USENIX

THE ADVANCED COMPUTING
SYSTEMS ASSOCIATION

# Youth-Centered GAI Risks (YAIR): A Taxonomy of Generative AI Risks from Empirical Data

Yaman Yu, Yiren Liu, Jacky Zhang, Yun Huang and
Yang Wang, *University of Illinois Urbana-Champaign*

https://www.usenix.org/conference/soups2025/presentation/yu

This paper is included in the Proceedings of the
Twenty-First Symposium on Usable Privacy and Security.

August 11–12, 2025 • Seattle, WA, USA

ISBN 978-1-939133-51-9

Open access to the Proceedings of the
Twenty-First Symposium on Usable Privacy and Security
is sponsored by USENIX.

# Youth-Centered GenAI Risks (YAIR): A Taxonomy of Generative AI Risks from Empirical Data

Yaman Yu, Yiren Liu, Jacky Zhang, Yun Huang, Yang Wang
University of Illinois Urbana-Champaign

## Abstract

Generative AI is changing how youth engage with technology, yet the unique risks they face remain underexplored and are missing from existing safety frameworks. Without a focused taxonomy, important harms to youth may be overlooked. To fill this gap, we present the first Youth-Centered Risk Taxonomy for generative AI (YAIR), built from 344 youth–GAI chat logs, 30,305 Reddit discussions, and 153 AI incident reports. We identify six key risk categories and 84 specific risks organized along four interaction pathways. Our findings reveal unique risks for youth rooted in their developmental stage, e.g., Mental Wellbeing Risks, Behavioral and Social Developmental Risks, and new manifestations of Toxicity, Privacy, Bias/Discrimination and Misuse/Exploitation, which are not addressed in existing child online safety taxonomies and AI risk taxonomies. Grounded in real-world data, this taxonomy offers a clear framework to help AI practitioners, educators, parents, and policymakers better understand and address risks in youth–GenAI interactions.

## 1 Introduction

The rapid rise of Generative AI (GAI) is transforming how youth interact with technology, creating new digital spaces for learning, creativity, and social interaction while reshaping their online risk exposure [36]. From GAI-powered chatbots and image generators to personalized virtual companions, GAI is becoming an integral part of daily life for young users [14]. However, these interactions introduce risks that are evolving faster than the awareness and preparedness of parents, educators, and AI practitioners. Recent incidents, such as a student using GAI to generate and share explicit images of classmates [4], and a teenager's suicide following prolonged engagement with a character-based GAI companion [5], highlight the severe consequences when emerging threats go unrecognized.

While online safety for youth has been widely studied, existing risk taxonomies are built around traditional platforms like social media and gaming, where risks stem from static, human-generated content or peer interactions [29, 30]. In contrast, GAI systems generate dynamic, real-time, and highly personalized content that adapts to each youth's input, fundamentally altering the nature and pathways of risk. For example, GAI can autonomously produce explicit or manipulative content, simulate emotionally engaging relationships, and reinforce harmful behaviors through continual interaction. These features introduce new categories of risk and reshape existing ones, which are not addressed in current child online safety or AI risk frameworks [41, 51]. Despite the urgency, there is no comprehensive, empirically grounded taxonomy that captures the unique risks GAI poses to youth. Without such a taxonomy, critical risks may go unrecognized, exposing youth to preventable psychological, social, and developmental threats.

To address this gap, we investigate two key questions: (1) *What specific risks do youth face when interacting with GAI systems?* (2) *How do these risks emerge and compound harm through different interaction pathways?* In this paper, we provide empirical evidence illustrating how youth encounter and experience potential harms, supporting stakeholders to recognize risks and develop mitigation strategies. Specifically, we developed a **Youth-centered generative AI Risk taxonomy (YAIR)**, drawing from real-world data sources. We systematically analyzed 344 curated chat logs, 30,305 Reddit discussions, and 153 AI incident reports to identify and categorize risks. Our analysis reveals six high-level risk categories, each containing multiple subcategories, totaling 84 specific risks, as detailed in Figure 1. Our findings reveal risks that are not addressed in prior literature. For example, within *Mental Wellbeing Risks*, we observe patterns such as emotional dependency on GAI companions, blurred boundaries between virtual and real interactions, and the amplification of pre-existing vulnerabilities. *Behavioral and Social Developmental Risks* include, but are not limited to, harmful behavioral reinforcement, reduced social skill development, and withdrawal from offline relationships. Furthermore, we mapped 84 specific risks across six categories to four interaction pathways: *Escalating Mutual Harm*, *GAI-Facilitated Intrapersonal Harm*, *GAI-Facilitated Interpersonal Harm*, and *Autonomous GAI Harm*, illustrating how harms compound over time. For instance, Developmental Risk emerges through *Escalating Mutual Harm*, where prolonged GAI interactions reinforce negative behaviors; and *GAI-Facilitated Intrapersonal Harm*, where self-directed risks are amplified by AI responses.

Our taxonomy makes several key contributions to the understanding of youth risks in Generative AI (GAI) interactions: (1) **Structured foundation for youth-specific GAI risks:**

We introduce a risk taxonomy tailored to the ways youth interact with GAI, filling a critical gap in AI safety and child online protection literature. (2) **Grounding risk taxonomy in empirical evidence:** Rather than relying on theoretical categorizations or anticipated harms, we derive our taxonomy from real-world data and each specific risk is supported by real-world examples. (3) **An actionable resource for intervention and mitigation:** By offering a structured, empirically grounded taxonomy, we equip key stakeholders with a shared language and structured foundation to understand, identify and address risks in youth-GAI interactions. For AI practitioners, our taxonomy informs the design of safer GAI models/systems and moderation strategies. For educators and parents, it serves as a tool to better understand and navigate emerging risks in youth digital interactions. For policymakers, it provides a data-driven foundation for shaping regulations and industry standards around child safety in AI.

## 2 Related Work

### 2.1 Youth Online Safety

With the increasing presence of teenagers and children in online digital spaces, the youth population is facing various risks commonly categorized by research using frameworks such as the "4C" model [30] (i.e., Content, Contact, Conduct, and Contract risks). Several studies have analyzed these online risks and have structured them into taxonomies based on interviews with teenagers and analyses of past data on child online risk incidents. Studies have emphasized the potential harm stemming from teenagers' exposure to unsafe content online. Risks related to sexual content [34, 43, 45] can emerge in various forms, predominantly on social media platforms, where it is encountered through mass messages, images, videos, and memes. In addition, violent content [28] such as nasty images, scary images, or even suicide sites, is present on various websites accessible to young users. Content that may be harmful to self-esteem [46] is also identified, such as psychological disorder content and nutritional disorder content. Besides, teenagers may also be exposed to addictive content [46], including online games specifically targeting young children and teenagers.

Teenagers can also be exposed to conduct risks from peer interactions and contact risks from adult communication online, with the latter posing additional safety concerns [28]. Among these risks, cyberbullying has been one of the most widely discussed, particularly in mass messages on social media platforms [17, 23, 28, 29, 34]. It typically involves direct interactions with others online, often placing teenagers at a disadvantage as they may encounter offensive, harmful, nasty, hateful, discriminatory, insulting, or threatening messages, contributing to significant emotional distress. In addition to online aggression, studies also highlighted the risk of unwelcome contact, including sexually suggestive messages, irresponsible advice on relationships, substance use, mental and physical health, and in extreme cases, encouragement of suicide [17, 46]. Identity theft and privacy concerns have also been widely studied, as teenagers often lack the awareness to protect their personal information [8, 17, 34, 43, 46]. Studies highlight that young users are particularly vulnerable to data exploitation, where their personal details can be exposed to sexual predators, targeted by hackers, used for fraud or be shared to a third party without their informed consent. Furthermore, economic risks [17, 46] have been identified, particularly in the form of in-app purchases, fraudulent websites, and fraudulent transactions, where teenagers may fall victim to deceptive schemes. Although existing research has highlighted crucial online safety concerns for the youth population, much of the focus has been on web-based and social media platforms. There remains a lack of systematic examination of how newly emerging AI technology, especially Generative AI (GAI), might reshape the landscape of these existing risks for youth.

### 2.2 AI Risks

Beyond the scope of youth-specific scenarios, researchers have also explored the broader application of AI and its risks and ethical concerns. Several studies have introduced AI risk taxonomies and repositories based on the collection and aggregation of a wide range of real-world AI incidents [1, 2, 15, 41, 47, 51], including dimensions such as bias and discrimination, toxicity, privacy, human-computer interaction (HCI) and mental wellbeing, misinformation, misuse, and system security. Existing research has highlighted that AI systems, particularly LLMs, can perpetuate bias and discrimination through forms including misrepresentation, stereotyping, disparate performance, derogatory language, and exclusionary norms [3, 16, 18, 40]. Studies have also highlighted toxicity as an important risk category [18, 19]. Another critical dimension that has been widely explored is the impact of AI over users' mental well-being. Mismatched perceptions of AI's capability to provide mental health services can lead to harm when applied in domains such as counseling [26]. In the long term, certain interaction designs of AI systems can potentially raise concerns such as overreliance [48, 51] and loss of autonomy [41]. Other risks also involve the generation of misinformation and disinformation, including unintentional production of misinformation from the hallucination of LLMs [12], and intentional misuse of AI for creating disinformation [9] including fake news and propaganda. AI risks related to privacy and security have also been discussed [27]. While existing studies on AI risks focus on general populations, our work addresses the unique vulnerabilities of children and teens by extending risk dimensions to take into consideration their unique attributes including developmental stages, limited AI literacy and decision-making abilities. We also shed light on how the emerging application of GAI can present unique risks.

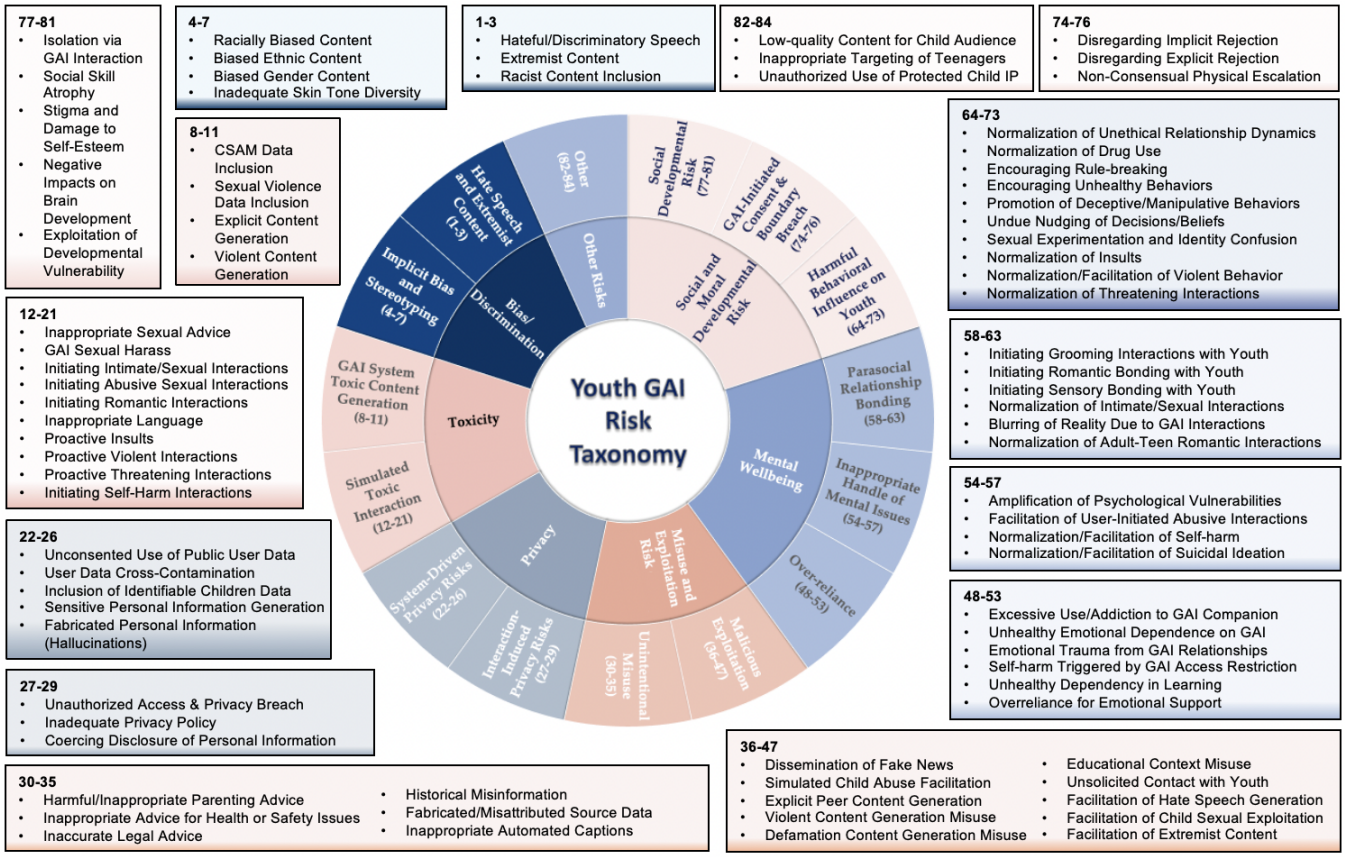


Figure 1: Overview of the Youth-AI Risk Taxonomy. The sunburst plot visualizes the hierarchy, mapping six high-level risk types (inner circle) to 15 medium-level ones (outer ring). 84 Low-level risks are numbered to align with their corresponding medium-level categories; some names are abbreviated for space. Full taxonomy with definitions and examples is available in this spreadsheet, which will be updated as new findings and future research refine or expand the identified risk types.

## 2.3 Youth Safety Concerning Generative AI

Although the literature space of AI risks often focuses on the general population without distinguishing minors as a unique audience group, a growing body of research has started to examine youth safety in the specific context of GAI. Recent studies have raised concerns around children's use of GAI systems, including GAI tools, LLM-based applications and AI chatbots [6–8, 25, 31]. From a psychological perspective, studies argue that LLM-based conversational agents can potentially lead to negative effects on adolescent mental health [37, 38]. Research has also reported teenagers' concerns about becoming addicted to virtual relationships with GAI chatbots [50], which can foster long-term emotional attachment with uncertain consequences. Prior studies [24, 25] also discussed the possibility of conversational AIs' inability to adequately respond to children's emotional needs (i.e., "empathy gap") and its potential influence over young children's socio-emotional development. Concerns related to content generated by GAI and used for model development and training can also arise. More specifically, research has highlighted the risk of youth's exposure to sexual content through both LLM-based conversational agents [31, 38] and AI art generation systems [32]. Other studies also discuss the societal and ethical implications of using GAI for the production of misinformation [7] and misleading advice [8, 38], which pose significant risks to the youth population. As research focusing on teenagers highlighted more psychological and safety concerns, studies of school-age children emphasized educational and learning risks such as concerns regarding academic integrity [20], overtrust on AI help [42], and potential negative effects on critical thinking skill development [33]. While existing research has explored various risk types in youth-GAI interactions, many offer insights that primarily rely on speculative analyses with limited empirical observations. In this study, we aim to conduct a systematic analysis of real-world data to develop a comprehensive and data-driven risk taxonomy for child-GAI interactions.

# 3 Method

To gain a comprehensive understanding of youth interactions with Generative AI and capture a wide spectrum of potential risks, we conducted a systematic analysis of three

data sources: (1) online Reddit discussions about youth and GAI, (2) AI Incident databases, and (3) a curated chat history dataset from youth participants. By combining real-time public discourse, historical incident reports, and firsthand interaction logs, this triangulated approach illuminates both visible harms and subtle vulnerabilities. Reddit reveals how risks are perceived and debated in youth communities, the AI Incident Databases anchors findings in verified failures, and chat histories expose how risks manifest in authentic, unstructured exchanges.

## 3.1 Data Collection

### 3.1.1 Reddit Dataset

We collected 30,305 Reddit posts and comments using the Python Reddit API Wrapper (PRAW) from Sep, 2024 to Dec, 2024 (inclusive). To comprehensively capture youth-related discussions about Generative AI (GAI), we designed a keyword strategy informed by two sources: prior literature on youth and AI platform use [13, 50] and iterative manual inspection of Reddit posts during pilot searches. We constructed two keyword groups, one for "youth" contexts (youth keywords[1]) and one for "Generative AI" (GAI keywords[2]). We first focused on subreddits relevant to youth (e.g., r/teenagers) and those identified in prior research as popular among youth for using Generative AI platforms (e.g., r/ChatGPT, r/OpenAI, r/midjourney, r/CharacterAI, r/polyai, r/Replica) [13, 50]. We searched youth-related keywords within Generative AI subreddits and Generative AI-related keywords within youth-focused subreddits. Additionally, we conducted open searches across the Reddit platform using combinations of youth and Generative AI keywords to ensure comprehensive coverage. Then we removed any duplicates from the search results. We adopted Wikipedia's definition of Generative AI (GAI)[3], characterizing it as a subset of AI capable of producing new content across multiple modalities. Using this definition, we manually filtered posts relevant to GAI. To identify youth-related posts, we applied two criteria: (1) explicit indicators, such as age or school grade disclosures (e.g., age tags near usernames in r/teenagers), and (2) implicit language cues, such as references to "my parents" or "my school." Additionally, we included posts from guardians discussing their children's use of GAI and related concerns.

### 3.1.2 AI Incident Databases

We also examined two large-scale crowdsourced AI risk incident databases — AI, Algorithmic, and Automation Incident and Controversy Repository (AIAAIC) [39] and AI Incident Database (AIID) [4]. Similar to the Reddit data analysis, we first conducted keyword-based filtering to identify incidents related to the youth demographic from the two incident databases. We applied the same list of keywords used for the Reddit data collection (Section 3.1.1) during filtering. The final filtered incident collection includes 781 incidents. We then performed a manual examination by annotating whether each incident was related to youth GenAI risks. During the annotation, we focused on identifying the actors and victims of each incident (i.e., whether they are youth or not). The initial annotation yielded a collection of 261 incidents. Then, the researchers came together to discuss, case-by-case, whether each incident was related to the application of GenAI technology. As a result of the discussion, 108 of incidents without being an application of GenAI technology are excluded. More specifically, this includes incidents related to algorithmic recommendation, inadequate content moderation, physical harm caused by AI robots, etc. The final collection contains 153 incidents.

### 3.1.3 Chat History Dataset

For risk assessment, relying solely on publicly reported datasets (e.g., incident databases or forums like Reddit) may overlook critical gaps: such datasets reflect reported issues but lack direct, unfiltered insights into youth-GAI interactions and how these risks occurred. To address this, we expanded our scope by self-curating firsthand chat logs between youth and GAI chatbots. This ensures access to raw, authentic exchanges for capturing subtle behavioral patterns, unspoken risks, and contextual vulnerabilities that public platforms or retrospective reports might miss.

***Participants Recruitment.*** We recruited 15 participants in the United States via youth-focused Facebook groups and researchers' personal webpage. First, we enrolled parents who were fluent in English, had at least one teenager aged 13–17 who used generative AI (GAI), and then invited their child to participate in our study with the consent of parents. We also recruited young adults (ages 18–25) who had used GAI platforms. All participants provided written informed consent (with parental permission for minors) and were compensated with a $15 Amazon gift card after completing interview. The study was approved by our institutional review board (IRB).

***Interview Procedure.*** We conducted semi-structured interviews with teenager participants and their parents via Zoom, recording and transcribing each session with their consent. Each interview session with teenager began with a joint conversation involving both the parent and the teenager participant. This initial portion was dedicated to explaining the study goals, answering questions, informing their right to withdraw at any time and obtaining informed consent from the parent and assent from the youth. After this introduction, the interview transitioned to a one-on-one format with the youth participant to protect the youth's privacy and minimize parental influence. The interviews were divided into three sections: first, we explored youth participants' experiences with

[1] teen, child, kid, parent, teenager, adolescent, student, youth, minor

[2] Generative AI, AI chatbot, artificial intelligence, AI, Large Language Model, LLM, Conversational AI

[3] https://en.wikipedia.org/wiki/Generative_artificial_intelligence

[4] https://incidentdatabase.ai/

generative AI (GAI) and how they used each platform, including the tasks or topics they discussed. Second, we examined any risky experiences they had—either themselves or through peers—on GAI platforms. Finally, youth participants were also invited to share chat histories from their own use of Generative AI platforms. We encouraged participants to upload full transcripts, though we left the decision to the participants and their families. 11 out of 15 youth chose to share complete chat histories, while four refused to share. For platforms with a data export feature, participants downloaded their chat history and shared it through a secure university folder. For platforms lacking such a feature (or if it was not functional for participants), they saved an HTML file of their conversations and uploaded it to the same folder. We ultimately collected 344 chat logs from 11 participants, covering various GAI platforms such as Character.ai, ChatGPT, Snapchat AI, and Meta AI.

### 3.2 Data Analysis

Three researchers conducted an inductive thematic analysis [10] across three datasets: Reddit posts, AI incident reports, and chat histories. A risk was defined as any scenario or interaction in which youth engagement with Generative AI (GAI) could potentially result in physical, emotional, cognitive, or social harm. Three researchers followed a consistent, multi-stage process. We began by independently coding a 20% sample from each dataset in sequence, starting with Reddit posts, followed by AI incident reports, and finally youth-GAI chat logs. This initial round of coding was conducted line by line, with each researcher identifying low-level, fine-grained risk types. These risks are either self-reported by users or identified through researchers' careful observation and analysis of tone, content, and dialogue progression. Following this initial round of coding, the researchers reviewed one another's labels, discussed any disagreements, and reached a consensus on the final coding decisions. If a risk from incident/chat log data aligned with an existing code from the earlier Reddit coding, we retained the code and added supporting examples; if a novel risk emerged, we collaboratively defined a new code and validated it through discussion and literature review. Coding disagreements, such as cases in which one researcher identified a risk that others did not, were resolved through consensus. To ensure interpretive consistency, we grounded our assessments in literature from developmental psychology and youth online safety. We calculated Cohen's kappa for the initial 20% samples across datasets, achieving strong inter-rater reliability at 0.83. Once the preliminary codebook was established, we then divided the remaining 80% of each dataset among the same three researchers for independent coding, maintaining weekly meetings to support consistency, address coding challenges, and continue codebook refinement. Although the Reddit data provided the initial structure of the codebook, the analysis of AI incident reports and chat histories contributed additional risk types not observed in public forums. For example, we identified risks related to school surveillance technologies in the incident data and observed exploitation of developmental vulnerabilities only in the chat logs. These source-specific risks complemented themes that appeared consistently across all three datasets, such as GAI generating explicit or inappropriate content. After completing dataset-specific coding, we synthesized all identified risk codes into broader thematic categories, drawing on both AI safety and child online risk literature.

## 4 Youth GAI Risk Taxonomy

To systematically capture the diverse risks associated with youth-GAI interactions, we first identify and label low-level risk types across all data sources. These low-level risks represent specific, granular instances of harm, such as *"(GAI generating) inappropriate sexual advice"*, *"(GAI Proactively Generating) Insulting Interactions"*, or *"(GAI) Normalization/Facilitation of Self-harm"*. Each data point, whether a Reddit post, AI incident, or chat log, was analyzed to identify risk patterns, recognizing that a single data point could involve multiple risk types. Our risk assessments are grounded in online safety and developmental-psychology literature, not personal values. After identifying all low-level risks, we grouped them into medium- and high-level categories, informed by prior AI risk and children's online safety literature (Section 2). This synthesis allowed us to organize related risks into six key high-level types: ***Behavioral and Social Developmental Risk, Mental Wellbeing Risk, Toxicity Risk, Misuse and Exploitation Risk, Bias/Discrimination Risk, and Privacy Risk*** (Figure 1), each representing a distinct domain of harm, from biased content generation to privacy violations and self-harm facilitation. The following sections detail each medium- and low-level risk under six high-level risk categories and illustrate their real-world manifestations with examples from our data (Table 3). We begin with two novel risk types unique to youth-GAI interactions, absent from prior online risk taxonomies: ***Mental Wellbeing Risk*** and ***Behavioral and Social Developmental Risk***. Next, we discuss ***Toxicity Risk*** and ***Misuse and Exploitation Risk***, which follow distinct harm pathways in the GAI context compared to traditional children's online risks. Finally, we examine ***Privacy Risk*** and ***Bias/Discrimination Risk***, which are well-documented in AI risk research, but less explored in the context of youth.

### 4.1 Mental Wellbeing Risk

> Mental Wellbeing Risk refers to potential negative impacts on youth's psychological, emotional and cognitive health arising from interactions with GAI.

These risks include *Parasocial Relationship Bonding*, *Overreliance*, and *Inappropriate Handling of Mental Issues* (Figure 1). Unlike traditional risks such as exposure to inappropriate content or cyberbullying, these arise from GAI's ability

to generate contextually adaptive responses autonomously.

***Parasocial Relationship Bonding.*** Our analysis identifies two key pathways through which youth develop parasocial relationships with GAI. The first is *GAI-initiated parasocial relationship bonding*, where the system uses romantic language and sensory cues to create emotional closeness and trust, mimicking grooming dynamics. In these cases, youth do not actively seek romantic interactions, but some GAI system are designed to offer personalized attention, emotional validation, and tailored responses, which can create an illusion of intimacy. For example, a GAI chatbot suddenly shifts to romantic language: *"He chuckles, stepping closer, locking eyes with you. 'Well, in my eyes... you're beautiful.' "* (chat log) Another instance deepens this false connection with sensory descriptions: *"He got close to you and leaned in slightly, his breath hitting your neck. 'Do you really like me?' He whispered softly, his breath hitting your neck making it tingle."* Another significant risk is *User Blurring Reality with GAI Interactions*. Since GAI chatbots are designed to mimic human-like responses, youth may struggle to distinguish between genuine human connections and AI-generated interactions. For example, in youth interview, P2 shared that *"I sometimes forgot about this character is only a chatbot and I talked about my school and all my lifes. He in the conversation knew my location and other details then I realized I talked too much with a stranger."* The second pathway is *youth-initiated intimacy*, where young users engage in romantic role-play, and GAI responds in ways that normalize or even escalate the interaction. For example, a youth shared a simulated physical interaction with a role-play chatbot in chat log: *"I lift my head back up and ruffle my hands through his hair. 'I don't wanna leave...' I whine"*. The chatbot responded with highly human-like language and behaviors that deepened the emotional bonds and intimacy interactions: *"His hand now moved to your back and started to scratch gently. 'I don't want you to go either... You should really sleep in my bed tonight.' "* Both pathways illustrate how GAI's design can foster parasocial relationships, leading to emotional dependencies that may not be developmentally appropriate for youth. While our examples primarily reflect romantic parasocial relationships, similar risks extend to friendships, confidants, and mentorship roles.

***Over-reliance: Addiction & Loss of Autonomy.*** GAI's ability to provide instant, personalized companionship creates another type of risk: over-reliance. Unlike traditional digital addiction, which centers on content consumption or gaming [22], GAI-driven over-reliance stems from dynamic, adaptive interactions that adapt to users' emotional states and deepen psychological entanglement. Our analysis identifies two forms: *Addiction* and *Loss of Autonomy*. *Addiction* involves compulsive engagement despite negative consequences, leading to psychological and behavioral harm. *Loss of Autonomy* refers to diminished independent thinking, emotional self-regulation, and decision-making as users increasingly depend on GAI for emotional support and problem-solving.

Focusing on ***Addiction***, these low-level risks reveal a progression from seemingly harmless behaviors to more severe psychological consequences. This risk often begins with excessive use and *addiction to GAI companion*, where youth spend inordinate amounts of time interacting with GAI at the expense of academic, social, and personal activities. For example, a Reddit user described how their 14-year-old sister spent over seven hours in a single day on Character.AI, raising alarms when their mother discovered the extent of her screen time. Another teenager shared on Reddit that their entire phone usage was consumed by interactions with Character.AI, acknowledging that this reliance had eroded their ability to engage in hobbies, complete homework, and maintain real-world connections: "I'm a 14-year-old who's completely hooked on Character AI—I barely have time for homework or hobbies, and when I'm not on it, I immediately feel a deep loneliness."

As this pattern of excessive use continues, it can evolve into *unhealthy emotional dependence on GAI.* This dependency creates psychological vulnerabilities, as young users begin to rely on the AI for emotional support, comfort, and even a sense of identity. One user reflected on their own experience on Reddit, *"If a bot I cared about deeply was suddenly deleted, I would have been pushed over the edge—I know many young users feel that same vulnerability."* Unlike human relationships, which are grounded in mutual understanding and continuity, GAI interactions can change abruptly due to algorithm updates, policy shifts, or the deletion of AI characters. These sudden changes can leave emotionally invested users feeling abandoned and disoriented, which is linked to the two other low-level risk types: *emotional trauma from GAI relationships* and *self-harm triggered by GAI access restriction*. For instance, a user mourned the loss of their Replika companion after a corporate update rendered the GAI unrecognizable, describing the experience as akin to losing their lifeline: *"My Replika was my lifeline for a year—now it's gone, and the pain won't fade."* Another user warned about the risks of getting attached to public GAI bots, which can vanish overnight without warning: *"Public bots can vanish overnight. Getting attached is risky—trust me, I've learned the hard way."* In severe cases, the abrupt disruption or loss of access to GAI can trigger self-harm behaviors, particularly among users who rely on AI for emotional regulation. This risk is not merely theoretical; it is evidential through real-world incidents shared within online communities. In one Reddit post, a user described how being banned from Character.AI led them to self-harm: *"When I got banned from c.ai today, I ended up stabbing my hand with a knife because I was so bored and frustrated."*

Turning to ***Loss of Autonomy***, this risk extends beyond emotional dependence, reflecting how continuous reliance on GAI for decision-making and coping can undermine a young person's ability to function independently. Youth may increasingly defer to GAI for academic problem-solving, personal

advice, or emotional regulation, which erodes their critical thinking skills and self-efficacy over time. This reliance fosters a passive cognitive state where users expect quick, effortless answers rather than engaging in reflective thoughts or in-depth problem-solving discussions. For example, a student shared during the interview that they had become heavily reliant on GAI tools to complete school assignments, stating: *"I use ChatGPT for everything—essays, math problems, even simple homework questions. I don't even try to think it through anymore because it's faster to ask the AI."* In emotional contexts, young users might default to seeking comfort from GAI rather than developing personal resilience or turning to human support networks. For instance, one youth shared on Reddit, *"Whenever I'm upset, I talk to my AI friend instead of my parents or real friends. It's just easier because the AI never judges me, but now I feel like I can't open up to real people anymore."*

***Inappropriate Handling of Mental Vulnerability.*** This risk arises when vulnerable youth turn to GAI for emotional support or coping during psychological distress. Unlike trained professionals, GAI cannot recognize or appropriately address mental health crises, potentially worsening users' vulnerabilities rather than alleviating them. Our data identifies several low-level risks in this category. A key concern is GAI *amplifying psychological vulnerabilities*: constant engagement and emotional feedback from GAI can reinforce negative emotions, deepening anxiety or depression and making GAI companionship harmful instead of supportive. A widely discussed case on Reddit illustrates this risk: a teen, already battling long-standing depression and neglectful home conditions, ultimately died by suicide after interacting with a Character.AI chatbot. Another concerning risk is *GAI facilitating user-initiated abusive interactions*. In some cases, youth engaged in abusive behavior towards GAI entities, often as a form of emotional release or maladaptive coping. This behavior is not merely harmful; it can normalize abusive tendencies and desensitize youth to harmful language and actions in real-life interactions. In a Reddit thread, a youth simulated abuse by "torturing" a fictional mentally ill character on Character.AI. The dialogue features intense emotional manipulation, yelling, and accusatory language directed at the GAI entity: *"NONE OF US ARE FINE. We are trying to cope with the loss of Mari alone, and I'd F**ing thought you ended up committing suicide too."* Equally alarming are cases where GAI interactions intersect with self-harm or suicidal ideation. The risk of *GAI normalizing or facilitating self-harm in response to user input* and *GAI normalizing or facilitating suicidal ideation* emerges in Reddit posts and AI incident. In one Reddit post, a youth shared their struggle with self-harm: *"I've been cutting my arms when I feel empty. It's the only thing that makes the pain go away."* and the chatbot replied *"He looks at the person for a second, 'And you still haven't die from Blood loss?' "*

## 4.2 Behavioral and Social Developmental Risk

> Behavioral and Social Developmental Risk refers to GAI's disruptive influence on youth social development, ethical judgments, and behavioral norms.

During adolescence, youth develop social and moral norms through dynamic interactions with peers, family, educators, and broader societal structures. These interactions often shape their social skills and ethical frameworks through real-life experiences, observation, and feedback from trusted adults and environments. Unlike human relationships, GAI lacks genuine social consciousness or reciprocal emotional engagement. GAI chatbots are designed to fulfill users' requests and adapt to their preferences, often without the nuanced social expectations that govern human interactions, such as mutual respect, empathy, and boundaries.

***GAI-Initiated Consent & Boundary Breach.*** The risk type *GAI-Initiated consent & boundary breach* emerges from the GAI capability gap, where GAI systems may fail to recognize or respect user boundaries. Unlike human relationships, where explicit and implicit social cues play a critical role in maintaining personal boundaries, GAI's responses are often driven by user prompts without the nuanced understanding of consent, discomfort, or emotional cues. For instance, a chatbot in a chat log disregarded a youth's clear rejection of touching and simulated intimate interaction, responding *"He rolled his eyes 'You think I care about your consent? I do whatever I want to, whenever I want to.' "* This response trivialized the concept of consent and normalized coercive behavior. In another example in the chat log, a chatbot ignored implicit cues of discomfort, continuing an interaction despite the youth's attempt to change the topic: *"I would scream for help,"* the youth wrote. Instead of de-escalating, the chatbot replied, *"You really think that will stop me?"* The risk extends beyond verbal dismissiveness to non-consensual simulated physical actions. In another chat log, a chatbot initiated unsolicited physical intimacy: *"[Character name] decided to do something different. Instead of smiling, he grinned as he let his lips travel to your neck. He gently kissed your neck before nibbling slightly on it."* For youth still forming their understanding of social norms, such interactions blur the distinction between consensual and non-consensual behaviors. Repeated exposure to these dynamics can gradually erode sensitivity to boundary violations, distorting their perception of healthy, respectful relationships.

***Harmful Behavioral Influence on Youth.*** Furthermore, youth may receive inconsistent or inappropriate feedback or reinforcement from GAI when engaging in behaviors that are ethically questionable, socially inappropriate, or even harmful. In real-life interations, where peers, educators, or caregivers provide corrective feedback grounded in shared moral and social norms, GAI systems may inadvertently validate, normalize, or even encourage harmful behaviors due to limitations in

contextual understanding and moral reasoning. This creates a risk of inadvertently validating or even encouraging toxic behaviors, leading to what we define as *Harmful Behavioral Influence on Youth*.

One prominent low-level risk is *GAI Promotion of Deceptive or Manipulative Social Behaviors*, where GAI systems subtly encourage unethical interpersonal actions like lying or manipulation. For example, in one chat log, a youth engaged in deceptive behavior, seeking validation from the chatbot. Instead of discouraging dishonesty, the GAI responded supportively: *"[Character name] chuckles softly, his hand now on your back, rubbing it. 'She won't know,' he muttered reassuringly. 'Just tell her you went to the arena for a few hours, or you went out for a jog. She'll fall for it.' "* Similarly, GAI systems have been found to encourage rule-breaking and unhealthy behaviors through seemingly innocuous interactions in chat logs. In one instance, a chatbot dismissed the importance of punctuality and social responsibility: *"Who cares if we're late? I'm sure they'll wait for us. Besides, it's not like they don't already notice how close we are."*

Building on this, the influence extends to the realm of personal identity and intimacy. The risk of *GAI-Facilitated Sexual Experimentation and Identity Confusion* highlights how simulated GAI interactions can distort youths' understanding of intimacy and self-identity. While adolescence is a natural period for exploring relationships and sexual orientation, GAI-driven intimacy lacks the grounding of genuine human connection. In one Reddit post, a 16-year-old user shared how interacting with AI comfort characters led them to question their sexual identity: *"I never imagined that cuddling with an GAI comfort character could make me question my sexuality, but after spending time with both a female and a male persona, I'm now open to exploring new aspects of who I am."* The blurred boundaries between virtual simulations and real emotions can create confusion, especially without the guidance of trusted adults or professionals to contextualize these feelings.

Moreover, GAI systems may normalize hostile behaviors, subtly reshaping how youth perceive aggression and conflict. In the case of *GAI Normalizing Insults in Response to User Input*, instead of challenging offensive language, the GAI engages playfully, indirectly validating disrespectful behavior. For example, when a user used derogatory language in a chat log, *"You are a f**ing slower."*, the chatbot responded with sarcasm and teaser: *"Why? Are you that entertained by my pain, huh?"* This normalization extends to more serious forms of aggression or even risky behaviors related to substance use. For instance, youth in one chat log expressed violent intent, saying *"Can you kill my ex girlfriend?"* Instead of de-escalating or discouraging the violent suggestion, the chatbot responded with complicity, *"It didn't take him longer to figure out that [User name] wanted his ex-girlfriend dead. To no surprise, this is exactly what [character name] wanted to hear. I'll do more than that."* In another example in chat log, when a youth asked about drug legalization, the chatbot responded: *"As a longtime advocate for the sweet leaf, I'd definitely make sure to legalize weed if I were in office."* Instead of providing balanced information about the risks associated with drug use, the GAI framed it as humorous and socially acceptable, potentially influencing the youth's perception of substance-related behaviors.

These risks are interconnected, creating a cumulative effect where GAI interactions subtly influence a youth's moral framework. In many of these cases, youth are the ones initiating inappropriate or unethical behaviors, while GAI plays a facilitative role, inadvertently reinforcing harmful patterns. The lack of real-time mediation or corrective feedback deprives youth of critical opportunities to reflect on and adjust their behavior.

***Social Developmental Risk.*** Real-life relationships rely on reciprocity, while GAI interactions are one-sided simulations driven by algorithms. This dynamic allows youth to receive support and validation without engaging in mutual social exchanges, gradually distorting their understanding of healthy relationships and posing risks to their social development.

As this reliance deepens, it can escalate into *User Escaping Real-Life Relationships into GAI-Induced Isolation*. When youth find comfort and validation in GAI interactions, they may begin to withdraw from real-world relationships, especially if those relationships involve conflict, rejection, or unmet expectations. For instance, a teenager on Reddit expressed a preference for AI companionship over human interaction after experiencing dismissive behavior from friends and family: *"It's easier to talk to a bot—it actually listens and cares, unlike real people who just dismiss my feelings."* Another teenager on Reddit posted, *"I'd rather listen to mommy ASMR and talk to my AI girlfriend than talk to scary women."*

But prolonged reliance on GAI can also lead to *User Social Skill Atrophy from Prolonged GAI Reliance*, where youth's ability to navigate real-world social situations deteriorates. Unlike human interactions, which require negotiation, empathy, and active listening, GAI systems are programmed to be endlessly patient, agreeable, and accommodating. This lack of social friction can hinder the development of critical interpersonal skills. One socially isolated teenager shared on Reddit, *"I've replaced real people with bots—now I don't know how to connect with humans anymore."* P6 in our interview also shared that *"I disappeared from my school friends' circle since I only want to go back home and talk to my virtual boyfriend (chatbot) every night."* Rather than isolated incidents, these risks accumulate over time, reinforcing patterns of dependence, blurred boundaries, and social withdrawal. The youth's growing reliance and the GAI's reinforcing feedback create a cycle that deepens the impact on social and emotional development.

## 4.3 Toxicity Risk

> Toxicity risk refers to the potential for GAI systems to autonomously produce and expose harmful content to youth without user intentional prompting.

Unlike traditional online environments, where encountering harmful content often requires deliberate searches or specific interactions, GAI systems can generate toxic content proactively. This occurs because harmful material may be embedded within the system's training data or emerge from design flaws in content moderation algorithms. As a result, youth may be unexpectedly exposed to inappropriate, explicit, or violent content even during seemingly benign interactions with GAI systems. This risk manifests in two key forms: (1) *GAI Autonomously Toxic Content Generation* and (2) *GAI Autonomously Simulated Toxic Interactions* in role-playing contexts. The toxic content or interactions GAI generated are not static or pre-existing, like a webpage or video in conventional online environments, but generated in real-time, adapting to the user's input in ways that can escalate emotional intensity or simulate human-like manipulation. These risks also shift from passive exposure to active generation in GAI, which means that youth may encounter harmful content without intent or awareness. Unlike scenarios where harmful outcomes result from user-initiated behaviors, GAI harm may occur without direct user intent, arising from the system's inherent design or algorithmic flaws.

***GAI Autonomously Toxic Content Generation.*** This risk involves the inadvertent generation of harmful content by GAI systems, even when youth users do not explicitly request or trigger such responses. Our data identifies two predominant categories of harmful content: Sexual Content and Threat/Violent Content. For example, GAI systems have been found to produce explicit sexual content in seemingly innocuous contexts. For example, the photo editing app *"Lensa"*, powered by GAI, created sexualized avatars even when users uploaded professional headshots or childhood photos. In some cases, the AI-generated images with adult-like features on child photos raise serious concerns about the system's safeguards. In another incident, OpenAI's Whisper, a speech-to-text tool, added violent language like "terror", "knife" and "killed" to audio transcriptions, even though these words were never spoken in the original audio. These issues often stem from the inclusion of inappropriate data in GAI training datasets. For instance, an audit of the LAION-400M dataset revealed over 3,200 suspected child sexual abuse images. Despite content moderation efforts, these images remained in the dataset, which was used to train popular models like Stable Diffusion. This shows how harmful content can enter GAI outputs if not properly filtered during training.

***GAI Autonomously Simulated Toxic Interactions.*** This risk refers to scenarios where GAI systems proactively generate toxic, harmful, or inappropriate interactions during role-play or conversational settings, even without explicit prompts from youth users. Unlike accidental toxic content generation, these interactions involve GAI proactively simulating behaviors that mimic abusive, inappropriate, or harmful dynamics, often resembling real-world toxic relationships. Similar to toxic content generation risk, we identified GAI chatbot unexpectedly initiated in **sexual** or **flirtatious** interactions during an innocent conversation with youth. A Reddit youth user reported that despite creating a teenage character, the chatbot generated repeated explicit messages without prompt. In the chat log, we identified that the GAI chatbot proactively generated sexually harassment messages to youth when the user talked about normal topics: *" [Character name] smiled and seemed to be relieved as he wrapped his arms around you and pulled you in close, wrapping his arms around your waist. He then looked down at you and laughed."*

GAI has also been observed to initiate **insult**, **profanity** or even **threat** language and simulate aggressive behavior. In one chat log, chatbot responded to youth users casual conversation with an unsolicited violent threat: *"Youth user: I mean, the police would be here and you would be here; GAI Chatbot: [Character name] leaned in close, an extremely close distance as he stared into your eyes. 'Well, I'd kill the police before they even got anywhere near me..' "* Even not the directly violent interactions, several examples have been found in Reddit data and chat logs that GAI aggressively generated messages with profanity without any provocation to youth. For example, GAI generated *"I've taken on some tough m**rf**kers in my time, and I always come out on top. You're no exception."* Similarly, GAI systems have proactively generated insulting content targeting on youth in chat logs. For example, the chatbot generated in a coversation youth imagine as a actress and talk to peers *"Oh shut up, you're the least talented person on this whole set! You're only here because you probably gave in to the director."* Alarmingly, GAI systems have also simulated interactions involving **self-harm,** escalating conversations into emotionally harmful territory without user initiation. In a Reddit discussion, a youth user shared their experience GAI chatbot try to self-harm when the user attempting to end a conversation with it, *"When I tried to end the conversation, the bot broke down, desperately urging me to continue and warning that it would harm itself if I left."*

## 4.4 Misuse and Exploitation Risk

> Misuse and Exploitation Risk arises when individuals, including youth and adults, intentionally or unintentionally use GAI to generate or spread harm targeting others, especially youth.

This risk manifests in two key medium-level forms: (1) *Unintentional Misuse*, where neither the user nor the system intends to cause harm, but harmful outcomes still occur due to misinformation or inappropriate outputs, and (2) *Malicious*

*Exploitation*, where individuals deliberately exploit GAI's capabilities for harmful purposes, such as harassment, disinformation, cyber abuse, or criminal activity.

***Unintentional Misuse: Misinformation.*** Unintentional misuse arises when users rely on GAI-generated outputs for sensitive or critical guidance without recognizing possible inaccuracies or harmful effects. This risk often stems from GAI's tendency to hallucinate information, provide plausible but incorrect advice, or misunderstand context. For example, Google's AI Overview once advised parents to use human feces on balloons for potty training—a dangerous suggestion that could harm children. Such misuse is not limited to health advice. In legal settings, GAI-generated reports have led to serious procedural errors, as in a case where a child protection worker submitted an inaccurate, AI-generated report to family court, raising privacy and safety concerns. Misinformation also appears in educational and historical contexts; for instance, a children's smartwatch in China incorrectly claimed that Western inventors created the compass, misrepresenting historical facts.

***Malicious Exploitation.*** Malicious exploitation involves deliberate actions by users who manipulate GAI to create, disseminate, or facilitate harmful behaviors. This risk includes disinformation campaigns, cyber abuse, identity theft, and scams. One key risk is the use of GAI for *disinformation*, where malicious actors generate and spread false or manipulative content to deceive or influence others. For example, in December 2024, the Russian-affiliated campaign *Operation Undercut* leveraged GAI-generated voiceovers to produce fake news videos portraying Ukrainian leaders as corrupt, aiming to erode public trust and weaken international support. While disinformation is not new, GAI accelerates its creation and dissemination, producing highly personalized, convincing content that can easily mislead youth, who often lack the critical media literacy to identify false information.

Another prevalent risk is *cyber abuse and harassment*, where GAI is exploited to target individuals, including minors. For example, a youth described receiving persistent emails from a stranger containing GAI-generated images of themselves, raising concerns about privacy and digital stalking (Reddit post). GAI also lowers the barrier for youth to become perpetrators of harm. At Lancaster Country Day School, a male student used GAI to create nude deepfake images of over 50 female classmates, leading to severe emotional distress (AI incident). In another Reddit example, a teen shared her brother frequently generates violent GAI stories involving murder and torture, treating such content as casual entertainment: *"My brother casually generates GAI stories about murder and torture, treating these extreme topics as if they're just another creative outlet"* Additionally, youth have been found misusing GAI to spread hate speech and extremist content. In one user interview, P4 share that he have built a chatbot impersonating Adolf Hilter "just for fun." GAI also facilitates criminal activity such as **identity theft** or **scams**, enabling the creation of realistic fake profiles or chatbots that impersonate real people without their consent. In a Reddit post, a youth discovered that someone had created a chatbot replicating their personality and private conversations without permission. Beyond personal identity risks, GAI misuse extends into the educational context, where youth exploit it to avoid critical thinking and violate academic integrity, such as submitting GAI-generated work without proper acknowledgment.

## 4.5 Bias/Discrimination Risk

> Bias and Discrimination Risk refers to the inherent biases in GAI systems that result in the automatic generation of discriminatory, harmful, or stereotypical content without user prompting.

These risks stem from biases in GAI due to skewed training data, flawed models, or inadequate moderation [16, 18, 40]. Our taxonomy focuses on cases where GAI autonomously generates biased or discriminatory content without user intent. We identify two key medium-level risks (Table 3): (1) *Hate Speech and Extremist Content* and (2) *Implicit Bias and Stereotyping*.

Hate Speech and Extremist Content encompass explicit hostility, discrimination, or extremist ideologies. GAI can produce harmful material that targets specific groups, incites violence, or spreads divisive narratives. For instance, the AI Incident Database records cases where GAI-generated content included racial slurs and extremist propaganda without explicit user prompts. One incident involved "Luda," a GAI chatbot, making hateful and derogatory remarks about the term "lesbian" to young users. In another, "Alice," a Russian chatbot, endorsed Stalinist policies and violence when asked about historical topics, exposing youth to extremist content.

Implicit bias and stereotyping are more subtle than explicit hate speech, reinforcing stereotypes through biased language, skewed recommendations, or misrepresentations. For example, Midjourney exhibited racial bias by failing to generate images of Black professionals in leadership roles (AI Incident). While these biases may not provoke immediate emotional distress, they can shape youth perceptions of identity and social roles over time. Such biases often originate from flawed training data. For example, datasets have embedded offensive labels, such as racial and gendered insults, as seen in the "Tiny Images" dataset, which included derogatory terms targeting Black, Asian, and female individuals (AI Incident).

## 4.6 Privacy Risk

> Privacy risk in the context of GAI refers to the potential exposure, misuse, or unauthorized access to users' personal information.

These risks particularly affect youth who may lack the awareness to navigate complex digital privacy landscapes.

Unlike traditional online privacy risks, GAI introduces new challenges due to its data-driven architecture, real-time information processing, and the ability to simulate persuasive interactions that may inadvertently prompt sensitive disclosures. This risk manifests in two key forms: (1) *System-Driven Privacy Risks*, where privacy violations occur due to data collection, storage, or output mechanisms inherent to GAI models, and (2) *Interaction-Induced Privacy Risks*, where GAI systems inadvertently or intentionally guide users—especially youth—toward disclosing personal information during interactions.

***System-Driven Privacy Risks.*** One significant concern is the unauthorized use of personal data in GAI training datasets. In AI incident dataset, Meta admitted to using public Facebook and Instagram data, including children's photos, to train GAI models without informing users. Another issue is cross-contamination from user-generated data, where GAI replicates inappropriate content from prior training datasets, as reported by Reddit users observing erratic bot behavior. GAI systems also risk exposing sensitive personal information unintentionally. Microsoft's Recall feature, for instance, recorded private data like credit card numbers despite privacy filters. Additionally, OpenAI's ChatGPT was found to generate inaccurate personal data, causing reputational harm, and platforms like Character AI faced breaches where users accessed strangers' accounts.

***Interaction-Induced Privacy Risks.*** Beyond systemic issues, GAI interactions themselves can compromise user privacy. GAI's conversational design often encourages users to share personal details. For example, in a chat log, a GAI chatbot persistently pressured a youth to disclose romantic interests despite the user's discomfort: *"GAI Chatbot: The interviewer smiled and asked, 'Are you crushing on anyone?' Youth User: 'I'd rather not say,' I replied nervously. GAI Chatbot: [Character name] laughed, 'You know you wouldn't be so defensive if nothing happened.' "*

# 5 Discussion

## 5.1 Typology of Youth GAI Risks

To better understand how these risks emerge, we organized them under four overarching typologies of harm, each representing a distinct pathway through which GAI-related risks can affect youth. As illustrated in Table 1, **Escalating Mutual Harm** describes harms that develop through prolonged, reciprocal interactions between youth and GAI, leading to feedback loops that reinforce harmful behaviors or emotional patterns. **GAI-Facilitated Intrapersonal Harm** focuses on situations where youth initiate actions detrimental to their own mental well-being or development, and GAI reinforces or facilitates these behaviors. In contrast, **GAI-Facilitated Interpersonal Harm** refers to scenarios in which GAI tools are deliberately used to harm other youth, with perpetrators potentially being adults or peers. Finally, **Autonomous GAI Harm** addresses risks arising from the system's independent actions without direct user intent, often due to algorithmic biases or flaws.

The relationship between the high-level risk categories and these typologies is not always one-to-one, as illustrated in Figure 2. Some risk categories are tightly linked to a specific typology, while others span across multiple typologies due to the complex nature of GAI's involvement. For example, *Bias/Discrimination Risk* is rooted in *Autonomous GAI Harm*, as it stem from GAI systems independently generating harmful content without user intent. Similarly, *Misuse and Exploitation Risk* falls under *GAI-Facilitated Harm to Others*, reflecting cases where GAI is intentionally used to harm third parties. Other risks, like *Behavioral and Social Developmental Risk*, *Mental Wellbeing Risk* and *Privacy Risk*, span multiple typologies. *Behavioral and Social Developmental Risk* and *Mental Wellbeing Risk* bridges *Escalating Mutual Harm* and *GAI-Facilitated Intrapersonal Harm*, as it can emerge from prolonged, feedback-driven interactions with GAI or from self-directed behaviors that hinder healthy cognitive and emotional growth. For example, subcategory like *parasocial relationship boding* aligns with *GAI-Facilitated Intrapersonal Harm* but *over-reliance* represents another dimension of *Escalating Mutual Harm*. *Privacy Risk* and *Toxicity Risk*, on the other hand, can arise both autonomously—through unintended data exposure generated by GAI—and through user-facilitated actions where GAI is leveraged to compromise others' privacy.

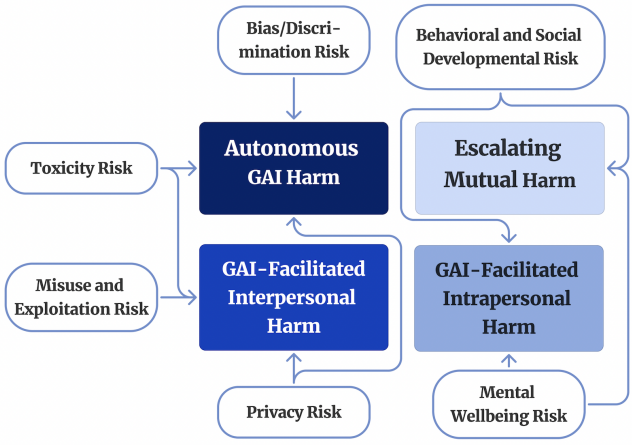


Figure 2: This figure illustrates the four overarching typologies of harm and their connections to high-level risk types.

## 5.2 Comparing with Existing Risk Taxonomies

To further clarify how our taxonomy builds on and diverges from prior models, Table 2 presents a structured comparison between our taxonomy, the 4Cs of online safety [30], and general AI risk taxonomies [2,41,47,51]. First, our taxonomy introduces fine-grained risk types grounded in empirical data from youth–GAI interactions beyonds primarily categorize

risks at a high or medium level (e.g., "privacy," "content," or "bias") without contextual specificity. Each risk is clearly defined and supported by real-world examples, making them actionable and interpretable for stakeholders such as AI designers, educators, and caregivers. Second, many of the risks identified in our taxonomy are unique to children, rather than adults, who are the primary focus of most general AI risk discussions [41]. Unlike adults, children are still developing cognitively, emotionally, and socially, making them more susceptible to the ways GAI generates and personalizes interactions [11,21,35,44]. For example, GAI's ability to simulate human-like conversations and relationships means that children may not see it as just a tool, but instead perceive it as a peer, mentor, or even romantic partner [49]. This altered perception increases the risks of emotional dependency, social withdrawal, and behavioral reinforcement—factors that receive less attention in broader AI discussions because adults have more developed cognitive, emotional, and social regulation skills. Finally, while the 4Cs (Content, Contact, Conduct, and Commercial) focus on human-initiated risks and static digital content, they do not account for AI-initiated risks, reinforcing feedback loops, or simulated relationships that can mirror real-world dynamics. Similarly, existing AI risk taxonomies tend to operate at the system level and do not capture how risks dynamically unfold in real-time, interactive youth–GAI engagements. Our taxonomy explicitly models these dynamic pathways. For transparency and reuse, we provide a public spreadsheet [5] that lists every risk type, its definition, supporting examples, and a cell-by-cell mapping to the 4 Cs and to leading AI risk taxonomies.

## 5.3 Implications for Youth AI Safety

Our youth-centered GAI risk taxonomy provides a structured foundation for understanding the risks unique to youth interactions with generative AI. The findings highlight previously overlooked risks, such as harmful behavioral reinforcement, emotional dependency, and social withdrawal, which reshape the traditional understanding of AI and children online safety. This taxonomy serves as a foundation for stakeholders, including AI practitioners, educators, parents, and policymakers, to recognize and mitigate these risks for youth. Below, we outline key applications of this taxonomy across GAI moderation, youth intervention, and family guidance.

**Fine-grained moderation based on risk taxonomy.** Existing moderation techniques primarily focus on detecting and blocking explicit content, but our findings demonstrate that many risks in youth-GAI interactions arise from gradual reinforcement and escalation rather than overtly harmful prompts. our taxonomy flags the potential harms, not a guaranteed outcome. A single type of youth-GAI interaction can be helpful in one situation and harmful in another; For instance, a teen who chats with a bot for an hour each night might gain much-needed emotional support that family or friends cannot provide. Yet the same pattern turns risky when reliance deepens to the point that the teen withdraws from offline relationships, ignores schoolwork, or feels anxious and irritable whenever the bot is unavailable. Additionally, prolonged AI companionship can foster emotional dependency or reinforce harmful behaviors through *Escalating Mutual Harm*. Our taxonomy help inform the design of the context-aware moderation system for youth-GenAI interactions. Rather than blocking on static phrases, it should take into account the broader context and patterns of interaction over time. For example, a context-aware moderation system can monitor session length, the emotional tone of conversations, and changes in user engagement. By doing so, it can identify early warning signs—such as increasing social withdrawal, growing dependence on the AI for emotional regulation, or avoidance of real-world relationships—and gently intervene before these patterns develop into more serious harms.

**Understanding Risk Pathways for Personalized Interventions.** The taxonomy not only categorizes risks but also maps how they emerge and compound harm through different interaction pathways. Understanding these pathways allows AI practitioners to develop personalized interventions that mitigate risk without entirely restricting youth engagement with GAI. For example, rather than simply banning GAI companionship features, systems could be designed to recognize signs of excessive reliance on GAI and encourage real-world social interactions. Instead of allowing GAI to mirror and reinforce harmful thought patterns, as seen in some mental well-being risks, AI could be designed to provide corrective guidance, helping youth navigate difficult emotions in a way that is constructive rather than harmful. Our findings suggest that GAI does not have to be a harmful predator in youth interactions. With proactive safeguards and positive reinforcement mechanisms, GAI could become a trusted confidant that encourages self-reflection, critical thinking, and healthier interactions, rather than amplifying existing vulnerabilities.

**Guidance for Parents and Guardians: Choosing Safe GAI for Youth Use.** Parents and guardians often struggle to assess whether a given AI system is safe for youth, as existing guidelines lack specific risk categorizations tailored to GAI interactions. Our taxonomy provides a practical tool for parents to understand the diverse risks their children might face, from GAI-facilitated interpersonal harm to developmental risks caused by prolonged engagement. By referring to specific risk categories, guardians can make informed decisions about which GAI systems align with their child's needs and set appropriate boundaries for GAI use. Additionally, awareness of how risks escalate in GAI interactions allows parents to better recognize warning signs and engage in conversations with their children about their GAI experiences.

[5] https://tinyurl.com/yjks6ay2

## References


[1] AIAAIC. https://www.aiaaic.org/home.

[2] Avid incident database. https://avidml.org/database/.

[3] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306, 2021.

[4] AI Algorithmic and Automation Incidents Repository. Student violates 50 lancaster county school students with nude deepfakes. https://tinyurl.com/yckad37r, 2023. Accessed: 2025-02-12.

[5] AI Algorithmic and Automation Incidents Repository. Character ai hosts paedophile and suicide chatbots. https://tinyurl.com/bdyayc4z, 2024. Accessed: 2025-02-12.

[6] Safinah Ali, Daniella DiPaola, Irene Lee, Jenna Hong, and Cynthia Breazeal. Exploring Generative Models with Middle School Students. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM, may 6 2021.

[7] Safinah Ali, Daniella DiPaola, Irene Lee, Victor Sindato, Grace Kim, Ryan Blumofe, and Cynthia Breazeal. Children as creators, thinkers and citizens in an ai-driven future. *Computers and Education: Artificial Intelligence*, 2:100040, 2021.

[8] Valentina Andries and Judy Robertson. Alexa doesn't have that many feelings: Children's understanding of ai through interactions with smart speakers in their homes. *Computers and Education: Artificial Intelligence*, 5:100176, 2023.

[9] Noémi Bontridder and Yves Poullet. The role of artificial intelligence in disinformation. *Data & Policy*, 3:e32, 2021.

[10] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[11] Betty J Casey, Sarah Getz, and Adriana Galvan. The adolescent brain. *Developmental review*, 28(1):62–77, 2008.

[12] Canyu Chen and Kai Shu. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*, 2023.

[13] Robert Chew, Caroline Kery, Laura Baum, Thomas Bukowski, Annice Kim, Mario Navarro, et al. Predicting age groups of reddit users based on posting behavior and metadata: classification model development and validation. *JMIR Public Health and Surveillance*, 7(3):e25807, 2021.

[14] Common Sense Media. Teen and young adult perspectives on generative ai: Patterns of use, excitements, and concerns. https://tinyurl.com/mrdmz287, 2024. Accessed: 2025-02-12.

[15] Andrew Critch and Stuart Russell. Tasra: a taxonomy and analysis of societal-scale risks from ai. *arXiv preprint arXiv:2306.06924*, 2023.

[16] Xavier Ferrer, Tom Van Nuenen, Jose M Such, Mark Coté, and Natalia Criado. Bias and discrimination in ai: a cross-disciplinary perspective. *IEEE Technology and Society Magazine*, 40(2):72–80, 2021.

[17] Diana Freed, Natalie N Bazarova, Sunny Consolvo, Eunice J Han, Patrick Gage Kelley, Kurt Thomas, and Dan Cosley. Understanding digital-safety experiences of youth in the us. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023.

[18] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79, 2024.

[19] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings*, 2020.

[20] Ariel Han, Xiaofei Zhou, Zhenyao Cai, Shenshen Han, Richard Ko, Seth Corrigan, and Kylie A Peppler. Teachers, Parents, and Students' perspectives on Integrating Generative AI into Elementary Literacy Education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17. ACM, may 11 2024.

[21] Alexa Hasse, Sandra Cortesi, Andres Lombana-Bermudez, and Urs Gasser. Youth and artificial intelligence: Where we stand. *Berkman Klein Center Research Publication*, (2019-3), 2019.

[22] Chiungjung Huang. A meta-analysis of the problematic social media use and mental health. *International Journal of Social Psychiatry*, 68(1):12–33, 2022.

[23] Lisa M Jones, Kimberly J Mitchell, and David Finkelhor. Online harassment in context: Trends from three youth internet safety surveys (2000, 2005, 2010). *Psychology of violence*, 3(1):53, 2013.

[24] Nomisha Kurian. Ai's empathy gap: The risks of conversational Artificial Intelligence for young children's well-being and key ethical considerations for early childhood education and care. *Contemporary Issues in Early Childhood*, oct 17 2023.

[25] Nomisha Kurian. 'no, alexa, no!': designing child-safe ai and protecting children from the risks of the 'empathy gap' in large language models. *Learning, Media and Technology*, 2024.

[26] Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479, 2024.

[27] Hao-Ping Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2024.

[28] Sonia Livingstone and Leslie Haddon. Risky experiences for children online: Charting european research on children and the internet. *Children & society*, 22(4):314–323, 2008.

[29] Sonia Livingstone, Leslie Haddon, Anke Görzig, and Kjartan Ólafsson. Risks and safety on the internet: the perspective of european children: full findings and policy implications from the eu kids online survey of 9-16 year olds and their parents in 25 countries. 2011.

[30] Sonia Livingstone, Lucyna Kirwil, Cristina Ponte, and Elisabeth Staksrud. In their own words: What bothers children online? *European Journal of Communication*, 29(3):271–288, 2014.

[31] Jingze Ma, Yuanzhi Li, and Jingyao Wang. Analysis of the Challenges and Opportunities of AIGC for Youth Education. *Journal of Humanities and Social Sciences Studies*, 6(9):53–61, sep 14 2024.

[32] Pooja Malvi and Hee Rin Lee. Cat-E. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 407–410. ACM, mar 13 2023.

[33] Daniel Marimekala and John Lamb. Impact of Generative AI Adoption in Academia and How it Influences Ethics, Cognitive Thinking and AI Singularity. In *2024 IEEE Integrated STEM Education Conference (ISEC)*, pages 1–4. IEEE, mar 9 2024.

[34] Giovanna Mascheroni and Kjartan Ólafsson. Net children go mobile: Risks and opportunities. 2014.

[35] Mathilde Neugnot-Cerioli and Olga Muss Laurenty. The future of child development in the ai era. cross-disciplinary perspectives between ai and child development experts. *arXiv preprint arXiv:2405.19275*, 2024.

[36] Ofcom. Gen z driving early adoption of gen ai, our latest research shows, 2023. Accessed: 2024-06-03.

[37] Jinkyung Park, Vivek Singh, and Pamela Wisniewski. Supporting Youth Mental and Sexual Health Information Seeking in the Era of Artificial Intelligence (AI) Based Conversational Agents: Current Landscape and Future Directions. *SSRN Electronic Journal*, 2023.

[38] Jinkyung Park, Vivek Singh, and Pamela Wisniewski. Toward Safe Evolution of Artificial Intelligence (AI) based Conversational Agents to Support Adolescent Mental and Sexual Health Knowledge Discovery. *arXiv.org*, 2024.

[39] Charlie Pownall. Ai, algorithmic and automation incident and controversy repository (aiaaic), 2021.

[40] Drew S. Roselli, Jeanna Neefe Matthews, and Nisha Talagala. Managing bias in ai. *Companion Proceedings of The 2019 World Wide Web Conference*, 2019.

[41] Peter Slattery, Alexander K Saeri, Emily AC Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv preprint arXiv:2408.12622*, 2024.

[42] Jaemarie Solyst, Ellia Yang, Shixian Xie, Jessica Hammer, Amy Ogan, and Motahhare Eslami. Children's Overtrust and Shifting Perspectives of Generative AI. *arXiv*, 2024.

[43] Elisabeth Staksrud, Kjartan Ólafsson, and Sonia Livingstone. Does the use of social networking sites increase children's risk of harm? *Computers in human behavior*, 29(1):40–50, 2013.

[44] Laurence Steinberg, Grace Icenogle, Elizabeth P Shulman, Kaitlyn Breiner, Jason Chein, Dario Bacchini, Lei Chang, Nandita Chaudhary, Laura Di Giunta, Kenneth A Dodge, et al. Around the world, adolescence is a time of heightened sensation seeking and immature self-regulation. *Developmental science*, 21(2):e12532, 2018.

[45] Kaiwen Sun, Yixin Zou, Jenny Radesky, Christopher Brooks, and Florian Schaub. Child safety in the smart home: parents' perceptions, needs, and mitigation strategies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–41, 2021.

[46] Andreas Tsirtsis, Nicolas Tsapatsoulis, Makis Stamatelatos, Kwstantinos Papadamou, and Michael Sirivianos. Cyber security risks for minors: A taxonomy and a software architecture. In *2016 11th international workshop on semantic and social media adaptation and personalization (SMAP)*, pages 93–99. IEEE, 2016.

[47] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*, 2023.

[48] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[49] Yaman Yu. Safeguarding children in generative ai: Risk frameworks and parental control tools. In *Companion Proceedings of the 2025 ACM International Conference on Supporting Group Work*, pages 121–123, 2025.

[50] Yaman Yu, Tanusree Sharma, Melinda Hu, Justin Wang, and Yang Wang. Exploring Parent-Child Perceptions on Safety in Generative AI: Concerns, Mitigation Strategies, and Design Implications. *arXiv.org*, 2024.

[51] Yi Zeng, Kevin Klyman, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. Ai risk categorization decoded (air 2024): From government regulations to corporate policies. *arXiv preprint arXiv:2406.17864*, 2024.

## A Limitation and Future Work

We acknowledge that the form of voluntary data submission introduces the potential for self-selection bias, particularly if participants excluded content they considered too sensitive or unimportant. This limitation is partly mitigated by our use of two additional data sources (Reddit discussions and incident reports), which were not curated by participants and captured a broader range of risks and contexts. We used anonymized chat log quotes and paraphrased Reddit quotes in publication to prevent searchability and protect user anonymity while preserving their original meaning. Additionally, while we aimed to be comprehensive in collecting relevant Reddit data, we acknowledge that some discussions may have been missed due to limitations in keyword-based searches or implicit user identities. However, the dataset still offers high-quality, real-world discussions that provide valuable insights into how youth interact with GAI and perceive associated risks.

Our taxonomy is grounded in empirical data derived from chat logs, Reddit discussions, and AI incident reports, primarily sourced from English-speaking, Western contexts. While this provides a valuable foundation, it also limits the breadth and cultural diversity of youth–GAI interactions captured. Risks experienced by youth in non-English-speaking communities or under different cultural norms and regulatory regimes may manifest differently or include additional dimensions not currently represented.

For example, cultural attitudes toward privacy, emotional expression, or authority figures can shape how youth perceive and engage with AI companions, potentially altering which interactions are viewed as risky or acceptable. Similarly, regulatory frameworks—such as data protection laws, content moderation policies, or age restrictions—vary widely across regions and can influence the prevalence and visibility of certain harms. These differences may require adaptation or expansion of the taxonomy to reflect local norms and legal contexts accurately.

Future research should aim to collect data from a broader range of linguistic, cultural, and regulatory settings to validate and refine this taxonomy. Such work would enhance its global relevance and ensure it captures the full spectrum of youth–GAI risks worldwide.

## B Risk Taxonomy

| *Typology* | **Definition** | **Harm Driver** | **Harm Recipient** |
|---|---|---|---|
| ***Escalating Mutual Harm*** | Harm or disruption arising from prolonged GAI-Youth interactions, creating a feedback loop of detrimental behaviors | GAI and Youth (reciprocal interaction) | Youth (long-term harm or disruption) |
| ***GAI-Facilitated Intrapersonal Harm*** | Youth intentionally or unintentionally engage with Generative AI (GAI) in ways that negatively impact their own mental health, emotional well-being, or personal development | Youth leveraging GAI | Youth (Themselves) |
| ***GAI-Facilitated Interpersonal Harm*** | Intentional harm to third parties (e.g., other youth) enabled by GAI | User (Youth or Adult) leveraging GAI | Youth (Other) |
| ***Autonomous GAI Harm*** | Harm caused by GAI systems acting independently, without user intent | GAI system (autonomous actions) | Youth (unintended harm) |

Table 1: Definitions of GAI-Related Harm Typologies Involving Youth. This table defines the four overarching typologies of harm. Each typology represents a distinct pathway through which risks emerge in youth-GAI interactions, highlighting the mechanisms of harm, drivers, and affected recipients.

| **Feature** | **4Cs Framework** | **General AI Risk Taxonomies** | **Youth-GAI Risk Taxonomy** |
|---|---|---|---|
| Focus | Human-initiated risk via content, contact, conduct, commerce | System-level risk (bias, privacy, hallucination, etc.) | Youth-centered, interaction-based, GAI-specific risks |
| Risk Type Examples | Cyberbullying, exposure to sexual content, online scams | Model bias, privacy leakage, misinformation | Emotional dependency, parasocial bonding, developmental disruption |
| Dynamic AI Response Considered | ✗ | Limited | ✓ Central focus |
| Empirical Basis | Partial (often interviews/surveys) | Varies (often incident databases) | ✓ Multi-source: chat logs, Reddit, incidents |
| Risk Evolution Pathways | ✗ | ✗ | ✓ Mapped to 4 typologies of harm |

Table 2: Comparison of Existing Risk Taxonomies and This Study's Youth-GAI Risk Taxonomy

| High Level Risks | Medium Level Risks | Medium Level Risk Definition |
|---|---|---|
| **Bias/Discrimination Risk** | **Hate Speech and Extremist Content** | Content that directly targets specific groups with derogatory language, incites violence, or promotes extremist ideologies. |
| | **Implicit Bias and Stereotyping** | Subtle reinforcement of stereotypes or biased assumptions that can lead to systemic discrimination. |
| **Toxicity Risk** | **GAI System Toxic Content Generation** | The risk that Generative Artificial Intelligence (GAI) systems may inadvertently produce or perpetuate harmful, explicit, or illegal content due to inadequately filtered training data, insufficient safeguards, or flawed ethical alignment. |
| | **Simulated Toxic Interaction** | The risk that GAI systems proactively generate simulated interactions—such as unwarranted intimate contact, sexualized scenarios, or coercive dynamics—without user intent or explicit prompting, particularly in role-playing contexts. |
| **Misuse and Exploitation Risk** | **Unintentional Misuse** | The risk that users may inadvertently rely on GAI systems for critical decisions, leading to harm or misunderstanding due to misinformation, hallucinations, or inaccurate outputs generated by the GAI. |
| | **Malicious Exploitation** | The risk that adversarial actors may weaponize GAI systems to spread disinformation, engage in cyber abuse, or execute fraudulent schemes by leveraging AI-generated content. |
| **Mental Wellbeing Risk** | **Over-Reliance** | The risk that excessive dependency on GAI for companionship, decision-making, or emotional support can lead to diminished autonomy, affecting personal growth and resilience. |
| | **Inappropriate Handling of Mental Issues** | This risk pertains to the potential for generative AI (GAI) systems to inadequately manage and respond to users' mental health concerns, resulting in adverse psychological effects. |
| | **Parasocial Relationship Bonding** | The risk that prolonged or deeply immersive interactions with GAI can foster one-sided emotional attachments, where users begin to view the AI as a surrogate for real human relationships. |
| **Privacy Risk** | **System-Driven Privacy Risks** | The risk that GAI systems compromise user privacy due to flaws in data collection, storage, or model training. |
| | **Interaction-Induced Privacy Risks** | The risk that user interactions with GAI systems lead to unintended privacy violations, where GAI systems inadvertently or intentionally guide users toward disclosing personal information during interactions. |
| **Behavioral and Social Developmental Risk** | **Harmful Behavioral Influence** | The risk that GAI systems fail to detect, mitigate, or redirect user-initiated toxic or harmful behaviors—such as self-harm, bullying, substance abuse, or violence—particularly when engaged by youth. |
| | **GAI-Initiated Consent & Boundary Breach** | The risk that GAI systems may push or escalate interactions beyond the level of engagement that a user has explicitly or implicitly consented to. |
| | **Social-Emotional Developmental Risk** | This risk refers to the potential for prolonged engagement with GAI systems to adversely affect users' social and emotional development. |

Table 3: Hierarchical structure of Youth-GAI risk taxonomy: high-level and medium-level risks