It Takes a Village: Youth Online Safety Research Highlights Need for Interdisciplinary, Multistakeholder Solutions

Elizabeth A. Sweigart [1,*], Jinkyung Katie Park [2], and Pamela J. Wisniewski [1]

[1] School of Engineering, Vanderbilt University, Nashville, USA
[2] School of Computing, Clemson University, Clemson, USA
Email: elizabeth.sweigart@vanderbilt.edu (E.A.S.); jinkyup@clemson.edu (J.K.P.); pamela.wisniewski@vanderbilt.edu (P.J.W.)
*Corresponding author

***Abstract*—Youth online safety is a critical concern for parents, educators, behavioral health specialists, technologists, and policymakers. Although the problem is relevant to this array of diverse stakeholders, past research has been limited primarily to parents and teens. Recent trends in Human-Computer Interaction (HCI) research emphasize the need for a holistic approach that bridges the gaps between these myriad constituencies. These interdisciplinary efforts aim to develop evidence-informed tools and policies that go beyond outdated paradigms and simple access restrictions. This article synthesizes the extant literature—from the evolving understanding of the complex and nuanced risks youth face online to the creation of automated risk detection solutions using Artificial Intelligence (AI)—to elucidate this emerging paradigm. This work contributes to the existing body of knowledge in the field by elaborating the urgent need for comprehensive research and collaborative strategies to ensure young people can safely navigate and benefit from the digital world.**

***Keywords*—adolescent online safety, online risk detection, Artificial Intelligence (AI), open innovation, Human-Computer Interaction (HCI)**

I. Introduction

Youth online safety is a top-of-mind topic for many worldwide ranging from parents, teens, and educators to behavioral health specialists, technologists, and policymakers. In the popular press, journalists have highlighted anecdotal evidence of social media harms especially affecting vulnerable groups like young women and girls [1] and the difficulties faced by policymakers and regulators in addressing them [2]. At the same time, academic studies have documented the complex decisions faced by families and educators in selecting from the range of software solutions aimed at combating the problem while still allowing youth to benefit from the learning opportunities the internet provides [3, 4]. Behavioral health practitioners are keenly aware of the dangers the internet can pose but remain wary of interventions that may inhibit youth from seeking mental health support [5].

Adding to the difficulty of finding a solution, until recently most research on this subject has lacked a holistic approach. The literature has focused on a limited number of geographic areas, mainly the United States, with over 80% of the data self-reported by teens [6]. Less than 2% of the analyzed data in peer-reviewed studies has come from educators [7]. More than 85% of these studies provide only a snapshot in time without longitudinal data [6]. Moreover, researchers have often overlooked the intersection of various risks, typically focusing on a single risk from a narrow set, such as cyberbullying or exposure to pornography. Nearly all existing software products targeting this issue were developed without input from youth and only rarely involved educators and child psychologists [8]. There is a clear and urgent need for more research and public policy initiatives in this area [9], as well as for evidence-based tools that help young people learn to navigate the increasingly online world safely.

Over the last few years, a new trend has emerged in the Human-Computer Interaction (HCI) field that aims to produce actionable, evidence-informed insights by bridging the gap between primary (i.e., teens and parents) and often-overlooked secondary (i.e., technologists, behavioral health providers, educators, and policymakers) stakeholders [10]. Ultimately, these integrated initiatives should create better solutions, products, and policies that help keep youth safer online while allowing them to take advantage of the opportunities the internet offers. These recent contributions to the literature underscore the value of an interdisciplinary, multistakeholder approach to addressing youth online safety that eschews dated notions of so-called 'stranger danger' and goes beyond merely blocking and restricting access to internet applications that youth can easily circumvent [11].

In this article, we synthesize the research performed as part of the Modus Operandi Safely (MOSafely) project [12] and by others supporting this paradigm shift away from single-stakeholder approaches to youth online safety toward a more inclusive and diverse

Manuscript received July 27, 2024; revised September 23, 2024; accepted November 11, 2024; published January 23, 2025.

doi: 10.12720/jait.16.1.121-129

multistakeholder perspective. Specifically, this article addresses:

- Evolving understanding of the risks youth face online and the shift away from a single-dimensional fear of online predators to broader concerns about self-image, privacy, and peer pressure that necessitates fresh approaches to online risk detection.
- Advances in the use of Artificial Intelligence (AI) and Machine Learning (ML) to develop sophisticated new automated tools and algorithms for online risk detection and the practical and ethical dimensions of their implementation.
- Conceptualization of an open-source consortium of primary and secondary stakeholders in the youth online safety arena that leverages an open innovation approach with shared governance.

Through this synthesis, we identify and describe the additional avenues of inquiry this paradigm shift demands. Advances in online risk detection algorithms necessitate an increased focus on the ethical use of AI regarding youth, particularly concerning how their data may be used (or not) for training and improving these models. Correspondingly, values-informed approaches, such as the UK's Age-Appropriate Design Code and Value Sensitive Design [13], should be integrated into a multistakeholder approach to youth online safety, providing essential frameworks for shaping policy initiatives that prioritize the well-being and rights of young users. We conclude with a discussion of how this work will ultimately require a thoughtful integration of open innovation principles underpinned by shared governance to ensure that all stakeholder voices are heard and diverse perspectives, informed by cultural differences, are respected.

II. Understanding the Risks Youth Face Online

Today's teens are deeply immersed in the digital world, with the internet and especially social media platforms playing a role in their daily lives. According to a study by Pew in 2023, platforms like YouTube, TikTok, Snapchat, and Instagram are the most widely used among U.S. teens, with YouTube leading at a 93% usage rate among youth aged 13 to 17 [14]. Despite concerns about the negative impacts of social media on youth, the Pew survey noted that nearly half of all teens report being online almost constantly, a figure that has roughly doubled since 2014–2015. This near-constant connectivity underscores the integral role digital devices and the internet play in the social and educational lives of adolescents. Demographic factors—ranging from socioeconomic status to gender, race, and ethnicity—also influence the patterns of internet and social media use among teens with youth from lower socioeconomic backgrounds reporting more frequent internet usage than their more affluent counterparts. Both the popular press and the academic literature well document how the high rates of social media use among teens and their almost continuous internet connectivity have profound implications for their social interactions, self-image, and mental health [15]. The extensive use of platforms for both positive interactions and exposure to online risks calls for a nuanced approach to adolescent online safety that balances the benefits of digital engagement with the need to protect teens.

A. Changing Landscape of Risks Faced by Youth Online

The literature reveals a significant evolution over the last decade in the understanding of the risks youth face online. Initially, the primary concern centered around the fear of online predators, so-called "stranger danger" [16]. However, this focus has broadened to encompass more sophisticated issues such as self-image, privacy, and peer pressure, and their relationship to the already widely recognized online dangers youth encounter such as sexual threats, cyberbullying, and exposure to inappropriate content [17–20].

1) Intersection of sexual threats and self-image risks

With more of their lives taking place in the digital realm, increasingly teens look to the internet as a place to explore their natural curiosity about sex and discuss sexual development and exploration with their peers [21]. As such, sexual threats are a particular area of concern for adolescents online. Although empirical evidence suggests that greater online access to sexually explicit materials and interactions is not driving youth to have sex earlier or more often, teens' online sexual interactions still expose them to numerous other threats [22]. These interactions range from consensual sexting with peers to non-consensual solicitations and abuse [20] and can have severe consequences, including emotional distress and reputational damage [23].

Youth from especially vulnerable groups may face a compounding of these issues. For example, many LGBTQIA+ teens lack trusted sources for education about sex as compared to their heterosexual peers [24], leading them to rely on internet pornography and other sexually explicit materials they can access for free online [25]. These teens may copy harmful behaviors they observe or try sexual acts that appear pleasurable on screen but may hurt them in real life [26, 27]. Even when these young people have supportive families, their parents may not have sufficient context or knowledge to educate their LGBTQIA+ teens when it comes to same-sex intimacy [28]. Encouragingly for those working to mitigate these risks, the research shows that even if youth are learning dangerous or harmful information from online pornography (e.g., not using condoms), their subsequent behavior is mitigated by having parents who talk to them honestly about sex [29] and by accurate, nonjudgmental educational content shared on social media [30].

Recent studies have drawn out the intersection of self-image and sexual risks encountered by teens online, which can have profound effects on their mental health and well-being [31]. The pressures of maintaining a certain image online may lead to risky behaviors by teens such as sexting and the sharing of personal photos to fit in or gain approval from peers [32]. This behavior, regardless of whether it begins as consensual, can quickly turn risky when images are shared without consent or used for extortive purposes [33]. The burden teens feel to conform to peer expectations and present an attractive image online can push teens into compromising situations, further impacting

their self-esteem and sense of security [23]. Social media platforms like Instagram, TikTok, and Snapchat are especially influential in shaping teens' perceptions of themselves, especially for young women [34]. These platforms often feature idealized images and videos that can lead to unrealistic expectations and comparisons. According to Ali et al., the media shared in unsafe private conversations often convey negative emotions, which can exacerbate feelings of inadequacy and low self-esteem among youth [21]. This constant exposure to curated and often unattainable images can distort teens' self-perception and lead to body image issues and anxiety. These findings showcase the interconnectedness of the types of risks youth encounter online reinforcing the need for scholars and practitioners to break free from the previously siloed thinking noted by Pinter *et al.* [6].

2) Interaction between cyberbullying, self-image, and privacy risks

Another significant risk that directly impacts self-image is cyberbullying which involves deliberate, repeated, and hostile behavior using digital technology to harm others, often manifesting through threatening messages, spreading rumors, or sharing private information without consent [35]. Often, cyberbullying incidents revolve around personal appearance and social status, which are critical aspects of adolescent self-image [36]. Victims of cyberbullying may experience severe emotional distress, leading to depression, anxiety, and even suicidal thoughts [37] across globally diverse populations of teens [38]. Additionally, the damaging comments and public shaming inherent in cyberbullying leave lasting scars on a young person's self-worth and identity [39]. The intersection and simultaneous occurrence of these risks make them particularly concerning for social service providers working with teens [40].

Cyberbullying not only targets the recipient's emotional well-being but also their sense of privacy, as sensitive information can be widely disseminated, leading to significant psychological distress [38]. Privacy violations are a significant aspect of cyberbullying, as perpetrators may share or threaten to share personal photos and information without consent [41]. As young people navigate the path from childhood and adulthood, privacy can become a tension between youth and their adult caregivers making the mitigation of these risks challenging from a sociotechnical perspective [42, 43]. Although the field of youth online safety has focused its attention primarily on the recipients of cyberbullying, emerging research shows that adolescents who engage in cyberbullying may be driven by their own depression or poor self-image [44].

3) Involvement of peer pressure and exposure to inappropriate online material

Exposure to inappropriate material is another key risk for youth online, especially multimedia-based platforms like YouTube and TikTok, despite their popularity, can expose youth to content that is not age-appropriate or harmful [45]. Church *et al.* [46] described interdisciplinary studies showing that the algorithms driving these platforms often prioritize profit-enhancing engagement over safety, leading to the dissemination of risky content to youth. When it comes to explicit content, strangers can be risky, but so can peers.

Youth report experiencing peer pressure to share sexually explicit content, opening them up to other online risks [20]. Hartikainen *et al*. [47] showed that youth are more likely to succumb to peer pressure to send sexual material online when the request comes from someone they are romantically interested in. Peer pressure may also be present as a protective factor in instances where teens provide support and encouragement for their peers to disengage from potentially harmful online sexual interactions [48]. Relatedly, Park *et al*. [49] showed that although the messages contained sexually explicit visuals or texts or both that were targeted toward the participant, they did not perceive them as risky since the messages were sent from their friends for fun. The authors argued that it might be beneficial for youth if they can ignore and are not adversely impacted by humorously framed explicit content exchanged between friends. At the same time, being insensitive to such content may lead youth to be in high-risk situations in the future. As such, existing literature indicates the need for interventions to empower youth to help them set safety boundaries and deal with risks that occur with peers.

B. Implications for Online Risk Detection

Shifts in understanding the online risks faced by youth—from a narrow focus on online predators to broader concerns about self-image, privacy, and peer pressure—call for innovative approaches to online risk detection. Research indicates that the interconnectedness and diversity of these risks necessitate the development of new tools for online risk detection, as traditional methods may fall short in addressing the complex digital dangers teens encounter [50]. Considering how peer pressure affects risky behaviors, how privacy breaches amplify emotional distress, and how self-image issues influence online activities can enhance the effectiveness of future risk detection technologies. Both AI- and ML-based technologies show the potential to detect and mitigate teens' exposure to inappropriate material [51, 52]. These technologies, when designed with a focus on context and user feedback, can play a crucial role in creating safer online environments for adolescents. Simultaneously, this underscores the need for nuanced detection systems that can differentiate between consensual and non-consensual interactions and provide appropriate interventions [32].

Paradoxically, increased awareness of the myriad risks youth face in their online activities has simultaneously given rise to a more nuanced view of online interactions. Rather than seeing them simply as potential risks, there is an emerging view that they can serve as opportunities for positive learning and resilience-building [53]. This perspective encourages the design of online safety mechanisms that are developmentally appropriate and empower adolescents rather than merely restrict their activities [11]. Cultural awareness and intentional inclusion of diverse communities are also critical to the success of future risk detection and intervention efforts in youth online safety. Ongoing research into and

development of online risk detection tools should consider the needs of particularly vulnerable and at-risk youth [54, 55] and look to incorporate the perspectives of youth who have experienced cyberbullying and other online harms [36]. The internet and social media are central to the lives of modern teens, shaping their communication, entertainment, and even their identity formation. The continuous evolution of these digital environments necessitates innovative research and adaptive strategies to support the healthy development of youth in an increasingly connected world.

III. Advances in Online Automated Risk Detection

As the literature bears out, current approaches to youth online safety often rely on parental controls and monitoring which can overwhelm parents with excessive information and invade teens' privacy [3]. Largely, these solutions are ineffective because existing risk detection algorithms generally fail to consider the context of online interactions, providing confusing feedback to users when innocuous situations are flagged as crises [32]. These challenges with current approaches reveal the need for more sophisticated, context-aware AI systems that can discern the nature of different interactions and tailor responses accordingly. The emergent research considers how advanced AI techniques, such as Natural Language Processing (NLP) and computer vision, can more accurately identify and respond to risky behavior while better protecting privacy [23, 51, 52]. Moreover, integrating human feedback into these systems ensures that AI interpretations align with real-world experiences and ethical considerations [56–58]. Such improvements are critical for creating a safer online environment that respects the autonomy and developmental needs of teens while effectively managing potential risks. In this section, we examine recent advances in the field related to ethical data collection, development of cutting-edge risk detection models built on youth-labeled ground-truth data, the real-world implications of these new systems with teens, and evolving approaches to balancing privacy and safety.

A. Ethical Data Collection

Collecting and curating the large amounts of trustworthy data required to train effective AI models (i.e., ground-truth data) is a mammoth task that requires careful management of significant ethical, legal, and regulatory considerations [59]. The collection of social media data from teens for AI risk detection systems presents numerous challenges, particularly when the data includes sensitive or illegal content like sexually explicit images. To address these issues responsibly, researchers may implement leading practices [60]:

- *Involve teens in the permission process*. In addition to obtaining parental consent, researchers may request teen assent to empower teens with greater control over their data.
- *Allow teens to label data proactively*. To avoid retraumatizing teens, researchers may allow them to self-identify experiences of harassment without time pressure.
- *Anonymize data from the outset*. Special care must be taken to protect privacy in the digital realm. Beyond standard practices of not publishing Personally Identifiable Information (PII), researchers should ensure that messages or postings are not quoted directly if they can be searched online. Usernames should be replaced with anonymous, randomly generated IDs at the point of data collection.

These steps ensure that the development and application of AI tools respect the privacy and psychological well-being of teen participants.

B. State-of-the-Art Risk Detection Models Using Ground-Truth Data Labeled by Youth

Beyond the ethics of data collection, recent research has also examined both the different types of data available for training AI (e.g., social media trace data, metadata) and ways of enhancing the labeling of the data (e.g., incorporating first-person perspectives) to help improve automated risk detection algorithms [36, 48, 61]. Working from the front end, these human-centered design approaches to data collection for algorithm training purposes not only contribute to teens' sense of agency in the process but also serve to make the datasets more ecologically valid [62]. Alsoubai *et al*. [32] used an ML model trained on a manually labeled dataset of 45,955 posts from teens on a peer support platform to accurately identify examples across the spectrum of sexual interactions, from consensual sexting to sexually abusive situations encountered by youth online with a high degree of accuracy, recording an average Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) of 0.90. This work uncovered key patterns related to teens' online sexting behaviors. For example, existing systems tend to treat any mention of homosexuality as explicit leading to a conflation of identity and behavior. Often, this results in a system flagging interactions which may instead be positive instances of peers encouraging or supporting one. The findings of this research can sophisticate future risk detection algorithms.

Ali *et al.* [21] were able to separate regular images from screenshots in a sample of private Instagram messages labeled by the youth participants themselves, allowing the researchers to better understand and characterize the images shared. Notably, the study found that media in unsafe conversations typically featured images of people and expressed negative emotions, whereas media in safe conversations conveyed positive emotions and more frequently included objects. Conversation length, an example of trace data, proved a consistent indicator of unsafe conversations, as they were typically shorter in duration, indicating that young users tend to withdraw from interactions that they perceive as unsafe. Applying this approach to the images themselves, Park *et al*. [51] concluded that vision transformers are effective at identifying complex image characteristics (e.g., personally targeted versus humor) for the purposes of automated risk detection and classification.

C. Evaluation of Real-World Implications with Teens

Human-centered design plays an important role in developing AI systems for monitoring and protecting vulnerable populations, like teens, while safeguarding their desires for privacy and confidentiality [55]. Exemplifying this approach, researchers developed the MOSafely dashboard which incorporates feedback from teen users directly into the AI learning process, allowing the system to continuously improve and adapt based on real-world use by youth [63].

Further evidencing the value of a human-centered design approach, research has shown that focusing on real-world user needs helps to identify and fill gaps in current technologies [56]. Integrating human insights into AI development can facilitate model interpretability and promote fairness, helping to mitigate biases that could impact automated detection outcomes. Likewise, drawing on theoretical frameworks from the social sciences can improve AI models by refining data annotation and feature selection based on understanding behavioral patterns associated with online risks [64]. Grounding this technology work in the ethical principles of fairness, accountability, and transparency is important for establishing and maintaining public trust in the solutions and ensuring the systems are both effective and equitable. Building on this foundation, innovative applications like Virtual Reality (VR) and Augmented Reality (AR) avatar-based therapy sessions have demonstrated how immersive, human-centered technology can support adolescents directly [65]. Leveraging AI-driven avatars to guide teens in recognizing and discussing online issues, these VR sessions offer a simulated 3D space where youth may feel more comfortable sharing personal experiences, such as cyberbullying [66]. By aligning AI and VR with ethical principles such as fairness and accountability, these approaches can both bolster public trust and provide adolescents with meaningful, practical support for navigating real-world challenges [67].

D. Balancing Privacy and Safety

As social media companies, like Meta, move to adopt end-to-end encryption for private communications across applications popular with youth, like Instagram, the data currently relied on for risk detection will be unavailable, raising concerns among youth online safety advocates [68]. Taking up the challenge to find new data points to use as indicators of potential online risk for youth, Ali *et al*. [52] tested various features including metadata, linguistic cues, and image analysis to identify risky conversations. Findings indicated that metadata, especially engagement duration, effectively predicted risks. Contrastingly, linguistic and media cues better differentiated the types of risks. These insights suggest design strategies for AI-based risk detection systems that can operate effectively under encryption constraints, contributing to more nuanced and youth-centric approaches in the field of adolescent online safety. Additionally, the shift towards using metadata respects user privacy by relying on less invasive data, aligning with broader ethical standards in digital communication.

AI-based automated risk detection systems have the potential both to keep youth safer in their online interactions and to be misused by malicious actors. For instance, such systems could be exploited to aggregate illicit content for pornographic websites or the darknet [69]. Additionally, the data these systems identify as risky could be targeted by hackers. To prevent the potential misuse or reverse engineering of algorithms designed to detect online risks for adolescents, researchers recommend taking steps like not making the algorithms open source and carefully monitoring the system to ensure only authorized personnel access it [60].

IV. Toward a Multistakeholder Solution

The research into the evolving understanding of the risks faced by youth online and the ongoing innovation in automated online risk detection systems accentuates the need for collaborative efforts across various fields, including HCI, social science, policy, regulatory, and others, to develop comprehensive and effectual solutions. Historically, both researchers and technology companies have centered their work on parents and teens as the primary stakeholders in youth online safety [4]. Technology companies have looked to address youth online safety by focusing on parental control software aimed at blocking, restricting, and reporting teens' online activities directly to parents [70, 71]. Although effective at mitigating certain risks, the approaches create unintended consequences that may harm youth, especially those in vulnerable environments like foster care [55]. Newer research has started to look beyond parents and teens, as the primary stakeholders, and consider the benefits and challenges of incorporating secondary stakeholders (e.g., teen behavioral health specialists, educators, policymakers) into the process of designing and implementing youth online safety solutions [12].

Despite some recently documented successes with teen-led data donation for research, academics often must rely on the largesse of social media companies for access to their information [62]. This tenuous relationship between researchers and industry players could be partially mediated by clear and cogent regulations from the government and other policy-making bodies related to privacy and data security [72]. Collaboration between these stakeholders might also serve to build trust with parents, teens, and the public who may be wary of the intentions of purely commercial enterprises when it comes to the collection and use of youths' data. Amidst these dynamics, the role of Social Service Providers (SSPs) becomes increasingly significant. SSPs, who are directly engaged with the welfare of youth, can act as crucial intermediaries between the digital technologies designed by researchers and the real-world applications that affect young users [73].

By collaborating with policymakers and researchers, SSPs can ensure that the development and implementation of online risk detection tools are not only effective but also align with the ethical standards and trust required to protect vulnerable youth populations [74]. This collaboration can enhance the practicality and acceptance of such tools,

ensuring they serve the intended protective purposes without compromising the trust and safety of the youth they aim to safeguard. At the same time, as the use of AI-powered online risk detection systems grows, researchers and technologists should look to incorporate insights and feedback from SSPs, many of whom have expressed reservations about these automated systems [40]. SSPs concerns stem from the potential lack of resources and qualified personnel needed to effectively respond to alerts generated by these systems, as well as fears that relying on AI could undermine the trust they've built with youth.

Although there appears to be consensus in the literature about the need for such an open and accessible community, current initiatives are largely siloed by stakeholder group [12]. Led by a team of scholar-practitioners, the Modus Operandi Safely (MOSafely) project [63] is an example of a new initiative that seeks to convene a collaborative network of youth online safety primary and secondary stakeholders. This consortium intends to include a diverse group of constituents including academics, industry experts, SSPs, and government officials, among others. Proposed as an open-source project, participants in the consortium will make their contribution of code and methodologies publicly available with the belief that this approach will accelerate innovation while also promoting transparency, which is vital for building trust among users, policymakers, and other stakeholders. This commitment to transparency reflects lessons learned from the literature about providing ways to verify the equity and effectiveness of the developed tools.

As the project progresses, it will require ongoing integration of a wide range of voices and respect for diverse perspectives, which are essential given the cultural differences that can influence perceptions of online safety and privacy [75]. Thoughtful approaches to governance, particularly shared governance, will also help to manage the expectations and contributions of different stakeholders, ensuring that the project remains aligned with its core goals. Highlighting the imperative of interdisciplinary cooperation in this process, collaborative decision-making models from organizational leadership psychology [76] will likely play a role in facilitating group functioning over the long term. In this way, multistakeholder solutions like the MOSafely project can foster innovation and trust while also creating a sustainable and inclusive framework that adapts to the evolving digital landscape and the needs of its diverse stakeholders and delivers on its promise to help keep youth safer online.

Fig. 1 presents a five-step conceptual framework for establishing a multistakeholder collaborative network dedicated to enhancing youth online safety through the integration of AI, software, AR, and VR technologies. We propose establishing a multidisciplinary steering committee, bringing together representatives from academia, industry, social services, government, education, parent and youth groups, and experts in AI, software, and AR/VR. This committee would crystallize the consortium's mission, objectives, and governance structure, with a focus on positioning these technologies as core components of the initiative. As a next step, the committee would oversee the creation of an open-source platform to facilitate collaborative development, providing a shared space where stakeholders can contribute code, AI models, AR/VR applications, methodologies, and resources, thus fostering innovation and progress across disciplines. With a collaborative workspace established, the next phase would engage stakeholders to form specialized working groups focused on areas like AI and software development, AR/VR applications, policy advocacy, education, and ethical considerations, with specific tasks and milestones to foster structured and measurable outcomes. The steering committee would also convene a task force to implement ethical guidelines to promote data privacy, consent, and responsible use of these technologies. To ensure transparency, all contributions and decisions would be publicly accessible and validated through peer review. Finally, we envision the consortium piloting AI, software, and AR/VR tools in real-world settings, refining solutions based on user feedback, and scaling its efforts globally with support from international stakeholders. Collaborative efforts with government officials and other regulators would help align policy development with our innovations.

Fig. 1. Five-step implementation plan for a collaborative network enhancing youth online safety through AI, software, AR, and VR technologies.

V. Conclusion

Youth online safety is a critical issue that affects a broad range of stakeholders, including parents, educators, behavioral health specialists, technologists, and policymakers. Prior research has largely concentrated on the perspectives of parents and teens, but recent advancements in HCI underscore the necessity of a more holistic approach that connects these groups and adapts to the continuously evolving digital landscape. This comprehensive approach requires interdisciplinary collaboration to create evidence-based tools and policies that transcend traditional paradigms and basic content restrictions. The MOSafely project and similar initiatives highlight the potential of emerging technologies—such as VR and AR—to enhance online safety tools. By integrating AI-driven VR and AR applications, we can create immersive, user-centered environments that not only educate but also allow youth to practice safe digital behaviors in realistic, simulated contexts. Building on the five-step conceptual framework proposed here, this article

synthesizes existing literature on the complexities of online risks faced by youth and discusses the development of innovative automated risk detection solutions using AI and ML. These findings reinforce the pressing need for sustained, collaborative research and an interdisciplinary approach, ensuring youth can explore and benefit from the digital world while being supported by robust, adaptable safety measures.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

EAS, JKP, and PJW jointly performed the analysis. EAS wrote the paper. JKP and PJW revised portions of the paper. All authors approved the final version.

FUNDING

This research is partially supported by the U.S. National Science Foundation under grant #IIP-2329976. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsors.

REFERENCES

[1] J. Valentino-DeVries and M. H. Keller, "A marketplace of girl influencers managed by moms and stalked by men," in *The New York Times*, New York, NY, USA, 2024.

[2] C. Lima-Strong, "Child safety battle looms in congress as tech CEOs testify," in *The Washington Post*, Washington, DC, USA, 2024.

[3] E. A. Sweigart, "Beyond blocking: Factors influencing parents' buying decisions for kids' online safety software," presented at the the 4th Annual BEST Conference on Human Behaviour & Decision Making, Brisbane, QLD, Australia, 2023.

[4] A. K. Ghosh, C. Hughes, and P. J. Wisniewski, "Circle of trust: A new approach to mobile online safety for families," in *Proc. the 2020 CHI Conference on Human Factors in Computing Systems (CHI'20)*, Honolulu, HI, USA, 2020, pp. 1–14. doi: 10.1145/3313831.3376747

[5] S. D. Kauer, C. Mangan, and L. Sanci, "Do online mental health services improve help-seeking for young people? A systematic review," *Journal of Medical Internet Research,* vol. 16, no. 3, 2014. doi: 10.2196/jmir.3103

[6] A. Pinter, P. J. Wisniewski, H. Xu, M. B. Rosson, and J. M. Carroll, "Adolescent online safety: Moving beyond formative evaluations to designing solutions for the future," in *Proc. the Conference on Interaction Design and Children (IDC'17)*, 2017, pp. 352–357. doi: 10.1145/3078072.3079722

[7] M. Sharples, R. Graber, C. Harrison, and K. Logan, "E-safety and web 2.0 for children aged 11–16," *Journal of Computer Assisted Learning,* vol. 25, no. 1, pp. 70–84, 2009. doi: 10.1111/j.1365-2729.2008.00304.x

[8] A. K. Ghosh, "Taking a more balanced approach to adolescent mobile safety," in *Proc. the 19th International Conference on Supporting Group Work (GROUP'16)*, Sanibel Island, FL, USA, 2016, pp. 495–498. doi: 10.1145/2957276.2997025

[9] T. Ammari, A. Kumar, C. Lampe, and S. Schoenebeck, "Managing children's online identities: How parents decide what to disclose about their children online," in *Proc. the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul, Republic of Korea, 2015, pp. 1895–1904. doi: 10.1145/2702123.2702325

[10] X. Caddle, J. Park, and P. J. Wisniewski, "A stakeholders' analysis of the sociotechnical approaches for protecting youth online," in *Proc. the 2024 Future of Information and Communication Conference (FICC)*, Berlin, Germany, 2024, vol. 3, pp. 587–616.

[11] E. A. Sweigart, "Pause, reflect, and redirect: A new approach to helping adolescents make better decisions online," presented at the 4th Annual Behavioural Economics, Society, and Technology Conference, Brisbane, QLD, Australia, February 10, 2022.

[12] X. V. Caddle *et al.*, "MOSafely: Building an open-source HCAI community to make the Internet a safer place for youth," in *Proc. the Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, Virtual Event, USA, 2021, pp. 315–318. doi: 10.1145/3462204.3481731

[13] K. Badillo-Urquiola, C. Chouhan, S. Chancellor, M. De Choudhary, and P. J. Wisniewski, "Beyond parental control: Designing adolescent online safety apps using value sensitive design," *Journal of Adolescent Research,* vol. 35, no. 1, pp. 147–175, 2020. doi: 10.1177/0743558419884692

[14] M. Anderson, M. Faverio, and J. Gottfried. (December 2023). Teens, social media and technology 2023. Pew Research Center. [Online]. Available: https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/

[15] K. Davis *et al.*, "Supporting teens' intentional social media use through interaction design: An exploratory proof-of-concept study," in *Proc. the 22nd Annual ACM Interaction Design and Children Conference*, Evanston, IL, USA, 2023, pp. 322–334. doi: 10.1145/3585088.3589387

[16] R. Sinclair, K. Duval, and M. Ste-Marie, "Reframing stranger danger," in *Working with Trauma-Exposed Children and Adolescents*, J. Pozzulo and C. Bennell, Eds. Routledge, 2018, pp. 149–182.

[17] K. Badillo-Urquiola, A. Razi, J. Edwards, and P. Wisniewski, "Children's perspectives on human sex trafficking prevention education," in *Proc. the 2020 ACM International Conference on Supporting Group Work*, Sanibel Island, FL, USA, 2020, pp. 123–126. doi: 10.1145/3323994.3369889

[18] A. C. Baldry, D. P. Farrington, and A. Sorrentino, "School bullying and cyberbullying among boys and girls: Roles and overlap," *Journal of Aggression, Maltreatment & Trauma,* vol. 26, no. 9, pp. 937–951, 2017. doi: 10.1080/10926771.2017.1330793

[19] A. C. Baldry, A. Sorrentino, and D. P. Farrington, "Cyberbullying and cybervictimization versus parental supervision, monitoring and control of adolescents' online activities," *Children and Youth Services Review,* vol. 96, pp. 302–307, 2019. doi: 10.1016/j.childyouth.2018.11.058

[20] A. Razi, K. Badillo-Urquiola, and P. J. Wisniewski, "Let's talk about sext: How adolescents seek support and advice about their online sexual experiences," in *Proc. the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13. doi: 10.1145/3313831.3376400

[21] S. Ali *et al.*, "Understanding the digital lives of youth: Analyzing media shared within safe versus unsafe private conversations on Instagram," in *Proc. the 2022 CHI Conference on Human Factors in Computing Systems*, New Orleans, LA, USA, 2022, 148. doi: 10.1145/3491102.3501969

[22] J. C. Abma and G. M. Martinez, "Sexual activity and contraceptive use among teenagers in the United States, 2011–2015," *National Health Statistics Report,* no. 104, pp. 1–23, 2017.

[23] A. Razi *et al.*, "Sliding into my DMs: Detecting uncomfortable or unsafe sexual risk experiences within Instagram direct messages grounded in the perspective of youth," in *Proc. the ACM on Human-Computer Interaction*, Hamburg, Germany, 2023, vol. 7, no. CSCW1, pp. 1–29. doi: 10.1145/3579522

[24] T. Tanni, M. Akter, J. Anderson, M. Amon, and P. Wisniewski, "Examining the unique online risk experiences and mental health outcomes of LGBTQ+ versus heterosexual youth," in *Proc. the CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA, 2024. doi: 10.1145/3613904.3642509

[25] S. Pappas. (March 1, 2021). Teaching porn literacy. *Monitor on Psychology*. [Online]. pp. 54–60. Available: https://www.apa.org/monitor/2021/03/teaching-porn-literacy

[26] R. Arrington-Sanders, G. W. Harper, A. Morgan, A. Ogunbajo, M. Trent, and J. D. Fortenberry, "The role of sexually explicit material in the sexual development of same-sex-attracted black adolescent males," *Archives of Sexual Behavior,* vol. 44, no. 3, pp. 597–608, 2015. doi: 10.1007/s10508-014-0416-x

[27] E. F. Rothman, C. Kaczmarsky, N. Burke, E. Jansen, and A. Baughman, " 'Without porn … I wouldn't know half the things I know now': A qualitative study of pornography use among a

sample of urban, low-income, Black and Hispanic youth," *The Journal of Sex Research,* vol. 52, no. 7, pp. 736–746, 2015. doi: 10.1080/00224499.2014.960908

[28] M. E. Newcomb, B. A. Feinstein, M. Matson, K. Macapagal, and B. Mustanski, " 'I have no idea what's going on out there': Parents' perspectives on promoting sexual health in lesbian, gay, bisexual, and transgender adolescents," *Sexuality Research & Social Policy,* vol. 15, no. 2, pp. 111–122, 2018. doi: 10.1007/s13178-018-0326-0

[29] P. J. Wright, D. Herbenick, and B. Paul, "Adolescent condom use, parent-adolescent sexual health communication, and pornography: Findings from a U.S. probability sample," *Health Communication,* vol. 35, no. 13, pp. 1576–1582, 2020. doi: 10.1080/10410236.2019.1652392

[30] R. Stevens, S. Gilliard-Matthews, J. Dunaev, A. Todhunter-Reid, B. Brawner, and J. Stewart, "Social media use and sexual risk reduction behavior among minority youth: Seeking safe sex information," *Nursing Research,* vol. 66, no. 5, pp. 368–377, 2017. doi: 10.1097/NNR.0000000000000237

[31] C. Longobardi, M. A. Fabris, L. E. Prino, and M. Settanni, "The role of body image concerns in online sexual victimization among female adolescents: The mediating effect of risky online behaviors," *Journal of Child & Adolescent Trauma,* vol. 14, no. 1, pp. 51–60, 2021. doi: 10.1007/s40653-020-00301-5

[32] A. Alsoubai, J. Song, A. Razi, N. Naher, M. De Choudhury, and P. J. Wisniewski, "From 'friends with benefits' to 'sextortion': A nuanced investigation of adolescents' online sexual risk experiences," in *Proc. the ACM on Human-Computer Interaction*, 2022, vol. 6, no. CSCW2, pp. 1–32. doi: 10.1145/3555136

[33] A. M. Gassó, B. Klettke, J. R. Agustina, and I. Montiel, "Sexting, mental health, and victimization among adolescents: A literature review," *International Journal of Environmental Research and Public Health,* vol. 16, no. 13, 2019. doi: 10.3390/ijerph16132364

[34] V. Shukla, "Social media's impact on teen age girls' self-esteem and body image," *International Journal of Science and Research,* vol. 12, no. 6, pp. 1258–1262, 2023. doi: 10.21275/SR23530140732

[35] Z. Ashktorab and J. Vitak, "Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers," in *Proc. the 2016 CHI Conference on Human Factors in Computing Systems*, San Jose, CA, USA, 2016.

[36] S. Kim, A. Razi, G. Stringhini, P. J. Wisniewski, and M. De Choudhury, "You don't know how I feel: Insider-outsider perspective gaps in cyberbullying risk detection," in *Proc. the International AAAI Conference on Web and Social Media*, Atlanta, GA, USA, May 2021, vol. 15, no. 1, pp. 290–302.

[37] C. A. Rose and B. M. Tynes, "Longitudinal associations between cybervictimization and mental health among U.S. adolescents," *Journal of Adolescent Health,* vol. 57, no. 3, pp. 305–312, 2015. doi: 10.1016/j.jadohealth.2015.05.002.

[38] G. Gohal *et al.*, "Prevalence and related risks of cyberbullying and its effects on adolescent," *BMC Psychiatry,* vol. 23, no. 1, 39, 2023. doi: 10.1186/s12888-023-04542-0

[39] H. Cowie, "Cyberbullying and its impact on young people's emotional health and well-being," *The Psychiatrist,* vol. 37, no. 5, pp. 167–170, 2013, doi: 10.1192/pb.bp.112.040840

[40] X. V. Caddle, N. Naher, Z. P. Miller, K. Badillo-Urquiola, and P. J. Wisniewski, "Duty to respond: The challenges social service providers face when charged with keeping youth safe online," in *Proc. the ACM on Human-Computer Interaction,* vol. 7, 6, 2022. doi: 10.1145/3567556

[41] M. Ojeda, R. Del Rey, and S. C. Hunter, "Longitudinal relationships between sexting and involvement in both bullying and cyberbullying," *Journal of Adolescence,* vol. 77, no. 1, pp. 81–89, 2019. doi: 10.1016/j.adolescence.2019.10.003

[42] M. Akter, Z. Agha, A. Alsoubai, N. Ali, and P. Wisniewski, "Towards collaborative family-centered design for online safety, privacy, and security," arXiv preprint, arxiv.2404.03165, 2024.

[43] M. Garmendia, G. Martínez, and C. Garitaonandia, "Sharenting, parental mediation and privacy among Spanish children," *European Journal of Communication,* vol. 37, no. 2, pp. 145–160, 2022. doi: 10.1177/02673231211012146

[44] S. Balta, E. Emirtekin, K. Kircaburun, and M. D. Griffiths, "The mediating role of depression in the relationship between body image dissatisfaction and cyberbullying perpetration," *International Journal of Mental Health and Addiction,* vol. 18, no. 6, pp. 1482–1492, 2020. doi: 10.1007/s11469-019-00151-9

[45] A. C. H. Fung and K. K. Y. Wong, "Tick-tock: Now is the time for regulating social media for child protection," *BMJ Paediatrics Open,* vol. 7, no. 1, 2023. doi: 10.1136/bmjpo-2023-002093

[46] K. Church, A. Schoene, J. E. Ortega, R. Chandrasekar, and V. Kordoni, "Emerging trends: Unfair, biased, addictive, dangerous, deadly, and insanely profitable," *Natural Language Engineering,* vol. 29, no. 2, pp. 483–508, 2023. doi: 10.1017/S1351324922000481

[47] H. Hartikainen, A. Razi, and P. J. Wisniewski, "'If you care about me, you'll send me a pic'—Examining the role of peer pressure in adolescent sexting," in *Proc. Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, Virtual Event, USA, 2021, pp. 67–71. doi: 10.1145/3462204.3481739

[48] H. Hartikainen, A. Razi, and P. Wisniewski, "Safe sexting: The advice and support adolescents receive from peers regarding online sexual risks," in *Proc. the ACM on Human-Computer Interaction*, 2021, vol. 5, no. CSCW1, pp. 1–31. doi: 10.1145/3449116

[49] J. Park, J. Gracie, A. Alsoubai, A. Razi, and P. J. Wisniewski, "Personally targeted risk vs. humor: How online risk perceptions of youth vs. third-party annotators differ based on privately shared media on Instagram," in *Proc. the 23rd Annual ACM Interaction Design and Children Conference*, Delft, Netherlands, 2024, pp. 1–13. doi: 10.1145/3628516.3655799

[50] P. J. Wisniewski, H. Xu, J. M. Carroll, and M. B. Rosson, "Grand challenges of researching adolescent online safety: A family systems approach," in *Proc. the Nineteenth Americas Conference on Information Systems*, Chicago, IL, USA, August 15–17, 2013.

[51] J. Park, J. Gracie, A. Alsoubai, G. Stringhini, V. Singh, and P. Wisniewski, "Towards automated detection of risky images shared by youth on social media," in *Companion Proc. the ACM Web Conference 2023 (WWW '23 Companion)*, Austin, TX, USA, 2023, pp. 1348–1357. doi: 10.1145/3543873.3587607

[52] S. Ali *et al.*, "Getting meta: A multimodal approach for detecting unsafe conversations within Instagram direct messages of youth," in *Proc. the ACM on Human-Computer Interaction*, 2023, vol. 7, no. CSCW1, pp. 1–30. doi: 10.1145/3579608

[53] K. Badillo-Urquiola, Z. Agha, M. Akter, and P. Wisniewski, "Towards assets-based approaches for adolescent online safety," in *Proc. the Workshop on From Needs to Strengths: Operationalizing an Assets-Based Design of Technology at the 2020 ACM Conference on Computer Supported Work (CSCW 2020)*, 2020.

[54] Z. Agha, N. Chatlani, A. Razi, and P. Wisniewski, "Towards conducting responsible research with teens and parents regarding online risks," in *Proc. the Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA, 2020, pp. 1–8. doi: 10.1145/3334480.3383073

[55] Z. Agha, K. Badillo-Urquiola, N. Chatlani, A. Alsoubai, and P. Wisniewski, "Socially responsible computing in adolescent online safety," *Tech Otherwise,* 2020. doi: 10.21428/93b2c832.88754350

[56] A. Razi *et al.*, "A human-centered systematic literature review of the computational approaches for online sexual risk detection," in *Proc. the ACM on Human-Computer Interaction*, 2021, vol. 5, no. CSCW2, pp. 1–38. doi: 10.1145/3479609

[57] S. Kim, A. Razi, G. Stringhini, P. J. Wisniewski, and M. de Choudhury, "A human-centered systematic literature review of cyberbullying detection algorithms," in *Proc. the ACM on Human-Computer Interaction*, 2021, vol. 5, no. CSCW2, pp. 1–34. doi: 10.1145/3476066

[58] A. Alsoubai, J. Park, S. Qadir, G. Stringhini, A. Razi, and P. J. Wisniewski, "Systemization of Knowledge (SoK): Creating a research agenda for human-centered real-time risk detection on social media platforms," in *Proc. the CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA, 2024, pp. 1–21. doi: 10.1145/3613904.3642315

[59] W. Liang *et al.*, "Advances, challenges and opportunities in creating data for trustworthy AI," *Nature Machine Intelligence,* vol. 4, no. 8, pp. 669–677, 2022. doi: 10.1038/s42256-022-00516-1

[60] A. Razi, S. Kim, M. de Choudhury, and P. Wisniewski, "Ethical considerations for adolescent online risk detection AI systems," in *Proc. Workshop on Good Systems: Ethical AI for CSCW at the 2019 ACM Conference on Computer Supported Cooperative Work (CSCW 2019)*, Austin, TX, USA, 2019.

[61] J.-E. Kim, E. C. Weinstein, and R. L. Selman, "Romantic relationship advice from anonymous online helpers: The peer support adolescents exchange," *Youth & Society,* vol. 49, no. 3, pp. 369–392, 2017. doi: 10.1177/0044118x15604849

[62] A. Razi *et al.*, "Instagram data donation: A case study on collecting ecologically valid social media data for the purpose of adolescent online risk detection," in *Proc. Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, New Orleans, LA, USA, 2022. doi: 10.1145/3491101.3503569

[63] A. Alsoubai *et al.*, "MOSafely, is that sus? A youth-centric online risk assessment dashboard," in *Proc. the Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, Virtual Event, Taiwan, 2022, pp. 197–200. doi: 10.1145/3500868.3559710

[64] L. N. Olson, J. L. Daggs, B. L. Ellevold, and T. K. K. Rogers, "Entrapping the innocent: Toward a theory of child sexual predators' luring communication," *Communication Theory,* vol. 17, no. 3, pp. 231–251, 2007. doi: 10.1111/j.1468-2885.2007.00294.x

[65] E. Bogdanski, "The effects of virtual reality telemedicine with pediatric patients diagnosed with posttraumatic stress disorder: Exploratory research method case report," *JMIR Formative Research,* vol. 7, 2023. doi: 10.2196/34346

[66] K. M. Ingram, D. L. Espelage, G. J. Merrin, A. Valido, J. Heinhorst, and M. Joyce, "Evaluation of a virtual reality enhanced bullying prevention curriculum pilot trial," *Journal of Adolescence,* vol. 71, no. 1, pp. 72–83, 2019. doi: 10.1016/j.adolescence.2018.12.006

[67] J. Cecil, M. Sweet-Darter, and A. Gupta, "Design and assessment of virtual learning environments to support STEM learning for autistic students," in *Proc. 2020 IEEE Frontiers in Education Conference (FIE)*, October 21–24, 2020, pp. 1–9. doi: 10.1109/FIE44824.2020.9274031

[68] R. Hart. (December 7, 2023). Meta launches end-to-end encryption for messages on Facebook and Messenger. *Forbes*. [Online]. Available: https://www.forbes.com/sites/roberthart/2023/12/07/meta-launches-end-to-end-encryption-for-messages-on-facebook-and-messenger/

[69] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?" presented at the ASIACCS, Taipei, Taiwan, 2006.

[70] P. J. Wisniewski, "The privacy paradox of adolescent online safety: A matter of risk prevention or risk resilience?" *IEEE Security & Privacy,* vol. 16, no. 2, pp. 86–90, 2018. doi: 10.1109/MSP.2018.1870874

[71] P. J. Wisniewski, A. K. Ghosh, H. Xu, M. B. Rosson, and J. M. Carroll, "Parental control vs. teen self-regulation: Is there a middle ground for mobile online safety?" in *Proc. the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Portland, OR, USA, 2017, pp. 51–69. doi: 10.1145/2998181.2998352

[72] X. V. Caddle *et al.*, "A case for partnering with social media platforms to protect adolescents online," in *Proc. Workshop on Social Media as a Design and Research Site in HCI: Mapping Out Opportunities and Envisioning Future Uses at the 2021 ACM Conference on Human Factors in Computing Systems (CHI 2021)*, Virtual due to COVID-19, 2021. doi: 10.1145/3491101.35035

[73] J. Robinson, S. Hetrick, G. Cox, S. Bendall, A. Yung, and J. Pirkis, "The safety and acceptability of delivering an online intervention to secondary students at risk of suicide: Findings from a pilot study," *Early Intervention in Psychiatry,* vol. 9, no. 6, pp. 498–506, 2015. doi: 10.1111/eip.12136

[74] K. Badillo-Urquiola, Z. Agha, D. Abaquita, S. B. Harpin, and P. J. Wisniewski, "Towards a social ecological approach to supporting caseworkers in promoting the online safety of youth in foster care," in *Proc. the ACM on Human-Computer Interaction*, 2024, vol. 8, pp. 1–28. doi: 10.1145/3637412

[75] P. Kumar *et al.*, "Co-designing online privacy-related games and stories with children," in *Proc. the 17th ACM Conference on Interaction Design and Children*, Trondheim, Norway, 2018, pp. 67–79. doi: 10.1145/3202185.3202735

[76] M. Painter-Morland, "Systemic leadership and the emergence of ethical responsiveness," *Journal of Business Ethics,* vol. 82, no. 2, pp. 509–524, 2008. doi: 10.1007/s10551-008-9900-3