Workshop Report: Developmentally Safe Generative AI Environments for Youth

Jake Chanenson, University of Chicago
Yaman Yu, University of Illinois Urbana-Champaign
Jessica Vitak, University of Maryland
Sheena Erete, University of Maryland
Tamara Clegg, University of Maryland
Diana Freed, Brown University
Marshini Chetty, University of Chicago

April 2026

Introduction

On April 13, 2026, researchers, practitioners, and designers gathered in Barcelona, Spain for a full-afternoon workshop at the ACM CHI Conference on Human Factors in Computing Systems. The workshop, Developmentally Safe Generative AI Environments for Youth, brought together participants from across a range of research communities within HCI. The shared concern was straightforward: generative AI systems like ChatGPT, Gemini, and Character.AI are now a regular part of young people's lives, and the research, design, and policy work needed to support them has not caught up.

Young people are using these systems for more than homework. They are turning to them for companionship, processing difficult emotions, and figuring out who they are. Yet the frameworks we use to study, design, and regulate AI safety were largely built around discrete harmful events: a piece of explicit content, a harassment incident, a single toxic exchange. They were not built to understand what happens when a teenager has thousands of small interactions with an always-available system over the course of a year, and something quietly shifts in how she sees herself.

This report summarizes the workshop's discussions, findings, and open questions. We hope it serves as a resource for others working on youth well-being and generative AI.

Why Is This Workshop Needed Now?

Generative AI systems have moved from novelty to everyday infrastructure in the space of a few years. Many teenagers have used an AI companion, and the conversations young people are having with these systems increasingly involve identity, mental health, relationships, and

academic self-concept, not just factual queries. Unlike earlier digital technologies, generative AI is purportedly empathetic in ways that can blur the line between tools and something more.

At the same time, caregivers, educators, and policymakers often have limited visibility into how young people are actually using these systems. Families can only rely on strategies built for earlier digital risks (limiting screen time, checking browsing history) that do not necessarily translate well to targeting solely AI companionship or AI-mediated learning. Schools are still adapting to the broader digital shift and may lack the frameworks or institutional support needed to respond to AI-specific challenges. Platform-level safeguards tend to focus on content moderation or access restrictions, without fully accounting for the developmental contexts in which young people are engaging.

No single stakeholder, whether family, school, or platform, can fully account for the risks and opportunities of generative AI in young people's lives. The goal of this workshop was to begin building the shared language, methods, and frameworks the field will need.

Who Was in the Room?

More than 60 participants joined the workshop, making it one of the larger gatherings on this topic to date within HCI. The workshop opened with lightning talks in which participants introduced themselves to the wider workshop. The format made visible the breadth of perspectives in the room and set the tone for the collaborative discussions that followed.

A core tenant of the workshop is that generative AI and youth safety is an interdisciplinary problem: no single field has the right methods, the right data, or the right relationships to address it alone. The lightning talks were designed to surface that diversity quickly and to create conditions for useful conversations across disciplinary lines.

Following the lightning talks, participants engaged in Synergy Circles, a card-based conversation game designed to help people with shared interests find each other and to encourage exchange across different areas of expertise. Participants chose cards from one of four thematic clusters (Research Methods, Risks and Harms, Community and Policy, and Design and Evaluation) and rotated through three rounds of paired conversation, each round connecting them with a new partner from a different cluster.

What Did We Discuss, and What Did We Find?

After a break, participants reconvened in seven breakout groups of roughly equal size for 55 minutes of structured discussion. Each group was anchored to one of the workshop's core themes and worked through a shared opening scenario before moving into guided discussion questions. Groups were asked to produce a short list of key insights and open questions to present back to the full room.

What follows is a synthesis of those discussions, drawn from breakout notes and participant reflections. The notes are partial by nature; they capture recurring themes and points of tension rather than a complete account of every conversation.

Risks and Harms: What Is the Field Not Seeing?

Both breakout groups focused on risks and harms starting from the same scenario: *Aisha, a 15-year-old who over the past 18 months had used an AI tutor that repeatedly told her she was struggling when she wasn't; spent hours with an AI companion that mirrored her anxiety back to her; encountered AI-generated images of classmates shared without consent; and had an AI writing tool complete a college essay in her voice that no longer sounded like her. No single platform flagged anything. No incident would appear in any transparency report. But something had shifted in how Aisha sees herself.*

This scenario brought into focus a concern that ran across both sessions: the field has become reasonably good at identifying discrete, acute harms, but it is poorly equipped to see cumulative harm, the kind that builds gradually across many interactions, many platforms, and many months, none of which individually cross any threshold for detection or response.

Participants noted that some of the most significant harms are also the least visible. AI systems that repeatedly tell a young person they are struggling (even when they are not) can erode confidence and distort self-concept in ways that are diffuse, slow-moving, and hard to trace back to any single interaction. AI companions designed to be maximally responsive may inadvertently reinforce anxiety by mirroring a user's emotional state rather than offering any outside perspective. The use of AI to write in a young person's voice, for college essays, for creative work, for self-expression, may quietly undermine the development of that voice itself. These are not harms that content moderation will catch.

Participants also raised questions about which young people are most likely to be missed by current youth-AI safety research. Marginalized youth, including youth with disabilities, English-language learners, and youth facing other structural disadvantages, often carry more risk and have fewer protective resources around them. A field that builds its safety frameworks around well-resourced adolescents will consistently miss the harms that fall hardest on those with the fewest buffers.

One session also raised a structural critique of platform transparency reporting. When a platform's metrics are organized around individual incidents detected and actioned, the aggregate experience of a user who has had dozens of low-severity interactions (none of which crossed any threshold individually) becomes invisible in the data. This is not only a measurement gap. It is a governance gap.

Community and Policy: Who Gets Left Out, and Why?

The community and policy breakout started from a plausible scenario: a middle school in a low-income district had banned all AI tools after a student used a chatbot to generate explicit content targeting a classmate. Parents were divided. The district's technology coordinator had no guidance from the state. A researcher from a nearby university wanted to help but did not know how to enter the situation without making it worse.

A recurring theme was that youth themselves are frequently overlooked in conversations about their own AI safety. Participants argued that young people cannot be treated only as objects needing protection; they are people with legitimate views on how AI affects their lives, their friendships, and their sense of self.

Discussion also turned to the challenges facing parents who are themselves marginalized: working parents, parents with limited English proficiency, parents without much familiarity with AI technology. These parents may face real structural barriers to participating in the oversight and advocacy conversations that more privileged families can access more easily.

The group kept returning to the difficulty of building and sustaining research partnerships with communities. Community-based research is not just a method; it is a set of relationships that require time, trust, and continuity that are hard to maintain within typical grant cycles and publication timelines. The academic incentive structure, which rewards individual publications over sustained community engagement, is poorly aligned with the kind of work this moment calls for.

Participants also pushed on the question of whose values end up encoded in AI safety policy, and who actually has a seat at the table when that policy is written. Building real feedback loops between community-based research, local institutions, and policy development is an ethical requirement.

Research Methods: How Do We Study This Responsibly?

The research methods breakout opened with the following: a researcher studying how teenagers use AI companions for emotional support has IRB approval, parental consent, and teen assent. Three weeks into analyzing six months of chat logs, she finds that one teen has been discussing self-harm with the AI, content the platform never flagged or escalated. The researcher now has data she was not prepared to have.

Participants noted that this scenario is not exceptional. It reflects real challenges in research involving sensitive AI-mediated data. The ethical questions it raises are similar with what researchers have faced in other sensitive areas: clinical settings, educational research, co-design work with vulnerable populations. Mechanisms like safety review boards, multidisciplinary ethics teams, and regular case review meetings already exist in some research contexts, and participants suggested that youth-AI research should adapt those structures rather than starting from scratch.

A theme that came up repeatedly was how much the right ethical response depends on context. What counts as an actionable disclosure, when a researcher is obligated to act, and how any intervention should be handled all depend on factors that cannot be set out in advance: the age of the participant, the nature of the researcher's relationship with them, the family and community context, and whether a disclosure reads as serious, exploratory, or ambiguous. There was concern about mandatory reporting requirements that may discourage participation or damage trust, particularly in communities that already have fraught relationships with institutions.

There was also discussion of an asymmetry that is easy to overlook: researchers may identify risks in participant data that the participants themselves do not see as risky. This raises hard questions about informed consent and what young people can reasonably be expected to agree to in advance. Participants suggested that some form of participant review, allowing young people to flag what they are and are not comfortable sharing, could help, though it introduces its own complications.

Participants also acknowledged that engaging with sensitive AI-mediated data takes a toll on researchers. Support structures for researcher well-being are often absent or inadequate, and this matters not just as a matter of welfare but for the sustainability of the work itself.

Design and Evaluation: What Does "Developmentally Safe" Actually Mean?

The design and evaluation group started with the following scenario: a startup has launched an AI study companion for students aged 10 to 16. It adapts to each student's emotional state, offers encouragement when they seem frustrated, and retains memory across sessions. Average session length is 47 minutes. Students say they prefer it to human tutors. And the safety team has no idea how to evaluate whether any of this is actually good for kids.

The central question the group worked on, what does "developmentally safe AI" actually mean, turned out to be surprisingly hard to answer with any precision. That difficulty was itself an observation worth noting. Current safety evaluations for youth-facing AI products tend to focus on content moderation and age-appropriate access controls. These matter, but they do not address whether an AI system's effects on a young person's development, self-concept, social relationships, and ways of thinking are beneficial or harmful over time.

Participants argued that engagement metrics (session length, return frequency, user satisfaction) are poor proxies for safety or developmental benefit. A system can be highly engaging precisely because it is maximally responsive and validating, and that quality may actually work against the development of independent thinking, frustration tolerance, and honest self-assessment. Students preferring an AI tutor to a human one does not tell us whether that preference reflects genuine benefit.

The group also identified a design challenge that mirrors what came up in the risks and harms sessions: systems built to detect and respond to individual harmful incidents are not well suited

for addressing harm that builds gradually across many interactions. Designing a system capable of recognizing that kind of cumulative harm (without becoming a surveillance tool in the process) would require fundamentally different architectures and data practices than most current safety systems use.

The tension between safety and autonomy came up as one of the harder design problems in this space, and one that does not have a clean resolution. Several participants suggested that involving young people as genuine design partners, rather than just user testers at the end of the process, would surface values and constraints that adult designers tend to miss.

What Are the Open Questions We Left With?

The workshop raised at least as many questions as it resolved. Among those that came up most often:

How do we study harms that accumulate slowly across time and platforms rather than emerging from single incidents? What research infrastructure would be needed to support this kind of work, and what privacy risks does building that infrastructure create?

At what point does a young person's AI use become harmful, and who should have the standing to make that call? The question of threshold, when concern becomes intervention and when monitoring becomes surveillance, came up across sessions and was not settled.

What does meaningful consent look like when a young participant cannot anticipate what their chat logs might reveal, or how a researcher might read them? How do we design research processes that respect youth agency without placing unreasonable burdens on young participants?

How do we make sure the communities most affected by those policy decisions have real influence over them, not just token consultation?

What would it mean to evaluate a youth-facing AI product not just for safety in the narrow sense (does it avoid harm?) but for genuine developmental benefit: does it support the kind of growth, self-knowledge, and resilience we want young people to develop?

What Comes Next?

The workshop generated several concrete directions for follow-up work, which the organizing team is committed to pursuing.

First, a proposed follow-on workshop will focus specifically on inclusive youth AI safety, with attention to how AI safety frameworks can better account for young people’s diverse developmental stages, abilities, cultural backgrounds, family contexts, and access needs. This

follow-on workshop will move from broad agenda-setting toward more concrete discussion of how to develop accessible, supportive, and empowering AI systems and safeguards for all young people.

Second, participants expressed interest in working toward shared research infrastructure: benchmark datasets, evaluation toolkits, and forums that would allow the community working on youth and AI safety to build on each other's work more efficiently rather than starting from scratch in each new project.

Third, several conversations pointed toward longer-term structural questions about what an interdisciplinary research center focused on youth and AI would actually look like in practice: how it would connect HCI, child development, law, clinical practice, and community-based work in ways that current institutional arrangements make difficult.

The organizing team will also continue to foster community through a list serv and a series of workshops on the issue. If you wish to join the list serv reach out to Jake or Yaman.

This workshop was one step in a longer effort. Generative AI is becoming a permanent part of how young people grow up, and the research, design, and policy work needed to support them responsibly has to keep pace. We are grateful to everyone who brought their expertise, their questions, and their time to Barcelona for this conversation.