingstone, Kjartan Ólafsson, Ellen J. Helsper, Francisco Lupiáñez-Villanueva, Giuseppe A. Veltri, and Frans Folkvord. 2017. Maximizing Opportunities and Minimizing Risks for Children Online: The Role of Digital Skills in Emerging Strategies of Parental Mediation. *Journal of Communication* 67, 1 (2 2017), 82–105. doi:10.1111/JCOM.12277
- [80] Sonia Livingstone and Mariya Stoilova. 2021. The 4Cs: Classifying Online Risk to Children. (*CO-RE Short Report Series on Key Topics*). Hamburg: Leibniz-Institut für Medienforschung | Hans-Bredow-Institut (HBI) (2021).
- [81] Renkai Ma, Yao Li, and Yubo Kou. 2023. Transparency, Fairness, and Coping: How Players Experience Moderation in Multiplayer Online Games. *Conference on Human Factors in Computing Systems - Proceedings* (4 2023). doi:10.1145/3544548.3581097
- [82] Renkai Ma, Zinan Zhang, Xinning Gui, and Yubo Kou. 2024. Labeling in the Dark: Exploring Creators' and Audience's Experiences with Classifying Content for Child Privacy Protection on YouTube. In *Designing Interactive Systems Conference (DIS '24)*. doi:10.1145/3643834.3661565
- [83] 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. *Conference on Human Factors in Computing Systems - Proceedings* (4 2020). doi:10.1145/3313831.3376666
- [84] Sarah McRoberts, Elizabeth Bonsignore, Tamara Peyton, and Svetlana Yarosh. 2016. "Do it for the viewers!" Audience engagement behaviors of young YouTubers. *Proceedings of IDC 2016 - The 15th International Conference on Interaction Design and Children* (6 2016), 334–343. doi:10.1145/2930674.2930676
- [85] Meta. 2023. How Meta products and tools support age-appropriate experiences. (2023).
- [86] Meta. 2024. Combating Financial Sextortion Scams From Nigeria. <https://about.fb.com/news/2024/07/combating-financial-sextortion-scams-from-nigeria/>
- [87] Meta. 2024. Introducing Stricter Message Settings for Teens on Instagram and Facebook. <https://about.fb.com/news/2024/01/introducing-stricter-message-settings-for-teens-on-instagram-and-facebook/>
- [88] Meta. 2024. New Tools to Help Protect Against Sextortion and Intimate Image Abuse. <https://about.instagram.com/blog/announcements/new-tools-to-help-protect-against-sextortion-and-intimate-image-abuse>
- [89] Melinda Wenner Moyer. 2022. Kids Are Using Social Media More Than Ever, Study Finds. <https://www.nytimes.com/2022/03/24/well/family/child-social-media-use.html>
- [90] Manoj Niverthi, Gaurav Verma, and Srijan Kumar. 2022. Characterizing, Detecting, and Predicting Online Ban Evasion. *WWW 2022 - Proceedings of the ACM Web Conference 2022* (4 2022), 2614–2623. doi:10.1145/3485447.3512133
- [91] 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 (4 2024). doi:10.1145/3637413
- [92] Richard J. O'Connor, M. Jane Lewis, Sarah E. Adkison, Maansi Bansal-Travers, and K. Michael Cummings. 2017. Perceptions of "Natural" and "Additive-Free" Cigarettes and Intentions to Purchase. *Health Education and Behavior*

- 44, 2 (4 2017), 222–226. doi:10.1177/1090198116653935
- [93] Jessica K. Pepper, Sherry L. Emery, Kurt M. Ribisl, Brian G. Southwell, and Noel T. Brewer. 2014. Effects of advertisements on smokers' interest in trying e-cigarettes: the roles of product comparison and visual cues. *Tobacco Control* 23, suppl 3 (7 2014), iii31–iii36. doi:10.1136/TOBACCONCONTROL-2014-051718
- [94] Mayana Pereira, Rahul Dodhia, Hyrum Anderson, and Richard Brown. 2024. Metadata-Based Detection of Child Sexual Abuse Material. *IEEE Transactions on Dependable and Secure Computing* 21, 4 (2024), 3153–3164. doi:10.1109/TDSC.2023.3324275
- [95] Pallav Pokhrel, Pebbles Fagan, Lisa Kehl, and Thaddeus A. Herzog. 2015. Receptivity to E-cigarette Marketing, Harm Perceptions, and E-cigarette Use. *American Journal of Health Behavior* 39, 1 (1 2015), 121–131. doi:10.5993/AJHB.39.1.13
- [96] Michael E. Porter and Mark R. Kramer. 2006. Strategy and Society: The Link between Competitive Advantage and Corporate Social Responsibility. *Harvard Business Review* (2006), 78–92.
- [97] 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 (4 2023), 29. doi:10.1145/3579522
- [98] Afsaneh Razi, Karla Badillo-Urquiola, and Pamela J. Wisniewski. 2020. Let's Talk about Sext: How Adolescents Seek Support and Advice about Their Online Sexual Experiences. *Conference on Human Factors in Computing Systems - Proceedings* (4 2020). doi:10.1145/3313831.3376400
- [99] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J. Wisniewski. 2021. A Human-Centered Systematic Literature Review of the Computational Approaches for Online Sexual Risk Detection. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 38. doi:10.1145/3479609
- [100] Renkai Ma, Yao Li, Sunhye Bai, Yubo Kou, and Xinning Gui. 2025. Weighing Benefits and Harms: Parental Mediation on Social Video Platforms. In *ACM CHI '25*. doi:10.1145/3706598.3713422
- [101] Hyejoon Rim and Sora Kim. 2016. Dimensions of corporate social responsibility (CSR) skepticism and their impacts on public evaluations toward CSR. *Journal of Public Relations Research* 28, 5-6 (11 2016), 248–267. doi:10.1080/1062726X.2016.1261702
- [102] Isabel Rodríguez-de Dios, Johanna M.F. van Oosten, and Juan José Igartua. 2018. A study of the relationship between parental mediation and adolescents' digital skills, online risks and online opportunities. *Computers in Human Behavior* 82 (5 2018), 186–198. doi:10.1016/J.CHB.2018.01.012
- [103] Kalhan Rosenblatt, Daysia Tolentino, Jason Abbruzzese, Saba Hamedy, Ben Goggin, Kate Snow, Jonathan Vanian, and Lora Kolodny. 2024. Senate hearing highlights: Lawmakers grill CEOs from TikTok, X and Meta about online child safety. <https://www.nbcnews.com/tech/live-blog/senate-hearing-online-child-safety-big-tech-live-updates-rcna136235>
- [104] Alexandra Ruiz-Gomez, Tama Leaver, and Crystal Abidin. 2022. Playing YouTube: How the Nancy YouTuber doll and app position children as aspiring YouTube influencers. *International Journal of Cultural Studies* 25, 2 (3 2022), 121–140. doi:10.1177/13678779211063222
- [105] Hagit Sasson and Gustavo Mesch. 2014. Parental mediation, peer norms and risky online behavior among adolescents. *Computers in Human Behavior* 33 (4 2014), 32–38. doi:10.1016/J.CHB.2013.12.025
- [106] Sarita Schoenebeck, Carol F. Scott, Emma Grace Hurley, Tammy Chang, and Ellen Selkie. 2021. Youth Trust in Social Media Companies and Expectations of Justice. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (4 2021), 1–18. doi:10.1145/3449076
- [107] Senate. 2023. S.1409 - Kids Online Safety Act. <https://www.congress.gov/bill/118th-congress/senate-bill/1409/text>
- [108] Wonsun Shin and Nurzali Ismail. 2014. Exploring the Role of Parents and Peers in Young Adolescents' Risk Taking on Social Networking Sites. *Cyberpsychology, Behavior, and Social Networking* 17, 9 (9 2014), 578–583. doi:10.1089/CYBER.2014.0095
- [109] Wonsun Shin and May O. Lwin. 2017. How does “talking about the Internet with others” affect teenagers' experience of online risks? The role of active mediation by parents, peers, and school teachers. *New Media and Society* 19, 7 (7 2017), 1109–1126. doi:10.1177/1461444815626612
- [110] Crystal R. Smit, Laura Buijs, Thabo J. van Woudenberg, Kirsten E. Bevelander, and Moniek Buijzen. 2020. The Impact of Social Media Influencers on Children's Dietary Behaviors. *Frontiers in Psychology* 10 (1 2020), 498050. doi:10.3389/fpsyg.2019.02975
- [111] Snapchat. 2023. How We Keep Our Snapchat Community Safe & Informed. <https://help.snapchat.com/hc/en-us/articles/7012398850836-How-We-Keep-Our-Snapchat-Community-Safe-Informed>
- [112] Snapchat. 2023. New features and transparency measures for Snapchatters in the European Union to comply with the Digital Services Act. <https://newsroom.snap.com/digital-services-act-snap>

- [113] Statista. 2023. Leading social apps children U.S. 2023. <https://www.statista.com/statistics/1339193/leading-social-apps-children-us/>
- [114] Kaiwen Sun, Hayley Simon, Carlo Sugatan, Susan A Gelman, Florian Schaub, Tanisha Afnan, and Jenny Radesky. 2021. They see you're a girl if you pick a pink robot with a skirt: A qualitative study of how children conceptualize data processing and digital privacy risks. *Conference on Human Factors in Computing Systems - Proceedings* (5 2021). doi:10.1145/3411764.3445333
- [115] TikTok. 2019. TikTok for Younger Users. <https://newsroom.tiktok.com/en-us/tiktok-for-younger-users>
- [116] TikTok. 2019. TikTok's Top 10 Tips for Parents. <https://newsroom.tiktok.com/en-us/tiktoks-top-10-tips-for-parents>
- [117] TikTok. 2021. Strengthening privacy and safety for youth on TikTok. <https://newsroom.tiktok.com/en-us/strengthening-privacy-and-safety-for-youth>
- [118] TikTok. 2022. Strengthening enforcement of sexually suggestive content. <https://newsroom.tiktok.com/en-us/strengthening-enforcement-of-sexually-suggestive-content>
- [119] TikTok. 2023. Updates to ads for teens and improved data control and transparency tools. <https://www.tiktok.com/business/en-US/blog/privacy-updates-improved-data-control-transparency-tools>
- [120] TikTok. 2023. Updating Family Pairing and establishing TikTok's Youth Council. <https://newsroom.tiktok.com/en-us/updating-family-pairing-and-establishing-tiktoks-youth-council>
- [121] TikTok. 2024. TikTok CEO Shou Chew's Opening Statement - Senate Judiciary Committee Hearing on Online Child Sexual Exploitation Crisis. <https://newsroom.tiktok.com/en-us/opening-statement-senate-judiciary-committee-hearing>
- [122] TikTok. 2024. TikTok highlights online safety features in new #SwipeOutHate advertising campaign during Guinness Six Nations. <https://newsroom.tiktok.com/en-gb/tiktok-highlights-online-safety-features-in-new-swipeouthate-advertising-campaign-during-guinness-six-nations>
- [123] TikTok. 2025. Direct messages. <https://support.tiktok.com/en/using-tiktok/messaging-and-notifications/direct-message-settings>
- [124] TikTok. 2025. Guardian's guide. <https://www.tiktok.com/safety/en/guardians-guide>
- [125] Jonathan Vanian. 2024. Meta's Zuckerberg apologizes to victims in tense D.C. hearing: 'I'm sorry for everything you've gone through'. <https://www.cnn.com/2024/01/31/ceos-of-meta-tiktok-x-and-other-social-media-firms-to-testify-at-hearing-on-child-safety.html>
- [126] Ge Wang, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2021. Protection or Punishment? Relating the Design Space of Parental Control Apps and Perceptions about Them to Support Parenting for Online Safety. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021). doi:10.1145/3476084
- [127] Ge Wang, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2022. 'Don't make assumptions about me!': Understanding Children's Perception of Datafication Online. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (11 2022). doi:10.1145/3555144
- [128] Ge Wang, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2023. 'Treat me as your friend, not a number in your database': Co-designing with Children to Cope with Datafication Online. *Conference on Human Factors in Computing Systems - Proceedings* (4 2023). doi:10.1145/3544548.3580933
- [129] Shannon Ward, Tara M. Dumas, Ankur Srivastava, Jordan P. Davis, and Wendy Ellis. 2021. Uploading Risk: Examining the Social Profile of Young Adults Most Susceptible to Engagement in Risky Social Media Challenges. *Cyberpsychology, Behavior, and Social Networking* 24, 12 (12 2021), 846–850. doi:10.1089/CYBER.2020.0846
- [130] Pamela Wisniewski, Arup Kumar Ghosh, Heng Xu, Mary Beth Rosson, and John M. Carroll. 2017. Parental control vs. teen self-regulation: Is there a middle ground for mobile online safety? *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (2 2017), 51–69. doi:10.1145/2998181.2998352
- [131] Ruth Wodak. 2022. Critical Linguistics and Critical Discourse Analysis. (10 2022), 426–443. doi:10.1075/HOP.M2.CRI1/HTML
- [132] Claes Wohlin. 2014. Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14* (2014). doi:10.1145/2601248
- [133] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. 2022. Sensemaking, Support, Safety, Retribution, Transformation: A Restorative Justice Approach to Understanding Adolescents' Needs for Addressing Online Harm. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* 15 (4 2022). doi:10.1145/3491102.3517614
- [134] YouTube. 2015. YouTube Kids. <https://support.google.com/youtube/answer/6173147>
- [135] YouTube. 2019. An update on our efforts to protect minors and families. <https://blog.youtube/news-and-events/an-update-on-our-efforts-to-protect/>
- [136] YouTube. 2019. More updates on our actions related to the safety of minors on YouTube. <https://blog.youtube/news-and-events/more-updates-on-our-actions-related-to/>

- [137] YouTube. 2020. Frequently asked questions about “made for kids”. <https://support.google.com/youtube/answer/9684541>
- [138] YouTube Kids. 2019. Resources, Tools, and Control Settings for Parents. <https://www.youtube.com/kids/parent-resources/>
- [139] 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 (11 2022). doi:10.1145/3555608
- [140] Jun Zhao, Ge Wang, Carys Dally, Petr Slovak, Julian Edbrooke-Childs, Max Van Kleek, and Nigel Shadbolt. 2019. ‘I make up a silly name’: Understanding children’s perception of privacy risks online. *Conference on Human Factors in Computing Systems - Proceedings* (5 2019). doi:10.1145/3290605.3300336
- [141] Wei Zhao, Ryan M. Kelly, Melissa J. Rogerson, and Jenny Waycott. 2024. Older Adults Imagining Future Technologies in Participatory Design Workshops: Supporting Continuity in the Pursuit of Meaningful Activities. *Conference on Human Factors in Computing Systems - Proceedings* 18, 24 (5 2024). doi:10.1145/3613904.3641887
- [142] Marc A. Zimmerman, Sarah A. Stoddard, Andria B. Eisman, Cleopatra H. Caldwell, Sophie M. Aiyer, and Alison Miller. 2013. Adolescent Resilience: Promotive Factors That Inform Prevention. *Child Development Perspectives* 7, 4 (12 2013), 215–220. doi:10.1111/CDEP.12042
- [143] Diana Zulli and David James Zulli. 2022. Extending the Internet meme: Conceptualizing technological mimesis and imitation publics on the TikTok platform. *New Media and Society* 24, 8 (8 2022), 1872–1890. doi:10.1177/1461444820983603

A Codebook for RQ2: Assessing Problematic Communication Practices

Received May 2025; revised November 2025; accepted December 2025

Table 3. Codebook for reflexive thematic analysis of the implementation or effectiveness of their described youth safety features

Theme	Sub-theme	Example Quote(s)
Safety Feature Implementation Claims Often Mismatch Rollout Reality	Provisional and evolving language makes it unclear whether features are implemented	<i>(2024 Snap):</i> "Parental tools: Soon, in Family Center... parents will be able to see if their 16- and 17-year old teens have an active public story..." <i>(2024 Snap):</i> "...Families will soon be able to choose up to three specific locations... and parents will receive notifications when their family member departs from or arrives..."
	Stated geographic limits make it unclear where features are implemented	<i>(2024 Meta):</i> "Alternate Topic Nudge Teens under 18 in certain countries will see a notification that encourages them to switch to a different topic if they've been looking at the same type of content on Explore for a while."
How Platforms' Safety Features Operate Remains Unclear and Inconsistent	The conditions triggering safety features to activate remain unclear and inconsistent	<i>(2023 Snap):</i> "If you're a parent over the age of 25, you can use Family Center..." <i>(2024 Snap):</i> "Family Center will provide parents over the age of 25 the ability to: See which Snapchat friends..."
	How safety features intersect with each other remains unclear	<i>(2023 Snap):</i> "...Easily and confidentially report any concerns directly to our Trust and Safety team" (Context is on Family Center). <i>(2024 Snap):</i> "When you report an account through Family Center, either from the 'Recent Conversations' section or from the 'Friends List,' you are taken to the account reporting page..." <i>(2024 Snap):</i> "...Easily and confidentially report any accounts parents may be concerned about directly to our 24/7 Trust and Safety team."
	Rationales for varying restriction levels regarding similar features remain unclear.	<i>(TT):</i> "...only people over 18 are allowed to host a livestream." <i>(TT):</i> "Creators must be 18+ to go LIVE..." <i>(TT):</i> "your teen has the control to decide who can make duets with or react to their content" <i>(TT):</i> "we give you the option to decide who can make duets with or react to your content."
How Platforms Present Safety Features Fails to Exhibit Effective Risk Mitigation	How platforms communicate about risks leaves the targeted risks unclear	<i>(Meta):</i> "...The Sensitive Content Control has three options... "Standard" is the default state... "More" enables people to see more sensitive content." <i>(Meta):</i> "...you may see content that doesn't break the rules, but could be upsetting... you can adjust the Sensitive Content Control..."
	Safety features' effectiveness in addressing risks lacks evidence	<i>(YT, 2019):</i> "suspend comments on videos featuring young minors... we had been working on an even more effective classifier... We accelerated its launch and now have a new comments classifier... that will detect and remove 2X more individual predatory comments."
	How platforms claim safety features work mismatches the scope of the risks they claim to address	<i>(YT, 2020):</i> "As a YouTube creator, you are required to set future and existing videos as made for kids or not... This will help ensure that we offer the appropriate features on your content." <i>(YT, 2020):</i> "Regardless of your location, you're legally required to comply with the Children's Online Privacy Protection Act... You're required to tell us that your videos are made for kids..."