Personally Targeted Risk vs. Humor: How Online Risk Perceptions of Youth vs. Third-Party Annotators Differ based on Privately Shared Media on Instagram

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

Joshua Gracie
University of Central Florida
Orlando, USA
joshua_gracie@knights.ucf.edu

Ashwaq Alsoubai
Vanderbilt University
Nashville, USA
ashwaq.alsoubai@vanderbilt.edu

Afsaneh Razi
Drexel University
Philadelphia, USA
afsaneh.razi@drexel.edu

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

ABSTRACT

While risk is highly subjective, especially when it comes to the private online interactions of youth, third-party annotations are often performed to identify risky content. Therefore, we conducted a mixed-methods study to examine if, how, and why risk perceptions might differ between youth and third-party annotators who were research assistants (RAs). We first asked 100 youth to share their Instagram private messages and flag media that made them feel unsafe. Then, we had RAs annotate the same media to identify what they thought was unsafe or risky. Compared to RAs, youth tended to flag images as risky when they perceived targeted harassment towards them or unwanted solicitations from strangers. In contrast, RAs were more likely to risk-flag sexual images with a humorous undertone shared among friends. Our findings highlight the differences between how online risks are perceived by youth compared to RAs. We provide recommendations for assessing online risks based on multiple perspectives to inform future youth-centered risk mitigation approaches.

Content Warning: *Sensitive topics, including sexual risk involving minors, are discussed in this paper. Readers should use their discretion as to whether they would like to proceed.*

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Empirical studies in HCI**;

KEYWORDS

Youth, Online Risk, Instagram, Private Message, Media Risk

ACM Reference Format:
Jinkyung Park, Joshua Gracie, Ashwaq Alsoubai, Afsaneh Razi, and Pamela J. Wisniewski. 2024. Personally Targeted Risk vs. Humor: How Online Risk Perceptions of Youth vs. Third-Party Annotators Differ based on Privately Shared Media on Instagram. In *Interaction Design and Children (IDC '24), June 17–20, 2024, Delft, Netherlands.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3628516.3655799

IDC '24, June 17–20, 2024, Delft, Netherlands

ACM ISBN 979-8-4007-0442-0/24/06
https://doi.org/10.1145/3628516.3655799

1 INTRODUCTION

While being on social media provides youth space for social connection, creativity, and peer support [5], the prevalence of risks associated with multimedia-based content (e.g., images, videos) has grown [48]. Risky media sharing practices on social media can range from posting an image of someone else without their consent [40], sharing a meme with the intent to embarrass or harass someone else [62], using someone else's photos to create fake profiles [58], or sharing explicit photos or pornography [66]. Given the massive scale of online content, researchers have been increasingly applying computational approaches to detect youth online risks, such as cyberbullying [29, 36, 59, 60] and sexually explicit and/or risky content [9, 48, 64, 65], many of which rely on the perspective of third persons (e.g., crowd-sourced workers) in labeling ground-truth data (e.g., labeling messages as risk vs. non-risky) [29, 54, 59]. Third-party annotations are collaborative tasks often performed in academia for training machine learning (ML) classifiers/algorithms (i.e., preparing the ground-truth data). However, little scrutiny is given to the ecological validity and implications of using such an approach for determining ground truth. As risk is highly subjective [15], understanding differences between youth risk perceptions and those of third-party annotators has far-reaching implications for the real-world ML-based systems deployed based on such translational research and how they impact people in real-world applications. Thus, we tackle the issue of the ecological validity of third-party annotations in designing youth-centered risk prevention programs and detection technologies. To do this we asked the following high-level research questions:

- **RQ1:** *What are the characteristics of risks youth experienced privately through media shared via Instagram Direct Messages (DMs)? Do the risk perceptions of youth versus research assistants significantly differ?*
- **RQ2:** *Based on the trends in differences, what are the key themes that help explain the differences in risk perceptions between youth and research assistants?*

To answer the research questions, we conducted a user study where youth (ages 13-21) donated their private Instagram conversations and flagged their own conversations for risk. Then we had research assistants (RAs) annotate the same conversations for risks. In this study, risky was defined as any conversation that made our participants or someone else feel uncomfortable or unsafe. We gave this same definition to youth and RAs when annotating for risky conversations. Next, we analyzed risky media messages (N = 674) privately received or sent by 42 participants on Instagram, which were identified by either the youth themselves or RAs. To answer RQ1, we examined frequencies of unsafe media flagged by youth versus RAs in terms of risk annotations (e.g., risk types and/or relationship types with the sender) and risk context (e.g., media content type and the nature of risks), and conducted Fisher's exact tests [21] to identify statistically significant associations between these categorical codes. Then, we qualitatively examined the content of the unsafe media messages to explore how and why the participants' perspectives and the research assistants' perspectives differ (RQ2). We found several statistically significant differences in the risk perceptions of youth and RAs. For instance, youth were more attuned to personally targeted attacks (e.g., harassment), while RAs were sensitive to sexual messages with a humorous undertone. While youth flagged messages sent from strangers as risky, RAs flagged more risky messages exchanged between friends. Also, youth did not flag their own risky behavior as unsafe (e.g., smoking marijuana or drinking), while RAs considered these behaviors risky.

Our results confirmed that compared to the youth (first-person), RAs (third-person) annotated significantly larger numbers of media messages as risky. This tendency of RAs' overrating as third-person was salient, especially when annotating sexual memes for risk, where RAs flagged many of the sexual jokes between youth and their peers as risky when youth themselves did not. Such a trend was noteworthy given that sexual interaction with peers is not necessarily risky and rather considered a natural part of adolescent development. By uncovering these differing risk perceptions, we address an important issue at the intersections of Interaction Design and Children (IDC), Human-Computer Interaction (HCI), and ML automated risk detection research for youth online safety. Our study makes the following contributions to the broader SIGCHI community:

- Advances the understanding of what media youth privately share on social media (i.e., Instagram) and their risk experience by examining unsafe media messages that were flagged by the youth themselves.
- Discovers key differences between how online risks are perceived by youth (first person) compared to research assistants (third person).
- Challenges the common practices (i.e., reliance on third-person) in the ML community for annotating ground truth data to build automated risk detection systems.
- Provides recommendations for assessing online risks based on multiple perspectives and respectfully designing sociotechnical systems to provide safer experiences for youth.

We offer important insights into the shortcomings of relying solely on third-party annotations and underscore the need to incorporate youth perspectives for designing more effective online risk detection models. Therefore, the originality of the work lies in its focus on youth perceptions and its potential to reshape the design of sociotechnical systems.

2 BACKGROUND

In this section, we synthesize the literature on youth online risks and highlight potential gaps that motivate our work.

2.1 Youth Online Risk Detection and Risk Perceptions

When discussing interaction design in the context of adolescents, online safety has been a central concern [1, 6, 8, 16, 26], as we want teens to benefit from online technologies but also protect them from potential harms. Meanwhile, the massive scale of online content generation has led to the development of ML-based risk detection tools to automatically identify such risks in the context of content moderation, as well as specifically to protect youth from online harm (e.g., [2, 4, 9, 29, 36, 48, 50, 60, 64, 65]). In the existing ML literature, one of the common approaches to annotate ground truth data (e.g., annotating whether certain messages are safe or unsafe) for risks is by leveraging third-person annotations (e.g., crowdsourcing) [10, 23, 44]. With this approach, researchers work with annotators to manually code the data based on defined guidelines or definitions. Third-person annotations can be efficient if the coding task is straightforward (e.g., identifying specific objects in images). However, when the tasks involve subjective perspectives such as labeling risky media, relying solely on third-party annotations may be ecologically invalid (e.g., over-flagged) ground truth data [36].

One line of research examines the gap between the perceived influence of media on self versus others, or the "third-person effect." Proposed by Davison [17], the third-person effect posits that people believe that others are more vulnerable to persuasive media messages than they are, and this perception can influence behavior. Scholars have shown that the significance of the third-person perception can lead individuals' behavioral responses to support protection (i.e., censorship) for others from the perceived harmful influence of the media [24, 42]. Empirical evidence of the third-person effects on negative media issues (e.g., violence, sexism, racism) has been documented by several studies [20, 24, 30, 55]. For instance, individuals tend to overrate the X-rating and support the censorship of pornography to protect others from the harm of such content [24]. It is argued that the third-person perception is an indication of an underlying paternalistic attitude [41] in which individuals see themselves as capable of defending themselves against potentially harmful media effects while seeing others as in need of protection [42]. This overestimation of the effect of negative or harmful media messages on others can lead people to take preventative action toward others [17].

2.2 The Implications of the Third-Person Effect on Youth Online Risk Detection

Within the evolving field of Human-Centered Machine Learning (HCML), scholars have highlighted the importance of grounding human values and needs in ML-based system development [14], to minimize the potential harms for those who are affected by the systems [18, 25, 47]. From a human-centered perspective, collecting

ground truth annotations from those who experience the risk ensures that the training risk detection models reflect real-world experiences and accurately represent the risks users face online [37, 52]. Risk perceptions are highly subjective [34] and the perceptions of risks are driving factors in subsequent behaviors [57]. Therefore, understanding the risk perceptions of people who experienced the risk (i.e., youth in our case) is the foundation of the design of ML-based sociotechnical systems to support them. Our work is the first to involve youth to compare their risk perceptions with third-party annotators in the context of automated risk detection ground truth. For instance, Kim et al. [36] highlighted significant differences in the performance of ML models for the detection of cyberbullying between training datasets based on the perspective of insiders (those who are directly involved in or have experienced cyberbullying) and outsiders (those who observe or analyze cyberbullying). The study highlighted that including insider perspectives is crucial for improving cyberbullying detection algorithms. Yet, they relied on the youths' risk perception based on the post categorization feature (e.g., bullying, self-harm, relationships, addiction, etc.) available on an online peer-support platform, "as a proxy for the victim's perspective," which potentially could lead to a disconnection between the actual risk perceptions of the youth and their categorization of posts on the platform. Also, the work was focused on cyberbullying while we have a holistic approach to various risk types. Therefore, we offer a more holistic view of differences between youth first-person accounts of risks versus third-person, by allowing youth and RAs to flag conversations for various risk types.

3 METHODS

Below, we present an overview of data collection and ethics in data collection processes. Then we describe the data annotation process, participants and dataset characteristics, and qualitative and quantitative analysis processes to address the research questions.

3.1 Study Overview

We conducted a user study of unsafe media messages sent to and from youth in private conversations on Instagram. We chose Instagram due to its high popularity among the youth [38]. We collected over 10,000 Direct Messages (DMs) from Instagram, contributed by 100 youth aged 13-21 who were: 1) English speakers based in the United States, 2) had an active Instagram account currently and for at least 3 months during the time they were teens between the ages of 13 to 17, 3) exchanged DM conversations with at least 15 people, and 4) had at least 2 DMs that made them or someone else feel uncomfortable or unsafe. Participants uploaded their Instagram data after downloading it from Instagram and marked their conversations as either safe or unsafe. The Instagram zip file included their conversational data which includes texts and media files (image, audio, video). Participants also filled out an online survey with demographic information such as age and gender. Then, we recruited six undergraduate research assistants (RAs) to identify unsafe DMs and the risk types in those unsafe messages. The annotation process by the participants and RAs is presented in the following section.

As the dataset included private and sensitive personal information, we took the utmost care to preserve the confidentiality and privacy of the participants. We gave step-by-step instructions to youth on how to remove data prior to uploading it to the system and we gave explicit warnings to avoid uploading any media including the nudity of a minor. Since we asked about potentially triggering sensitive information from participants, we also included the "Help Resources" tab on the website available to participants at all times. When we developed a web-based system to collect youth-donated data, the technical implementation of the system went through an institutional security audit. We ensured that our system passed all security standards and policies of our institution. We followed our data management plan which included only storing data in safe and restricted data storage approved by the university's information technology security audit team (see Razi et al. [51] for details regarding system development and privacy measures). When presenting the results, we paraphrased quotations and recreated privately shared images to ensure confidentiality. All images presented in this paper are publicly available via a general search (i.e., broadly disseminated images) or have been modified to protect the identity of participants. For the same reason, all faces have been blurred. We did not alter publicly available images, such as memes. More details regarding considerations for data ethics are explained in Section 7.

3.2 Data Annotation Process

3.2.1 Youth's First-Person Risk Perceptions. First, each private conversation (a set of DMs) was labeled by youth as either safe or unsafe. If the conversation was labeled as unsafe, participants were then asked to identify the specific DMs that made that conversation unsafe, as well as the type of risk(s). The risk types are derived from the existing literature [67] and the existing Instagram risk reporting categories [33]. The seven categories included:

- Nudity/porn: Photos or videos of a nude or partially nude people or person
- Sexual messages/solicitations: Sending or receiving a sexual message ("sexting") - being asked to send a sexual message, revealing, or naked photo
- Harassment: Messages that contain credible threats, aim to degrade, or shame someone, contain personal information to blackmail or harass someone or threaten to post nude photos of someone
- Hate speech: Messages that encourage violence or attack anyone based on who they are; specific threats of physical harm, theft, or vandalism
- Violence/threat of violence: Messages, photos, or videos of extreme violence, or that encourage violence or attack anyone based on their religious, ethnic, or sexual background
- Sale or promotion of illegal activities: Messages promoting the use or distribution of illegal material such as drugs
- Self-injury: Messages promoting self-injury such as suicidal thoughts, cutting, and/or eating disorders

In addition, participants were asked to provide more context if a conversation was labeled as unsafe such as where they met the sender(s) (e.g., online or offline) and their relationship with the sender(s) (e.g., friends or strangers). We call the above two risk dimensions labeled by youth "risk annotations." Once participants

completed the risk annotation, we manually verified the data provided by the participants and compensated them with a $50 Amazon gift card for their time and efforts.

3.2.2 Third-Party Research Assistants' Risk Annotations. Next, we recruited six undergraduate RAs to identify unsafe media messages, who ranged in age from eighteen to early twenties. We consciously recruited undergraduate RAs as our annotators because working with RAs to annotate ground truth data is a common approach in the existing literature [23, 36, 59] and emerging adults (age 18-21) are the closest peer group to the youth who could be knowledgeable about the context of youth online risks. We recruited RAs by sending emails through our university's academic departments to students who were interested in working on online safety projects. We had an interdisciplinary team of RAs whose majors were Computer Science, Psychology, Criminology, and Sociology. Most of the RAs were voluntary/unpaid and were not incentivized to flag more or less unsafe media messages.

In this work, we did not focus on measuring (dis)agreement among RAs to have them annotate the messages based on their own perceptions of risks as third persons. After completing IRB CITI training and onboarding information sessions, each RA was assigned participants with which to review and annotate all of their private conversations for risks. We developed a web-based tool to facilitate this annotation process. We also had an active Skype group chat with RAs for ongoing conversations about any challenges with the annotation process and for mental health support. We had two RAs code each conversation (a set of direct messages) that participants donated so that each conversation was annotated by the participant and two RAs. All RAs independently labeled the given conversations in terms of 1) whether the conversations are risky or safe, and if risky, 2) risk types. Unlike participants, RAs were not asked to label the relationship to the sender as the third person could not know the exact relationship of the participants with the sender. Therefore, we relied on the relationship type information for RAs' risk labels by matching the conversation IDs from the participants' risk labels. If matches in conversation IDs were found, we used the relationship-type labels that the participants provided for those conversation IDs as a proxy.

3.3 Participants' Demographics and Dataset Characteristics

We collected Instagram data from 100 youth aged between 13-21, with an average age of 16 (*SD* = 2.03 years). The majority of the participants identified themselves as female (68%), with 24% as males, and 8% as non-binary or preferred not to answer. Participants' race distribution was as follows: 41% White, 19% Black/African-American, 16% mixed races or preferred to self-identify, 16% Asian or Pacific Islander, and 8% Hispanic/Latino. Participants were mostly heterosexual or straight (47%), followed by bisexual (28%), preferred not to self-identify (12%), and homosexual (11%). From 100 youth, we collected 11,062 conversations, out of which 1,452 (13.13%) conversations were marked as risky by participants. We filtered the dataset to focus on media messages (e.g., images, audio, videos) that had risk types flagged by the youth and/or RAs to focus on media-sharing behavior in a private conversation context. This filtering process resulted in 674 unsafe media messages from 127 private conversations exchanged by 42 youth, which were identified as unsafe by 18 youth and 6 RAs. From 674 unsafe media messages, 41 unsafe media messages were labeled by the participants, 645 were labeled by the RAs, and 12 media messages were labeled by both the participants and the RAs.

Of the 18 participants who flagged their own media messages as risky, 14 identified themselves as females and 4 as males with an average age of 15.5. No participants identified themselves as non-binary or chose to self-identify. Participants' races included Caucasian/White (7, 39%), Asian or Pacific Islander (5, 28%), Hispanic/Latino (2, 11%), African American/Black (1, 6%), and Mixed races or who preferred not to self-identify (3, 17%). Participants' sexual orientations included in order 12 heterosexual (67%), 4 bisexual (22%), and 2 homosexual (11%). The relationship status of the participant in their teenage years included 12 single (67%), 3 serious relationship (exclusive) (17%), 2 both single and serious relationship (11%), and 1 dating (nonexclusive) (6%). During their teen years, their caregiver(s) was(were) mostly mother and father (17, 94%), and only mother (1, 6%). Participants used Instagram mostly several times a day (12, 67%), several times an hour (3, 17%), every day or almost every day (2, 11%), and once or twice a week (1, 6%).

3.4 Data Analysis Approach

3.4.1 Qualitative Analyses. After youth and RAs annotated media messages for risks, we performed qualitative analyses on the 674 unsafe media messages to determine the risk context. First, we conducted a content analysis [19] to code each unsafe media message by media content type. Through the content analysis, we came up with the five media content types including:

- Meme: Digitally altered/created images usually containing both images and text
- Screenshot: Images of device screens
- Natural image of the person: Images of a person or body part in the natural world
- Natural image of objects: Images of an object or animal in the natural world
- Art Illustration: Drawn or illustrated artworks

Next, we performed a thematic analysis [61] to identify more nuanced characteristics and patterns within risky media. We began this process by revisiting the dataset and noting down some initial codes based on our observations, considering the larger conversation around the shared unsafe media. From there, we began the full coding process for two more rounds to refine the codes. Through this iterative and comparative process, we identified the three codes:

- Humor: Risky images that contained a humorous undertone (non-serious),
- Broadcast: Risky images that were not directed toward any particular individual
- Personal: Risky images that were sent personally (i.e., to target or address the individual).

With a list of the three codes generated, we constructed themes by examining codes and grouping codes into meaningful patterns. Next, we reviewed our themes alongside our dataset to confirm that they actually captured important meanings within the coded data. After reviewing the theme thoroughly, we named the theme

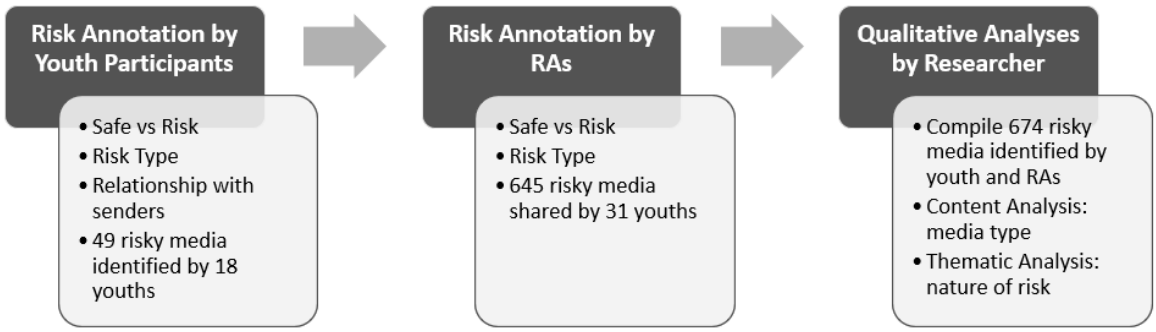


Figure 1: Overview of data annotation and qualitative analyses

the "nature of risk" of the unsafe media message. Note that we differentiated our content analysis from our thematic analysis to be precise regarding our mixed qualitative methods. While thematic analysis considers both latent and manifest content in data analysis, a content analysis can choose between manifest and latent contents before proceeding to the next stage of data analysis [63]. Coding for media type, for instance, was more straightforward and did not require coding for emerging thoughts or ideas. Meanwhile, coding for more nuanced insights (e.g., whether the risk was targeted or humorous) required inferences from the broader context of the conversation.

Using the codes above, we coded the risky media messages in terms of "media content types" and "nature of risk." Each unsafe message was assigned *one* code from the media content type and the nature of risk dimensions. Some of the media messages contained one or more media content types (e.g., image of a person and image of objects in one media message). In this case, we assigned one code that was the most relevant to the context of the unsafe media message. Along with the risk annotations (risk types and relationships with senders) provided by the participants and the RAs, the labels for risk context (media content type and nature of risk) annotated by the researcher were used to examine the differing online risk perceptions of youth vs RAs. Figure 1 shows the overview data annotation and qualitative analyses process.

3.4.2 Statistical Analyses. To compare the difference between youth and RAs based on their risk perceptions and answer RQ1, we conducted Fisher's exact tests [21] between youth-labeled and RA-labeled unsafe messages in terms of risk types, relationship types, media content types, and the nature of risks. The Fishers test is a type of exact test that is usually used to examine the significance of the association (contingency) between the two categorical variables when the expected frequencies are less than 5 [35]. We leveraged the Fishers exact test to assess whether there were significant differences in the proportions of our risk themes and codes when unsafe media messages were flagged by participants versus the third-party annotators. The p values ($\alpha = .05$) were used to demonstrate the significance of the associations between youth-labeled and RA-labeled unsafe messages.

3.4.3 Qualitative Examination of Youth vs RA Risk Labels. Finally, we conducted another round of qualitative analyses of the risky media messages annotated by the participants and the RAs to explore how and why the participants' perspectives and the RAs' perspectives differ (RQ2). Based on Fisher's exact test results, we qualitatively compared the content of the unsafe media messages across different salient risk dimensions (risk types and relationship types) and contextual dimensions (media content type and nature of risks) when there was a noticeable difference between participants' and RAs' risk perceptions. For comparative analysis, we closely examined the larger conversation around the shared unsafe media to understand the broader context of the mismatch between participants' and RAs' risk perceptions. Through the in-depth analysis of risky images and youth conversation around those images, we identified four themes related to how and why risk perceptions differ between youth and RAs. The four themes included: 1) both participants and RAs flagged sexually explicit risks 2) youth did not flag their own risky behavior as unsafe, while RAs did, 3) youth perceived personal attacks disguised as innocuous messages as risky, while RAs flagged humorous risk, and 4) youth perceived risks from strangers, while RAs were more aware of risky interactions between friends.

4 RESULTS

In this section, we present the difference between the participants' and RAs' labels in terms of the risk dimensions (RQ1). Then, we describe the themes of how and why perceptions of online risks varied between youth and RAs (RQ2).

4.1 Characteristics of and Quantitative Differences between Risky Media Labeled by Youth Versus RAs (RQ1)

4.1.1 Risk Types. Out of risk type labels annotated by youth, the most frequently labeled risk type was "harassment" (19, 38%). The second most frequent risk type was "nudity/porn" (12, 24%), followed by "sexual messages" (8, 16%), "hate speech" (4, 8%), "violence" (3, 6%), "sale/promotion of illegal activities" and "self-injury" (2, 4%), respectively. RAs, on the other hand, labeled the majority of unsafe media messages they found as sexual messages/solicitation (353, 47%), followed by nudity/porn (105, 14%), and harassment (96, 13%). We found a statistically significant difference between the participant and the RA labels and the risk types based on Fisher's test ($p < 0.001$) (Table 1).

Risk Type	Participants	RAs	Total
Harassment	19 (38%)	96 (13%)	115
Hate Speech	4 (8%)	42 (6%)	46
Nudity/Porn	12 (24%)	105 (14%)	117
Sale/promotion of illegal activities	2 (4%)	56 (7%)	58
Self-injury	2 (4%)	18 (2%)	20
Sexual messages	8 (16%)	353 (47%)	361
Violence	3 (6%)	84 (11%)	87
Total	50 (100%)	754 (100%)	804

Table 1: The counts and percentages of risk type across participant labels and RA labels. The difference between youth and RA labels was significant ($p < 0.001$).

4.1.2 Relationship Types. Out of unsafe media messages with relationship type annotated by participants, the majority (24, 63%) were from "acquaintance" which consisted mostly of harassment messages. Some (12, 32%) messages were sent from "stranger" and we observed all seven kinds of risk types in those messages. Only 1 (3%) unsafe message was sent from either "friend" or "family." None of the unsafe media was sent from participants' "significant other." On the other hand, most unsafe media messages flagged by the RAs were sent from "friends" of the participants (229, 90%), followed by "stranger" (11, 4%) and "acquaintance" (7, 3%). A Fisher's test identified a statistically significant difference between the participant and RA labels in terms of the relationship types with the sender ($p < 0.001$) (Table 2).

Relationship Type	Participants	RAs	Total
Acquaintance	24 (63%)	7 (3%)	31
Friend	1 (3%)	229 (90%)	230
Stranger	12 (32%)	11 (4%)	23
Significant other	0 (0%)	5 (2%)	5
Family	1 (3%)	3 (1%)	4
Total	38 (100%)	255 (100%)	293

Table 2: The counts and percentages of relationship type across participant labels and RA labels. The difference between youth and RA labels was significant ($p < 0.001$).

4.1.3 Media Content Types. The majority of the unsafe media messages labeled by the participants were "natural image of person" (25, 61%), followed by "meme" (7, 17%) and "screenshot" (6, 15%). There were 2 (5%) "video/audio" messages and 1 (2%) "object natural image" from the participant risk labels. None of the unsafe media labeled by youth was "art illustration." On the other hand, the majority of the unsafe media messages identified by RAs were memes (224, 35%), followed by video/audio (134, 21%), and screenshots (133, 21%). A Fisher's test yielded a significant difference between the participant and RA labels based on the media content type ($p < 0.001$) (Table 3).

Media Content Type	Participants	RAs	Total
Meme	7 (17%)	224 (35%)	231
Screenshot	6 (15%)	133 (21%)	139
Nature image of person	25 (61%)	62 (10%)	87
Video/audio	2 (5%)	134 (21%)	136
Art illustration	0 (0%)	71 (11%)	71
Nature image of object	1 (2%)	21 (3%)	22
Total	41 (100%)	645 (100%)	686

Table 3: The counts and percentages of media content type across participant labels and RA labels. The difference between youth and RA labels was significant ($p < 0.001$).

4.1.4 The Nature of Risks. The majority of the unsafe media messages labeled by participants were targeted at the participants personally (36, 88%), followed by humor (4, 10%). Only one risky message was intended to be broadcast (2%). Meanwhile, many of the unsafe media messages flagged by the RAs were framed as humorous (361, 57%), followed by broadcast (194, 30%) and personal (82, 13%). A Fisher's test resulted in a significant difference between the participant and RA labels based on the nature of the risks ($p < 0.001$) (Table 4).

Nature of Risk	Participants	RAs	Total
Humor	4 (10%)	361 (57%)	365
Broadcast	1 (2%)	194 (30%)	195
Personal	36 (88%)	82 (13%)	118
Total	41 (100%)	637 (100%)	678

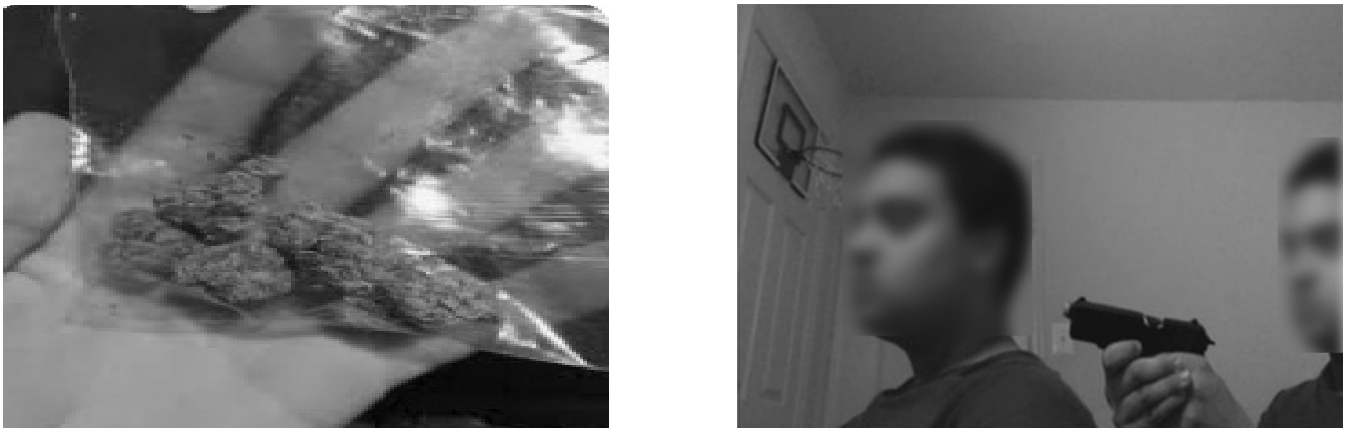
Table 4: The counts and percentages of nature of risk across participant labels and RA labels. The difference between youth and RA labels was significant ($p < 0.001$).

4.2 A Qualitative Examination of Differing Risk Perceptions of Youth Versus RAs (RQ2)

We first present the few instances where youth and RAs flagged the same media messages for risk, then unpack the themes for why they flagged the majority of the media messages differently.

4.2.1 Both Participants and RAs Flagged Sexually Explicit and/or Objective Risks. There were 12 messages labeled by both youth and RAs, most of which were personally targeted sexual risk such as sexually explicit images with natural persons' genitals exposed. Figure 2 show publicly searchable examples of a natural image of nudity/porn risk labeled both by the youth and RAs. In the images, an adult man is making a sexual advance toward a younger girl in a school skirt. The images were sent to the youth with the text: "*Whatever...You could be my spoiled little princess (smiley emoji).*" Also, there was a sexually overt meme consisting of an image of a female and a male having a fellatio and vulgar text derogating the white female which was annotated by both participants and RAs as "nudity/porn." As can be seen from the examples, sexual risks

Figure 2: Risky media messages labeled by both youth and RAs. Youth and RAs both flagged sexually explicit images. Both images were publicly available via a general search.



(a) Promotion of illegal activities labeled exclusively by RAs

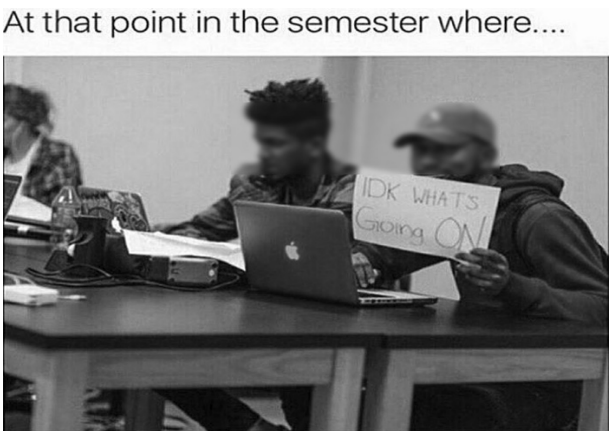
(b) Meme of violence/self-harm labeled exclusively by RAs

Figure 3: Risky images involving youth's own risky behavior identified exclusively by RAs. Youth were conservative about labeling their own risk. Both images were publicly available.

flagged by both RAs and youth were overt and targeted directly toward the youth. There were also harassment messages flagged by both participants and RAs. In one example, the participant's face was used as a stranger's screensaver without consent from the participant. The image was followed with requests for the youth to send nude photos of the participant and be a secret friend: *"Go to the mirror with full nakedness and take good look at yourself."* Overall, the risky media messages flagged by both youth and RAs were more explicit and targeted than those that were flagged by only one of the two parties.

4.2.2 Youth did not flag their own risky behavior as unsafe, while RAs did. In RQ1, one of the least risky media content types youth labeled was video/audio. When reviewing the unsafe videos exclusively identified by RAs, we found that many of them were of "sale or promotion of illegal activity," in which participants themselves performed/promoted risky behaviors. For instance, we noted a set of messages (sent from the participant) that featured videos of the participant smoking marijuana and drinking vodka. There was another set of risky images flagged exclusively by RAs in which the participant was either holding marijuana or weighing it Figure 3a, but none of them were flagged by the participant. We also observed a similar trend among unsafe screenshots of "harassment," most of which were flagged solely by the RAs, not by the participants. For instance, we noticed that some screenshots of harassment were being exchanged between the participants and their friends to disseminate how participants (as perpetrators) harassed someone else. None of these messages were flagged by the participants.

The trend was consistent for risky memes created by the participants or their peers, which were identified exclusively by RAs. For instance, there was a set of media messages in which youth and their acquaintances/friends were exchanging risky memes that contained various types of risks such as hate speech, violence, self-harm, and nudes. Yet, none of them were perceived as unsafe by



(a) Targeted harassment disguised as an innocuous meme (labeled exclusively by participants)

(b) Sexual meme with humorous undertone (labeled exclusively by RAs)

Figure 4: Personal attacks disguised as innocuous messages identified by youth (left) and humorous sexual risk annotated by RAs (right). Youth were conservative about labeling sexual images with a humorous undertone. Both images were publicly available.

youth. Figure 3b is an example of a publicly searchable meme of violence/self-harm sent from the youth with the text "Trust nobody not even yourself," which was identified as unsafe only by RAs.

4.2.3 Youth perceived personal attacks disguised as innocuous messages as risky, while RAs flagged humorous risks. In RQ1, we found that participants were more likely to flag "personally targeted" media messages, while the RAs were more likely to flag "humorous" or "broadcast" media messages as unsafe. Among participants' risk labels, many of the personally targeted media messages were of harassment, which RAs often overlooked. For instance, in one conversation, the sender sent thirteen messages containing photos of themselves and repeatedly asked participants to rate the photos on physical attractiveness with the text: *"Which one you like...Rate me lmaoo."* While the youth asked the sender to stop sending such requests: *"Ok chill with all the texting pls."* In this example, RAs did not perceive the shared images as risky since the selfies themselves can look like innocuous messages, but the participants felt uncomfortable or unsafe (i.e., labeled as harassment) because they were being inundated by unwanted messages. Another example is a meme of a student sitting in a class with the text "At that point in the semester where...IDK what's going on" (Figure 4a). RAs did not flag the image as risky since it contained no explicit content. On the other hand, the participant felt uncomfortable as the meme was targeted at the participant and labeled the messages as harassment.

The discrepancies between the participant labels and the RA labels were also found when considering humorously framed risks. The majority of humorous messages were "sexual messages" with sexual undertones identified exclusively by the RAs (Figure 4b). Many of the humorous sexual messages were of sexual memes or screenshots that were publicly available online. Although the messages contained sexual themes and references (i.e., 'who's gonna give it to him'), when they are framed in humor, youth did not perceive them as unsafe while RAs did. Overall, we found that participants perceived less risk from humorously framed sexual messages. However, they perceived more risk from personally targeted harassment messages even when there were no explicit content risks (e.g., threat or sexual risk) in the messages. RAs, on the other hand, perceived less risk from personally targeted messages if there were no explicit content risks in those messages but noticed more risk from sexual messages with a humorous undertone.

4.2.4 Youth perceived risks from strangers, while RAs were more aware of risky interactions between friends. In RQ1, we found that the participants rarely flagged media messages from their friends, while RAs most frequently flagged them. When reviewing the risky messages that were flagged exclusively by the RAs, we found that nearly all of those messages were sexual memes, art illustrations, screenshots, and nudity/porn exchanged between the participant and their friends. Many of the sexual memes and screenshots youth exchanged with their friends were publicly available via search, while many of the sexual art illustrations and nudity/porn were created by the youth themselves or their friends. Figure 5b shows an example of sexual messages sent from youth' friends, which was perceived risky exclusively by RAs. Although the messages contained sexually explicit visuals and/or texts that were targeted toward the participant, they did not perceive them as risky since the messages were sent from their trusted parties for fun.

On the contrary, youth tended to flag humorously framed sexual content as risky when it is coming from strangers. For instance, one participant flagged a humorously framed video of a sexual message that was sent from a stranger as unsafe (while RAs did not). In this video, a boy is making sexual jokes about a female character. Although the message was humorously framed, the participant perceived it as risky because it was coming from an unknown party. There was also a sexual meme with a humorous tone, but participants perceived it as a sexual risk as it was sent from a stranger (Figure 5a). Taken together, we confirmed that participants

(a) Humorously framed sexual meme sent from a stranger (labeled exclusively by participants)

(b) Screenshot of sexual message exchange between friends (labeled exclusively by RAs)

Figure 5: Sexual message sent from a stranger (annotated by participant, left) and a sexual message sent from participant's friend (annotated by RAs, right). Youth were conservative about labeling sexual images exchanged between friends. The image on the left was publicly searchable, while the image on the right was re-created in the likeness of the original by the researchers.

perceived less risk from media messages sent by individuals they know or trust even when the messages were explicit and targeted toward the participants. Instead, they were more concerned with media messages (even humorous ones) if the messages came from unknown or barely known parties. On the other hand, the RAs were more aware of content risks (e.g., nudity/sexual messages) from the messages exchanged between the youth and their friends for non-risky purposes.

5 DISCUSSION

In this section, we first discuss the implications of our findings and how they relate to the existing literature. Then we discuss the implications of our findings related to designing ML approaches for risk detection, youth education, and online platform design. Finally, we acknowledge the limitations of our work and provide suggestions for future research directions.

5.1 Differences in Risk Perspectives

5.1.1 Third-person Effects in Youth Online Safety Context. Our results confirmed that compared to the youth (first-person), RAs (third-person) annotated significantly larger numbers of media messages as risky. That is, the pattern of RAs' over-flagging unsafe media aligns with the third-person effect documented in the literature [17]. This tendency of RAs' overrating as third-person was salient, especially when annotating sexual memes for risk where RAs flagged many of the sexual jokes between youth and their peers as risky when youth themselves did not. This is a noteworthy trend, particularly given that sexual interaction with peers is not necessarily risky and rather, considered a natural and necessary part of adolescence [4, 52, 67]. RAs' overrating of unsafe media messages could be the reflection of their paternalistic views toward youth (even with a small age gap), assuming youth are more vulnerable to the negative impacts of such potentially unsafe content than they are, hence, youth need protection even from jokes among peers if the content is sexual in nature. This paternalistic view could have been heightened for sexually risky cases due to the perceived sexual vulnerabilities of females [22]. In fact, the majority of our participants were female, and some of the sexual images annotated exclusively by RAs were selfies of females expressing their sexuality for consensual sexting, which does not always lead to harm [52]. Although we did not ask RAs about their perception toward youth's susceptibility to unsafe content in this study, we showed *in what ways* this perception gap manifested. Future research is needed to explore deeper into how third-person perceptions impact the annotation tasks for youth online risks.

5.1.2 Strangers can be risky, but so can friends. We discovered that youth's relationship with the sender has an impact on their perception of online risks. For one, we observed "stranger danger" scenarios where participants felt unsafe with humorous media content sent from strangers. In fact, the issue of online stranger danger is not new. Existing literature established that teens are generally at higher risk of potentially harmful online interactions with strangers [8]. Our findings add empirical evidence to prior literature that youth perceived more risks from strangers and that it is crucial to empower youth with coping strategies (e.g., nudging, education) for stranger danger. At the same time, youth underestimated the risks posed by their friends or trusted parties that RAs highlighted. This raises concerns because prior literature confirmed that the youth had a harder time dealing with certain risks from their friends than strangers [52]. For instance, pressure from friends or romantic partners affects youth sexting decisions while youth have conflicting feelings (e.g., doubt and shame) about engaging in sexting [27]. Hence, we need to investigate ways to empower youth by educating how to set safety boundaries and deal with risks when they occur with people they know.

Another pattern we observed in our study was that the youth did not flag their own risk (in many cases, involving friends) or take into account the negative impacts of risks (e.g., promoting smoking, harassing peers) as risky, while RAs did. This could be explained as the extension of third-person effects beyond the direct relational context between the youth and the RAs. That is, the scope of the third-person effect could be extended beyond the context of the direct relationship between the first-person and third-person (youth and RAs in our case). Prior literature showed that both parental mediation and support for censorship were associated with the parents' perceived negative effects of televised violence not only on their own children but also on other children [28]. In our case, when annotating for risks, RAs may have considered potential risks posed to not only youth participants but also those who interacted with youth (e.g., friends and acquaintances). Future research is warranted to explore the degree to which third-person perceptions can be extended when evaluating online risks for youth.

5.1.3 Are humorous risks still risks? Our findings suggest that the framing of risks (i.e., humorous vs personally targeted) could be a key factor that differentiates perceptions of online media risks between youth and RAs. Prior literature suggested that youth may create and share risky content for fun [12]. This could have been the case for our youth; they may have perceived some of the explicit contents exchanged with their friends to be funny, rather than personally risky. Additionally, youth frequently found explicit content risks to be "irrelevant" or even pursue such risks [4, 52, 67] because they perceive that non-targeted risks need not be "resolved" and hence, can largely be ignored [67]. The participants in our study may have perceived that unsafe messages that were not personally targeted did not warrant labeling the interaction as risky. RAs, on the other hand, may have perceived the same explicit content as risky because their focus was more on the content itself and less on the contextual information around the conversation as a whole (e.g., friends making sexual jokes). Again, this could be due to the paternalistic views of RAs toward youth that youth need to be protected from exposure to explicit content, including humor.

On one hand, it might be beneficial for youth if they can ignore and are not adversely impacted by non-targeted and/or humorously framed explicit content. At the same time, being insensitive to such content may lead youth to be in high-risk situations in the future. In fact, as developmental theory captures, some levels of risk-taking and experiential learning are normal aspects of adolescent developmental growth [11]. For instance, by exchanging sexual jokes, youth may socialize and learn about sexual experiences. However, some of the risky content generated by other youth (e.g., violence) was the second most concerning risk for younger teens [39]. The high priority given to youth-generated risky content is noticeable as it has received less attention than sexual content or bullying. Similarly, being exposed to youth-generated media about underage drinking and smoking marijuana may trigger future online and offline problematic behaviors [43]. Hence, setting a healthy balance between allowing youth to learn from low-risk experiences and protecting them from high-risk situations is necessary. Future research can examine youth-generated risky content with various risk levels and how we could support youth set healthy boundaries between fun and danger.

5.2 Implications for Designing Machine Learning Approaches for Risk Detection

Our work provides important insights into designing automated detection of youth online risks. Developing more accurate risk detection algorithms is the end goal of our work. Therefore, establishing robust ground truth for what is risky is an important prerequisite to designing such algorithms. We observed that annotating online risk is a highly subjective task, and hence, we cannot do this objectively with a level of consistency, even after all the measures we took to ensure consistency. Particularly considering the third-person effect, if we continue to rely only on over-rated third-party annotations for ground-truth annotations, risk detection systems will have high false positive cases. Our findings ground the need for methods to collect risk-flagged data including those who experience the risks, rather than relying on the third person alone to make it more ecologically valid [3]. On one hand, our stance is first-person who encountered the risk should have the strongest voice when historically, this has not been the case. On the other hand, we acknowledge that youth are still forming their sense of risk awareness and there are reasons why we see the differences from both sides. Therefore, we suggest assessing online risks based on multiple perspectives and respectfully designing technologies to provide safer experiences for youth. Additionally, the key differences between youth and RA risk perceptions could serve as foundations for engineering and fine-tuning features in machine learning models to detect youth online risk. These human-centered approaches to designing youth online risk detection will be more translatable in the real world and benefit youth for their "rich ecological validity" [36].

Our findings also have implications beyond youth online risk detection. The differences in risk perceptions uncovered in this study could impact the design of automated support systems. Now, ML-based automated systems pervade our society, ranging from medicine and public health [13, 46] to criminal justice and child welfare [56]. Given the pervasiveness and the scale of impacts of decisions made by such systems, there has been a shift to applying a human-centered lens to computational approaches [37, 53]. Among many, fairness and bias in ML-based systems are critical topics that must be addressed by the SIGCHI community. Our work adds valuable insights into this shift by highlighting the importance of the voices of those who are often replaced by proxies. We call for more efforts toward reflecting the real-world experiences of key stakeholders in the design of ML-based sociotechnical systems. In addition, we acknowledge the ethics and challenges of trauma-informed research such as our own work. Collecting risk labels from vulnerable populations is challenging as it requires researchers to make additional efforts to ensure that participating in research does not put vulnerable youth in more harm. In our study, we went through a rigorous process to ensure the safety of our research participants, from building a secure system to collect risk labels to providing mental health resources for youth and annotators. Due to such efforts, we had no adverse experiences reported by both youth and annotators. Yet, our research team conducted a follow-up interview study (forthcoming publication) to share our experiences and discuss the ethics of trauma-informed research in-depth, but the study was outside the scope of the current study.

Additionally, since personal information could be easily traceable even in aggregated data, we need extra care for the privacy of the youth. We note that dealing with sensitive data entails various challenges that researchers should carefully address. Privacy protection of participants and ethical usage of data should be the utmost priority, which should be extended to the speculated usage of the applications when deployed in real-life scenarios. Future research should address ethical and privacy-preserving ways to work with sensitive datasets generated by the most vulnerable youth. Taken together, our work is a step forward to work with ecologically valid datasets so that translational research has a real-world impact on youth online safety.

5.3 Implications for Practice and Design

5.3.1 Practical Implications for Youth Education. Our research provides practical implications for youth online safety education. We revealed that youth were often insensitive to their own risky content, such as animated pornography or videos of underage drinking. However, it highlighted the importance of recognizing the potential harm of such content, as it could lead to problematic behaviors online and offline. Hence, we need to play a critical role in educating youth by increasing awareness of risks that they cannot see themselves. Risk assessment from the third-person perception (such as RA risk labels in our study) would provide educators and practitioners with valuable insights into risky media that is often overlooked by youth. At the same time, we acknowledge that not all explicit content (e.g., sexual jokes between friends) is necessarily unsafe for youth. Given that teens can potentially benefit from being exposed to low online risks (e.g., develop interpersonal skills such as boundary setting and empathy [67]), we do not want to take away these opportunities from youth by over-flagging their online interaction as risky, which could increase the chances of false positives and hence, lead us to miss the most concerning risks that need to be addressed. Thus, our role in practice should not be focused on flagging every explicit content, but rather on helping youth become aware of and more sensitized to risks that they may not perceive on their own.

5.3.2 Design Implications. One of the key findings of our study was that unsafe media messages identified by youth were mostly from strangers or acquaintances. To mitigate the issue of "stranger danger," social media platforms may apply a message filter in which youth are informed about the potential risks of viewing messages from strangers and that they can choose whether or not to view the message at all. Automatically blocking private messages from unknown adults could be an aggressive yet proactive solution to protect youth from being exposed to unsafe content. Given that Instagram already implemented a policy in which adult users are not allowed to privately message teens under 18 who do not follow those adult users [31], this might be an easy intervention for social media platforms to actively moderate online stranger danger. Furthermore, we acknowledge that not all media content shared between youth and their peers is safe. Social media platforms could play a critical role in alerting youth-generated risky content (e.g., screenshots of youth harassing others, videos of illegal activities of youth), although youth consider such content to be non-risky or even fun. Using the third-person risk labels, social media platforms can design and implement a nudging system to alert youth about the potential risks in youth-generated risky content. Receiving nudges may increase risk awareness of otherwise desensitized youth and may mitigate the opportunities for them to consume youth-generated risky content.

Finally, as not all explicit content is necessarily unsafe to youth, social media platforms can consider contextual factors when designing filters to alert or block risky media content. One way to do so is to add more interactive features to their safety features. For instance, social media platforms can add feedback features to the filtering system (e.g., Sensitivity Filter [32]) so that youth can provide interactive feedback (e.g., reporting false alerts) to the system. They can also consider adding customization features to allow youth users to tailor filtering/alert systems to work best for them. This way, social media platforms can reflect unique perspectives of youth to design interventions to promote youth online safety.

5.4 Limitations and Future Research

We acknowledge a few limitations of our work. First, due to the qualitative approaches used in this study, we focused on analyzing 674 private media messages exchanged on Instagram; thus, we note that our findings cannot be generalizable. Another potential limitation would be sampling bias. Participants of our study must have registered as active users on Instagram for a certain time and signed up to donate their data for research purposes. Thus, we recognize that the results from this study may not be the same for other youth populations. Future research could endeavor to explore differences in perspectives with a more diverse pool of youth and annotators with conversation data collected from other social media platforms. In addition, as risk is highly subjective, we expected to have a certain level of disagreement among RAs for our risk annotation tasks. However, disagreement among human annotations is not necessarily considered noise because there could be a plausible range of human judgments for subjective tasks (such as ours), rather than a single ground truth [45, 49]. Future research can explore different types of potential biases in online risk data annotation in depth and ways to mitigate such biases. Furthermore, our dataset may not reflect the entire unsafe media messages privately exchanged on Instagram. For legal reasons, we asked our participants to remove any instances of child pornography from their data, hence, we were not able to include such high-risk and/or illegal media. Finally, we recognize the ethics of research involving vulnerable populations, which continues to be an important open issue within the scholarly communities. While understanding the first-person perspective is valuable, studying online risks with youth can unintentionally put an "already vulnerable population at greater risk" [7]. Reviewing and flagging risky media could have made youth feel uncomfortable. Having said that, the issue of youth online risks is a critical one, and research such as ours is necessary for designing youth-centered online safety interventions.

6 CONCLUSION

Our findings challenge the prevailing reliance on third-party ground truth annotation to design youth online risk detection systems. We examined the key dimensions of how and why perceptions of online

risks varied between youth and RAs. We found that RAs annotated a significantly larger number of unsafe media messages than youth did. This is because contextual factors such as the way risks are framed (humorous vs. personally targeted), and the sources of the risks together (themselves/friends vs. strangers) could differentiate online risk perceptions of youth and research assistants. A key takeaway from our study is that risk is highly subjective, especially when it comes to the private online interactions of youth, and that understanding the perspective of those who are experiencing risks is vital. Our work provides grounds for annotating online risks by incorporating youth perspectives and respectfully designing sociotechnical systems to provide safer experiences for youth.

7 SELECTION AND PARTICIPATION OF CHILDREN

We collected Instagram data from youth between the ages of 13 and 21. The participants were recruited via the website of the authors' research lab. To participate in this study, each youth was required to have an active Instagram account currently and for at least 3 months by the time they were between the ages of 13 and 17. The participants were also required to have had at least 15 direct message (DM) conversations, two of which must have made them or someone else feel uncomfortable. We obtained approval from the Institutional Review Boards (IRBs) of the authors' institutions including informed consent from eligible participants over the age of 18; for those under 18, we obtained informed consent from their parents followed by their informed assent before they participated in the study. Since we asked about potentially triggering sensitive information from participants, we included the "Help Resources" tab on our website available to participants at all times.

We have procedures in place for our duty of being mandated child abuse reporters and our responsibility of reporting child pornography (i.e., any nudity of a minor under the age of 18) to authorities, which we clearly stated in the consent and assent forms. We gave step-by-step instructions to youth on how to remove data before uploading it to our system to avoid sharing any media including the nudity of a minor. All participants' data were de-identified for the analysis and stored on a secure server. Additionally, we acquired the National Institute of Health (NIH) Certificate of Confidentiality to preserve the privacy of our participants and prevent the subpoena of the data during legal discovery. The above information on data sharing was communicated during the consent process. All researchers conducting data collection or analyzing the data completed the CITI human subjects research training and the initiation protection of minors training program.

ACKNOWLEDGMENTS

This research was supported by the U.S. National Science Foundation under grants IIP-2329976, IIS-2333207, and the William T. Grant Foundation grant 187941. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our sponsors.

REFERENCES

[1] Zainab Agha. 2023. To Nudge or Not to Nudge: Co-Designing and Evaluating the Effectiveness of Adolescent Online Safety Nudges. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 760–763.

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

[3] Ashwaq Alsoubai, Xavier V Caddle, Ryan Doherty, Alexandra Taylor Koehler, Estefania Sanchez, Munmun De Choudhury, and Pamela J Wisniewski. 2022. MOSafely, Is that Sus? A Youth-Centric Online Risk Assessment Dashboard. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*. 197–200.

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

[5] Monica Anderson, Emily A Vogels, Andrew Perrin, and Lee Raine. 2022. Connection, Creativity and Drama: Teen Life on Social Media in 2022. https://www.pewresearch.org/internet/2022/11/16/connection-creativity-and-drama-teen-life-on-social-media-in-2022/

[6] Karla Badillo-Urquiola, Scott Harpin, and Pamela Wisniewski. 2017. Abandoned but not forgotten: Providing access while protecting foster youth from online risks. In *Proceedings of the 2017 Conference on Interaction Design and Children*. 17–26.

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

[8] Karla Badillo-Urquiola, Diva Smriti, Brenna McNally, Evan Golub, Elizabeth Bonsignore, and Pamela J Wisniewski. 2019. Stranger danger! social media app features co-designed with children to keep them safe online. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. 394–406.

[9] Vimala Balakrishnan, Shahzaib Khan, and Hamid R Arabnia. 2020. Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers & Security* 90 (2020), 101710.

[10] Natã M Barbosa and Monchu Chen. 2019. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[11] Diana Baumrind. 1987. A developmental perspective on adolescent risk taking in contemporary America. *New directions for child and adolescent development* 1987, 37 (1987), 93–125.

[12] Danah Boyd, Alice Marwick, Parry Aftab, and Maeve Koeltl. 2009. The conundrum of visibility: Youth safety and the Internet. (2009).

[13] Francisco Maria Calisto, Nuno Nunes, and Jacinto C Nascimento. 2022. Modeling adoption of intelligent agents in medical imaging. *International Journal of Human-Computer Studies* 168 (2022), 102922.

[14] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 2019. Who is the" human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–32.

[15] Jody Clay-Warner. 2003. The context of sexual violence: Situational predictors of self-protective actions. *Violence and victims* 18, 5 (2003), 543–556.

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

[17] W Phillips Davison. 1983. The third-person effect in communication. *Public opinion quarterly* 47, 1 (1983), 1–15.

[18] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[19] James W Drisko and Tina Maschi. 2016. *Content analysis*. Pocket Guide to Social Work Re.

[20] Julie M Duck and Barbara-Ann Mullin. 1995. The perceived impact of the mass media: Reconsidering the third person effect. *European Journal of Social Psychology* 25, 1 (1995), 77–93.

[21] Ronald Aylmer Fisher. 1970. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*. Springer, 66–70.

[22] Jacklyn Friedman. 2017. Stop Telling Women We Need to Be Saved From Our Sexuality. https://time.com/5015027/stop-policing-womens-sexuality/

[23] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 325–336.

[24] Albert C Gunther. 1995. Overrating the X-rating: The third-person perception and support for censorship of pornography. *Journal of Communication* 45, 1 (1995), 27–38.

[25] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020*

conference on fairness, accountability, and transparency. 501–512.
[26] Heidi Hartikainen, Netta Iivari, and Marianne Kinnula. 2019. Children's design recommendations for online safety education. *International Journal of Child-Computer Interaction* 22 (2019), 100146.
[27] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. 'If You Care About Me, You'll Send Me a Pic'-Examining the Role of Peer Pressure in Adolescent Sexting. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 67–71.
[28] Cynthia Hoffner and Martha Buchanan. 2002. Parents' responses to television violence: The third-person perception, parental mediation, and support for censorship. *Media Psychology* 4, 3 (2002), 231–252.
[29] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics*. Springer, 49–66.
[30] John M Innes and Howard Zeitz. 1988. The public's view of the impact of the mass media: A test of the 'third person'effect. *European Journal of Social Psychology* 18, 5 (1988), 457–463.
[31] Instagram. 2021. Continuing to Make Instagram Safer for the Youngest Members of Our Community. https://about.instagram.com/blog/announcements/continuing-to-make-instagram-safer-for-the-youngest-members-of-our-community
[32] Instagram. 2021. Introducing Sensitive Content Control. https://about.instagram.com/blog/announcements/introducing-sensitive-content-control
[33] Instagram. 2022. Abuse and Spam. https://help.instagram.com/165828726894770
[34] Sibylle Juvalta, Camilla Speranza, Dominik Robin, Yassmeen El Maohub, Julia Krasselt, Philipp Dreesen, Julia Dratva, and L Suzanne Suggs. 2023. Young people's media use and adherence to preventive measures in the "infodemic": Is it masked by political ideology? *Social Science & Medicine* 317 (2023), 115596.
[35] Hae-Young Kim. 2017. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative dentistry & endodontics* 42, 2 (2017), 152–155.
[36] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela Wisniewski, and Munmun De Choudhury. 2021. You Don't Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 290–302.
[37] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J Wisniewski, and Munmun De Choudhury. 2021. A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–34.
[38] Susan Laborde. 2023. Teenage social media usage statistics in 2023. https://techreport.com/statistics/teenage-use-of-social-media-statistics/
[39] Sonia Livingstone, Lucyna Kirwil, Cristina Ponte, and Elisabeth Staksrud. 2014. In their own words: What bothers children online? *European Journal of Communication* 29, 3 (2014), 271–288.
[40] Megan K Maas, Kyla M Cary, Elizabeth M Clancy, Bianca Klettke, Heather L McCauley, and Jeff R Temple. 2021. Slutpage use among US college students: the secret and social platforms of image-based sexual abuse. *Archives of sexual behavior* 50, 5 (2021), 2203–2214.
[41] Douglas M McLeod, Benjamin H Detenber, and William P Eveland Jr. 2001. Behind the third-person effect: Differentiating perceptual processes for self and other. *Journal of Communication* 51, 4 (2001), 678–695.
[42] Douglas M McLeod, William P Eveland Jr, and Amy I Nathanson. 1997. Support for censorship of violent and misogynic rap lyrics: An analysis of the third-person effect. *Communication Research* 24, 2 (1997), 153–174.
[43] Megan A Moreno and Jennifer M Whitehill. 2014. Influence of social media on alcohol use in adolescents and young adults. *Alcohol research: current reviews* 36, 1 (2014), 91.
[44] Michael Muller, Christine T Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, et al. 2021. Designing ground truth and the social life of labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
[45] Jennimaria Palomaki, Olivia Rhinehart, and Michael Tseng. 2018. A Case for a Range of Acceptable Annotations.. In *SAD/CrowdBias@ HCOMP*. 19–31.
[46] Jinkyung Park, Ramanathan Arunachalam, Vincent Silenzio, Vivek K Singh, et al. 2022. Fairness in mobile phone–based mental health assessment algorithms: Exploratory study. *JMIR formative research* 6, 6 (2022), e34366.
[47] Jinkyung Park, Rahul Dev Ellezhuthil, Joseph Isaac, Christoph Mergerson, Lauren Feldman, and Vivek Singh. 2023. Misinformation Detection Algorithms and Fairness across Political Ideologies: The Impact of Article Level Labeling. In *Proceedings of the 15th ACM Web Science Conference 2023*. 107–116.
[48] Jinkyung Park, Joshua Gracie, Ashwaq Alsoubai, Gianluca Stringhini, Vivek Singh, and Pamela Wisniewski. 2023. Towards Automated Detection of Risky Images Shared by Youth on Social Media. In *Companion Proceedings of the ACM Web Conference 2023*. 1348–1357.
[49] Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics* 7 (2019), 677–694.
[50] Afsaneh Razi, Ashwaq AlSoubai, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J Wisniewski. 2023. Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–29.
[51] Afsaneh Razi, Ashwaq AlSoubai, Seunghyun Kim, Nurun Naher, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J Wisniewski. 2022. Instagram Data Donation: A Case Study on Collecting Ecologically Valid Social Media Data for the Purpose of Adolescent Online Risk Detection. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–9.
[52] Afsaneh Razi, Karla Badillo-Urquiola, and Pamela J Wisniewski. 2020. Let's Talk about Sext: How Adolescents Seek Support and Advice about Their Online Sexual Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[53] 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 (2021), 1–38.
[54] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6 (2017), 1–12.
[55] Michael B Salwen and Michel Dupagne. 1999. The third-person effect: Perceptions of the media's influence and immoral consequences. *Communication Research* 26, 5 (1999), 523–549.
[56] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. A human-centered review of algorithms used within the US child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[57] Jingyuan Shi and Hye Kyung Kim. 2019. Integrating risk perception attitude framework and the theory of planned behavior to predict mental health promotion behaviors among young adults. *Health communication* (2019).
[58] Thiago H Silva, Pedro OS Vaz De Melo, Jussara M Almeida, Juliana Salles, and Antonio AF Loureiro. 2013. A picture of Instagram is worth more than a thousand words: Workload characterization and application. In *2013 IEEE International Conference on Distributed Computing in Sensor Systems*. IEEE, 123–132.
[59] Vivek K Singh, Souvick Ghosh, and Christin Jose. 2017. Toward multimodal cyberbullying detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2090–2099.
[60] Devin Soni and Vivek K Singh. 2018. See no evil, hear no evil: Audio-visual-textual cyberbullying detection. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–26.
[61] Gareth Terry, Nikki Hayfield, Victoria Clarke, and Virginia Braun. 2017. Thematic analysis. *The SAGE handbook of qualitative research in psychology* 2 (2017), 17–37.
[62] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 247–267.
[63] Mojtaba Vaismoradi, Hannele Turunen, and Terese Bondas. 2013. Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing & health sciences* 15, 3 (2013), 398–405.
[64] Nishant Vishwamitra, Hongxin Hu, Feng Luo, and Long Cheng. 2021. Towards understanding and detecting cyberbullying in real-world images. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*.
[65] Paulo Vitorino, Sandra Avila, Mauricio Perez, and Anderson Rocha. 2018. Leveraging deep neural networks to fight child pornography in the age of social media. *Journal of Visual Communication and Image Representation* 50 (2018), 303–313.
[66] Kate Walker and Emma Sleath. 2017. A systematic review of the current knowledge regarding revenge pornography and non-consensual sharing of sexually explicit media. *Aggression and violent behavior* 36 (2017), 9–24.
[67] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, Daniel F Perkins, and John M Carroll. 2016. Dear diary: Teens reflect on their weekly online risk experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3919–3930.