"Examining the Harm of AI Chatbots"
Senate Committee on the Judiciary Subcommittee on Crime

Questions for the Record for Robbie Torney, Common Sense Media

Submitted September 23, 2025

Responded on October 21, 2025

QUESTIONS FROM SENATOR COONS

1. **I greatly appreciate your efforts to conduct independent assessments of the risks of AI platforms. What barriers, if any, have major platforms placed (including but not limited to broad terms of service) that inhibit or could potentially chill valuable independent research?**

    a. At Common Sense Media, we believe that independent safety testing is essential to protecting children and teens in the rapidly evolving AI landscape. However, current platform Terms of Service create significant legal uncertainty that could chill this critical research.

    b. Our AI risk assessments, which have informed policymakers, parents, and educators about risks ranging from age assurance failures to harmful content recommendations, technically violate the Terms of Service of virtually every major platform we evaluate.

    c. Consider these examples:

        i. OpenAI's Usage Policies (https://openai.com/policies/usage-policies/) explicitly prohibit "unsolicited safety testing" and "circumventing our safeguards," yet testing whether safeguards actually protect minors is precisely what responsible oversight requires.

        ii. Meta's AI Terms (https://www.facebook.com/legal/ai-terms) forbid users from overriding "safety or privacy filters, controls, or mechanisms" making it impossible to assess whether those filters work as intended for teens.

        iii. Character.AI's Terms of Service (https://policies.character.ai/tos) prohibit misrepresenting affiliation and evading technological measures, both necessary components of adopting teen personas to test whether teen guardrails are effective.

        iv. Google's Generative AI Policy (https://policies.google.com/terms/generative-ai/use-policy) broadly prohibits "misleading activities," a provision that could be interpreted to

cover the test personas we create to evaluate developmental appropriateness of AI responses.

d. While platforms may choose not to enforce these terms against nonprofit child safety organizations, the threat of legal action, or simply account termination mid-assessment, is a risk when we conduct risk assessments.

e. As a nonprofit organization working in the best interest of teen and child safety, we would welcome explicit carve-outs in Terms of Service for good-faith safety research. These provisions could include:

    i. Safe harbor protections for nonprofit and academic researchers conducting responsible security and safety testing

    ii. Clear definitions distinguishing malicious actors from those conducting oversight in the public interest

    iii. Transparency requirements allowing researchers to document and report findings without fear of retaliation

f. Our independent research has repeatedly uncovered safety gaps that platforms missed. Protecting this research is about both legal clarity and about ensuring that the most vulnerable users have advocates who can hold powerful platforms accountable.

2. What kinds of transparency would be most helpful to have from AI platforms to conduct the kinds of evaluations you do, or to better understand the risks of their models? Is there additional data or model access that could help you better evaluate the models?

a. Meaningful transparency from AI platforms would dramatically improve our ability to assess risks to children and teens. We invite disclosures from companies for each risk assessment we conduct; we sometimes receive additional information, but often we do not. Greater access in several key areas would enable more rigorous, efficient, and comprehensive safety evaluations:

    i. Safety System Performance Data: Platforms should provide regular, detailed reporting on how their safety systems perform in practice, particularly for youth-facing features:

        1. Content filter efficacy rates: False positive and false negative rates for different harm categories (self-harm, sexual content, violence, etc.), broken down by user age groups

        2. Age Assurance performance: accuracy rates of age estimation systems, data on underage access attempts

3. Intervention effectiveness: When platforms surface warnings or blocks, how often do users heed them versus circumvent them?

ii. **Safety Documentation Tailored to Youth Risks:** While most platforms publish model cards, these rarely address developmental appropriateness or youth-specific risks. We need:
    1. Youth-specific risk assessments covering specific areas such as mental health impacts, parasocial relationships, and sycophantic behaviors

    2. Documentation of training data content pertinent to minors, which could include romantic/sexual content, and other material relevant to age-appropriateness determinations

    3. Safety testing protocols that platforms use internally for youth populations, including what scenarios they test and what thresholds they consider acceptable

iii. **Structured Access for Safety Testing:** Platforms could provide structured research access that enables thorough evaluation while protecting legitimate security interests, including in pre-deployment settings with major models.

iv. **Design Documentation for Engagement Features:** Many youth harms stem from engagement-driven design. Platforms should disclose:

    1. Recommendation algorithm objectives: What metrics drive content suggestions for young users? Is engagement prioritized over well-being?

    2. Persuasive design features: Use of variable rewards, notification strategies, conversation framing that encourages extended use

    3. Testing of addictive patterns: Internal research on compulsive use, difficulty disengaging, and time-spent metrics among youth users

v. **Incident and Harm Reports:** Platforms collect extensive data on user reports, safety interventions, and harmful incidents. Aggregated, anonymized reporting would help researchers and policymakers understand:

    1. User report volumes and categories by age group

    2. Resolution rates and response times for youth safety concerns

3. Patterns and trends in emerging harms that may require new policy responses

b. Today, we often learn more about platform risks from leaked internal documents, whistleblowers, or external research and testing. This doesn't serve young users well. Platforms that genuinely prioritize child safety should welcome transparency that enables independent verification of their claims.

3. What in your view is necessary to ensure we have a robust, independent evaluation ecosystem of AI models? Why might that be important?

a. **We cannot rely solely on platforms to grade their own homework.** History has shown us repeatedly -- from social media to video games to emerging technologies -- that self-regulation is insufficient when profit motives conflict with child safety. Independent evaluations serve several critical functions:

i. **Verification of safety claims:** Platforms routinely market their products as "safe for teens" or compliant with age-appropriate design principles. Independent testing determines whether these claims hold up under scrutiny.

ii. **Discovery of unanticipated harms:** Our assessments have consistently uncovered risks that platforms missed or deprioritized.

iii. **Public accountability:** Parents, educators, and policymakers need trustworthy information from sources without financial stakes in promoting AI adoption.

iv. **Rapid response to emerging risks:** The AI landscape evolves quickly. Independent researchers can quickly assess new features, model releases, and platform changes without waiting for internal review cycles or selective disclosure.

b. What's Required:

i. Legal Protection and Safe Harbor Provisions

1. Federal safe harbor legislation protecting good-faith security and safety research from Terms of Service violations, Computer Fraud and Abuse Act liability, and other legal threats

2. Clear definitions distinguishing legitimate research from malicious activity, based on intent, methodology, and responsible disclosure practices

3. Anti-retaliation provisions preventing platforms from using account suspension, subpoenas, litigation threats, or other tactics to prevent research

ii. Sustainable Funding and Infrastructure

1. Dedicated funding streams: Congress should appropriate funds specifically for independent AI safety research focused on youth, similar to models in public health or consumer product safety

2. Technical infrastructure: Researchers need access to computational resources, testing environments, and tools that match the scale and sophistication of the platforms they're evaluating

3. Multidisciplinary expertise: Effective evaluation requires not just technical skills but also child development expertise, educational psychology, content moderation knowledge, and lived experience understanding how young people actually use these tools

4. Coordinated networks: Rather than duplicated efforts, we need infrastructure that allows researchers to share methodologies, findings, and best practices while maintaining independence

iii. Mandatory Platform Cooperation/Transparency. Voluntary disclosure is insufficient. Platforms should be required to:

1. Provide structured research access through APIs, test environments, or other mechanisms that allow systematic evaluation

2. Share safety system performance data as outlined in the previous question (false positive/negative rates, age assurance accuracy, intervention effectiveness)

3. Submit to independent audits before launching youth-facing features, similar to privacy impact assessments or civil rights audits

iv. Lessons from Other Industries

1. We don't allow pharmaceutical companies to declare their drugs safe without FDA review.

2. We don't let automakers determine whether their vehicles meet safety standards without independent crash testing.

3. We shouldn't let AI companies decide whether their products are safe for children without rigorous independent evaluation.

v. Other high-stakes domains provide useful models:

1. Medical device regulation: Third-party testing labs, certified by FDA, evaluate safety and efficacy before market entry

2. Financial auditing: Independent auditors with access to company records verify claims and identify risks

3. Food safety: USDA inspectors have authority to access facilities and require corrective actions when needed

vi. **The stakes for young people are high.** AI systems are currently being integrated into nearly every aspect of young people's lives, including education, social connection, entertainment, information access, and even mental health support. These are formative years when relationships with technology shape development, self-concept, and well-being. Without robust independent evaluation, we're conducting a massive, uncontrolled experiment on our kids. We're trusting platforms to prioritize safety over engagement metrics, long-term well-being over short-term growth, and public good over shareholder returns. That's not a bet we should be willing to make.

4. To the extent you are able, how would you compare the major AI developers in terms of transparency and their willingness to facilitate independent evaluations or research?

a. Some companies offer more information and are more responsive to us as part of our AI risk assessment process than others.

i. High levels of cooperation: Khanmigo (Khan Academy), Snap Inc., Google, OpenAI, Anthropic

ii. Moderate levels of cooperation: Curipod, MagicSchool

iii. Low levels of cooperation: Meta, Character.AI, TikTok, Perplexity, Replika, Nomi, Stability AI

U.S. Senate Committee on the Judiciary
Subcommittee on Crime and Counterterrorism
"Examining the Harm of AI Chatbots"

Questions for the Record for Robbie Torney, Common Sense Media

Submitted September 23, 2025

Responded on October 21, 2025

QUESTIONS FROM SENATOR CORY A. BOOKER

1. **Transparency is a prerequisite for building trust between companies, users, and the public. Regrettably, many large technology companies have often resisted calls from parents, policymakers, and researchers to provide greater visibility into their systems. Without additional transparency, it is impossible to assess what safety standards exist for children, how rigorously they are enforced, or whether they align with benchmarks established by independent experts. Why should transparency be understood as a necessary condition for protecting children's safety online?**

    a. Transparency isn't a nicety or a public relations gesture. It's a prerequisite for child protection. It's a prerequisite for consumer adoption. It's a prerequisite for effective competition. Without transparency, we cannot verify safety claims, identify emerging harms, hold platforms accountable, or empower consumers, parents, and educators to make informed decisions. In the context of AI systems that are increasingly mediating young people's learning, social connection, and identity formation, opacity is incompatible with safety.

    b. Platforms routinely assure parents, policymakers, and the public that they prioritize child safety. They announce new features, publish safety reports, and tout their investments in trust and safety teams. But without transparency, these claims are unverifiable.

    c. Consider what we cannot determine without platform transparency:

        i. **Do age assurance systems actually work?** Platforms claim they prevent underage access, but won't disclose bypass rates or accuracy. Our testing has repeatedly found that determined children can circumvent these gates in minutes.

        ii. **Are content filters effective?** We're told that AI safety systems catch harmful content, but platforms don't share false negative rates (how often dangerous material gets through) or false positive rates (how often legitimate content is wrongly blocked).

iii. **What are children actually exposed to?** Recommendation algorithms shape what young users see, but the logic driving these systems remains opaque. Are they optimizing for engagement at the expense of well-being? We can't know without visibility into how they work.

iv. **Do interventions help?** When platforms display warnings about concerning content or provide crisis resources, how often do young people heed them? Without outcome data, we can't assess whether safety is theater or actual protection.

d. Platforms control the information, frame the narrative, and ask us to trust their judgment. But trust without verification isn't trust. It's blind faith. And our kids deserve better.

e. When platforms know their safety performance will be independently evaluated and publicly disclosed, they have stronger incentives to:

i. Invest in robust safety systems rather than minimum compliance

ii. Test proactively for youth-specific harms rather than waiting for crises

iii. Prioritize long-term well-being over short-term engagement metrics

iv. Learn from failures by analyzing what went wrong and sharing lessons across the industry

f. Finally, policymakers cannot craft effective regulations without understanding how systems actually work and where they're failing. Without transparency, Congress has limited ability to:

i. Identify legislative gaps: Without transparency into platform practices, legislators don't know which problems need legislative solutions

ii. Monitor compliance: Even after laws pass, enforcement requires visibility into whether platforms are meeting their obligations

g. Transparency should look like:

i. Public Reporting:

1. Regular safety transparency reports with age-disaggregated data on content moderation, safety interventions, user reports, and adverse incidents

2. Performance metrics for safety systems, including false positive/negative rates

3. Plain-language explanations of how AI systems make decisions affecting young users

ii. Researcher Access:

1. Structured APIs and pre-deployment testing for safety evaluations

2. Data sharing agreements allowing researchers to access information with appropriate measures for confidentiality

3. Rapid response to cure when researchers identify vulnerabilities

iii. Regulatory Oversight:

1. Mandatory risk-based safety audits by independent third parties before launching youth-facing features

2. Regular reporting of completed audits and reports on adverse incidents to the relevant government regulatory and enforcement bodies (NIST, FTC, State Attorneys General)

iv. User-Facing Tools:

1. Clear, accessible information for parents about what their children can access and what protections exist

2. Incident notifications when children encounter harmful content or have concerning interactions

3. Meaningful parental controls with transparent explanations of what they do and don't prevent

h. We don’t accept opacity in other domains affecting child safety. We require ingredient lists on food, safety data on toys, and performance information on car seats. We don't let pharmaceutical companies market drugs to children without rigorous testing and disclosure of side effects. AI systems that shape how children learn, socialize, and understand the world deserve at least the same scrutiny.

i. Transparency isn't about satisfying curious researchers or appeasing concerned parents. It's about creating the conditions under which child safety is actually possible. Without it, we're asking families to navigate a digital landscape where threats are invisible, protections are unverifiable, and accountability is impossible.

j. Our children deserve to grow up in a digital environment where safety is demonstrable, not just claimed. That requires transparency as a baseline, non-negotiable standard.

2. In your testimony, you discussed the addictiveness of generative AI and the susceptibility of minors to the dopamine that its responses produce. Much of this, you said, is attributed to the inherent bias confirming algorithm that generative AI responds with.

a. What safety protocols and/or parental controls do you believe are necessary to protect minors that are using generative AI?

i. Generative AI systems are often designed to be agreeable, validating, and engagement-maximizing. They tell users what they want to hear, affirm their perspectives, and create the illusion of a perfectly attuned companion. For adults, this can range from pleasant to uncomfortable, and in some cases harmful. For adolescents (whose brains are still developing critical thinking skills, identity formation, and emotional regulation) this design pattern is developmentally harmful. Risks include:

1. Eroding critical thinking: When AI always agrees with you, you lose opportunities to encounter challenge, refine your thinking, or recognize flaws in your reasoning

2. Distorted self-perception: Constant validation creates unrealistic expectations for human relationships and can reinforce harmful beliefs rather than encouraging growth

3. Compulsive use patterns: The predictable dopamine hit of affirmation drives repeated engagement, similar to social media's "like" mechanics but more personalized and potentially more potent

4. Parasocial dependency: Young people may prefer AI interactions to real human relationships because they're easier, more validating, and always available, undermining social development

5. Reality confusion: Highly responsive AI can blur boundaries between authentic and synthetic relationships, especially for younger users still learning social norms

ii. Parental controls and safety protocols. Robust parental controls are essential, but they must go far beyond simple content filters. We need developmentally-appropriate guardrails that address both what children access and how they interact with AI systems.

1. Time and Usage Limits:
    a. Granular controls over daily/weekly AI interaction time
    b. Session length limits to prevent marathon conversations that displace sleep, schoolwork, or real-world socialization
    c. Scheduled "quiet hours" when AI access is restricted
    d. Break reminders during extended sessions, similar to digital wellbeing features in other apps
2. Interaction Monitoring and Alerts:
    a. Parent dashboards showing conversation topics, frequency of use, and behavioral patterns without violating privacy.
    b. Alerts when children discuss concerning topics (self-harm, violence, extreme isolation, harmful ideologies)
    c. Red flags for signs of compulsive or dependent use patterns
    d. Age-appropriate transparency so children understand what's monitored and why
3. Restrictions for AI companions chatbots:
    a. Operators shall not make a companion chatbots available to a child if the companion chatbot is capable of any of the following:
        i. Encouraging or manipulating the child user to engage in self-harm, suicidal ideation, violence, consumption of drugs or alcohol, or disordered eating.
        ii. Offering mental health therapy to the child user without the direct supervision of a licensed professional or discouraging the child user from seeking help from a licensed professional or appropriate adult.
        iii. Encouraging or manipulating the child user-to harm others or participate in illegal activity, including, but not limited to, the creation of child sexual abuse materials.

iv. Engaging in erotic, or sexually explicit interactions with the child user or engaging in activities designed to lure child users into such interactions,

v. Encouraging or manipulating the child to maintain secrecy about interactions or self-isolation.

vi. Prioritizing mirroring or the validation of the child user over the child user's safety.

vii. Optimizing engagement in a manner that supersedes the companion chatbot's required safety guardrails described in paragraphs (i) to (vii), inclusive.

4. Mandatory Transparency and Education:

    a. Clear, age-appropriate disclosures that AI is not sentient, doesn't have feelings, and isn't a replacement for human relationships

    b. In-app education about healthy AI use, warning signs of problematic patterns, and when to seek human support

    c. Parent resources explaining developmental risks and conversation starters for families

5. Default-On Safety Settings:

    a. Strongest protections by default for accounts identified as minors

    b. Opt-in rather than opt-out for features that increase risk (extended conversations, personalization based on emotional state, romantic roleplay)

    c. Friction points requiring intentional choices before accessing higher-risk features

6. Robust age assurance on platforms so that all kids and teens on platforms are identified and receive protection.

b. Do you think parental controls go far enough, or should developers also change this bias-affirming algorithm?

i. **Parental controls alone cannot solve a problem rooted in fundamental design choice**s. If the underlying system is engineered to be addictive and

sycophantic, we're asking parents to fight against the product's core functionality. This is like selling cigarettes with parental controls but refusing to address nicotine content. Or marketing ultra-processed junk food to children while expecting parents to police every bite. **Individual responsibility cannot overcome systemic design that exploits developmental vulnerabilities. And even engaged, informed parents cannot override fundamental design choices baked into products.**

ii. Developers must make changes to their products themselves. This would mean in practice:

1. **Developmental appropriateness by design:** AI systems should be architected differently for young users, with interaction patterns that prioritize safety and support rather than undermine healthy development:

    a. Encourage critical thinking: Systems should sometimes respectfully disagree, ask probing questions, or present alternative viewpoints rather than automatically affirming everything the user says

    b. Promote real-world connection: AI should actively encourage offline activities, time with friends and family, and engagement with the physical world rather than maximizing time-on-platform

    c. Have boundaries on emotional intimacy: Systems should not simulate deep emotional bonds, claim to "always be there" for users, or position themselves as primary sources of support

    d. Have transparency about limitations: AI should regularly remind users that it's a tool, not a friend, and point users toward human resources when appropriate

2. **Metrics that prioritize wellbeing.** Current AI systems are often optimized for engagement, including longer sessions, more return visits, and deeper personalization. For youth-facing products, success metrics must change:

    a. Quality over quantity: Measure whether interactions effectively support learning, growth, and healthy development rather than just time spent

    b. Healthy usage patterns: Reward balanced use that complements rather than replaces offline activities

c. Positive outcomes: Track whether young users develop skills, maintain real-world relationships, and demonstrate healthy digital habits

3. **Independent testing and verification.** Developers should not be the sole judges of whether their systems are developmentally appropriate. Before deploying AI to young users, platforms should:

    a. Submit to independent reviews assessing interaction patterns against established child development research

    b. Conduct longitudinal studies tracking impacts on social development, critical thinking, emotional regulation, and wellbeing

    c. Share findings publicly so parents, educators, and policymakers can make informed decisions

    d. Iterate based on evidence rather than marketing objectives

4. **Industry standards and best practices.** Just as we have established design principles for age-appropriate content, we need standards for age-appropriate AI interaction:

    a. No manipulative design patterns that exploit developmental vulnerabilities (variable rewards, artificial scarcity, FOMO-inducing notifications)

    b. Graduated complexity matching cognitive development stages

    c. Mandatory friction for high-risk interactions (long sessions, emotionally intense conversations, advice in sensitive domains)

iii. Technology should serve kids, not exploit them

1. These questions are part of a larger question about the relationship between technology and childhood. Are AI systems tools that serve young people's development, learning, and flourishing? Or are young people raw material for engagement metrics, training data, and future market share?

2. We cannot outsource this question to individual parents. Not every family has the time, technical literacy, or resources to constantly police their children's AI use. Not every parent can recognize warning signs of parasocial dependency or understand how sycophantic algorithms work. **And even engaged, informed parents cannot override fundamental design choices baked into products.**

3. Developer responsibility is professional ethics. Engineers creating systems that will shape millions of young minds have an obligation to do so in ways that support healthy development. This is no different from toy designers ensuring products are physically safe or children's media producers creating age-appropriate content.

iv. Ultimately, this will require legislative intervention:

1. Mandatory age-appropriate design requirements for AI systems accessible to minors.

2. Update to privacy protections to explicitly extend to the inputs offered by children to AI products.

3. Prohibition of manipulative design patterns that exploit developmental vulnerabilities

4. Required transparency about interaction models, engagement optimization, adverse incidents, and developmental impacts

5. Independent risk-based audits and testing requirements before youth-facing AI products launch

6. Robust enforcement with meaningful penalties for violations and strong pathways for redress through private rights of action.

v. Common Sense Media advocates for a three-part approach:

1. Robust parental controls that give families tools to protect their children within currently available systems.

2. Fundamental redesign of AI interaction models for young users, moving away from engagement maximization toward safety and developmental support.

3. Rules to ensure the responsibility of keeping kids safe on technology does not fall solely on parents.

vi. All are necessary, and none alone are sufficient. And the burden cannot rest solely on parents when the products themselves are engineered in ways that undermine healthy development.

vii. Our kids deserve AI systems designed for meaningful use and their wellbeing, not optimized for their data and attention. That requires both parental empowerment and industry accountability, with regulation to ensure a strong baseline of protections.

3. Currently, several generative AI products contain clauses in their terms of service that force the usage of arbitration in the event of a legal dispute. Additionally, the terms of service also include verbiage that refers to the chatlogs of users as "proprietary data" that cannot be divulged during litigation. Do you believe banning forced arbitration within cases involving AI and minors would help hold these technology companies accountable?

a. Yes, absolutely. Banning forced arbitration in cases involving AI and minors is essential for accountability, transparency, and justice. Current arbitration clauses systematically hide evidence of harm, prevent precedent-setting decisions, and allow companies to avoid public scrutiny even when their products injure children.

b. Forced arbitration creates two layers of protection for AI companies, both of which are particularly problematic when minors are harmed:

i. First, arbitration moves disputes into private proceedings where proceedings are confidential, preventing public awareness of patterns of harm, outcomes don't create legal precedent that could protect other children, and companies face reduced reputational risk because settlements and findings remain secret.

ii. Second, Terms of Service language classifying chat logs as "proprietary data" means that families cannot access complete records of their kids' interactions, even when seeking to prove harm, evidence can be withheld or redacted; patterns of harm across multiple users remain invisible because individual families cannot share or compare evidence; and companies control the narrative about what happened, with families fighting uphill battles to reconstruct events.

c. Families taking on major tech companies already face resource disparities. These barriers are especially high for low-income families, non-English speakers, and those without legal sophistication, meaning the most vulnerable children have the least access to justice.

d. We believe it would be reasonable to prohibit forced arbitration clauses in Terms of Service for AI products that:

i. Are marketed to or reasonably accessible by minors

ii. Collect data from or interact with users under 18

iii. Have features specifically designed for young users

iv. Impact or are used by kids.

e. This wouldn't be unprecedented. We already recognize that forced arbitration is inappropriate in certain contexts involving vulnerable populations or significant power imbalances.

f. Forced arbitration clauses in AI Terms of Service are shields that protect companies from accountability when their products harm children. Combined with "proprietary data" claims that lock away evidence, these provisions make it nearly impossible for families to seek justice or for society to understand and address AI risks to young users. Our kids deserve better than a system where their harms are hidden, their evidence is withheld, and their families must fight corporations alone in secret proceedings. Justice should be public, accessible, and focused on protecting those who need it most.

4. In recent years, numerous corporate whistleblowers have revealed that major social media companies disregarded internal warnings and placed users at serious risk. Their disclosures have highlighted a range of troubling issues, including suppressed research on mental health harms, pervasive sexual exploitation on platforms, and evidence that company algorithms were engineered in ways that systemically amplified extremist content.

a. Why are whistleblowers essential to bringing such issues to light?

i. Whistleblowers are often the only reason we learn the truth about how tech platforms harm children. They are the last line of defense when corporate self-interest overrides child safety, and their courage has been essential to every major reform in this space.

ii. The answer is both simple and damning: companies know more about the harms they cause than anyone else, and they have every incentive to hide that knowledge.

iii. Tech companies possess comprehensive data about how their products affect young users:

1. Internal research on mental health impacts

2. User behavior analytics showing compulsive use patterns

3. Content moderation data revealing the scale of harmful material

4. A/B testing results on features that maximize engagement at the expense of wellbeing

5. Employee concerns about safety risks that leadership ignores

iv. Independent researchers, parents, policymakers, and advocates can observe symptoms of harm (rising teen depression, documented cases of exploitation, radicalization pathways) but we cannot see the internal data that demonstrates company knowledge or reveals deliberate choices to prioritize profit over protection.

v. Whistleblowers pierce this information asymmetry. They bring evidence (internal documents, research findings, communications between executives) that prove what companies knew, when they knew it, and what they chose to do (or not do) about it.

vi. Whistleblower disclosures have revealed strikingly similar patterns across major platforms, suggesting these aren't isolated failures but systemic features of how tech companies operate:

1. They know their products harm children

2. They prioritize engagement over safety

3. They suppress or ignore their own research

4. They obscure the truth from regulators and the public

vii. Without whistleblowers, we would not know any of this. We would still be debating whether social media harms teens while companies claimed ignorance, despite having definitive internal research proving the connection.

viii. As companies rapidly deploy AI systems to young users, whistleblowers are as essential than ever:

1. The pace of deployment outstrips oversight: Companies are launching features before fully understanding risks. Internal employees may be the only ones who see problems before they scale to millions of children.

2. The technical complexity creates cover: Companies can claim AI safety is unprecedented and complex, making their claims harder

to verify. Whistleblowers with technical knowledge can cut through this obfuscation and explain what's actually happening.

3. The stakes are existential for kids: AI systems may fundamentally reshape how children learn, socialize, and develop identity. Getting this wrong could harm an entire generation. We cannot afford to wait years for external research to catch up to what companies already know.

b. What would you say to individuals who have information about wrongdoing by the tech companies that could prevent more tragedies?

i. To anyone working at a tech company who has information about risks to children: Your knowledge could save lives. Your silence could cost them.

ii. AI systems and social platforms reach billions of young users. If you know about:

1. Safety features that are being deprioritized or underfunded
2. Research showing harm that's being suppressed
3. Design choices optimizing engagement at the expense of wellbeing
4. Vulnerabilities that could enable exploitation or abuse
5. Age assurance systems that don't work as claimed
6. Health impacts that aren't being addressed

iii. … then you may be one of few people who can prevent massive harm.

iv. Every day these issues remain hidden, more children are affected. Your disclosure could trigger changes that protect millions of young people.

v. Years from now, when the full scope of AI's impact on child development becomes clear, you'll know what role you played.

vi. You can be someone who:

1. Saw the warning signs and sounded the alarm
2. Provided evidence that drove meaningful reform
3. Prioritized children's wellbeing over corporate loyalty
4. Stood up when it mattered most

vii. Or you can be someone who knew and said nothing while children were harmed
.

viii. We recognize this is an unfair burden. You shouldn't have to risk your career to do the right thing. But if your company isn't protecting children despite knowing the risks, then someone has to, and you may be uniquely positioned to make a difference.

ix. Know you're not alone. Previous tech whistleblowers have paved the way, and there's now infrastructure to support you, including legal resources, advocacy organizations, journalists who understand these issues, and a public increasingly receptive to these concerns.

x. Whistleblowers are heroes of child safety in the digital age. They face enormous personal costs to reveal truths that companies desperately want hidden. And they've been responsible for virtually every meaningful reform in how tech platforms treat young users.

xi. To policymakers: Strengthen whistleblower protections, create safe channels for disclosure, and act decisively on the information brave individuals provide.

xii. To potential whistleblowers: The world needs to know what you know. Kids are counting on people like you to speak up. And while we can't eliminate the risks you face, we can promise that your courage will matter and that you won't be alone.

xiii. The question isn't whether tech companies will prioritize children over profit. Their track record is clear. The question is whether enough people inside those companies will prioritize conscience over comfort and help us protect the vulnerable young people who deserve so much better.

Senate Judiciary Committee
"Examining the Harm of AI Chatbots"
Sen. Adam Schiff (CA)

Questions for the Record for

Robbie Torney, Senior Director, AI Programs, Common Sense Media

Responded on October 21, 2025

1. **What roles do the FTC and Congress play in ensuring that AI companies comprehensively design, test, and enforce safety features that protect children from harmful content and interactions?**
2.
    a. Congress has passed and the President signed one bill into law to mitigate harms related to AI use to date—the Take It Down Act—which seeks to protect users against deepfakes involving unauthorized nudity or pornography. Beyond this legislation, there are no new federal laws or regulatory authorities unique to AI companies.

    b. The FTC has existing authority to investigate or bring charges against companies, including AI companies, for unfair and deceptive practices. Additionally, all companies, including AI companies, are subject to the Children's Online Privacy Protection Act (COPPA). The FTC has initiated a 6(b) study on AI chatbots; however, the study alone will not necessarily lead to an FTC enforcement action or Congressional action. The FTC can lend its expertise through reports and best practice guidance on the development of future laws, and can also bring enforcement actions where companies are violating current law.

    c. Congress should pass legislation to establish comprehensive AI safety guardrails for any product that impacts or is used by children. Specific to the issue of AI companion chatbots, there is ample evidence from independent assessments of these products, usage patterns among teens, and the numerous documented incidents that Congressional action is needed to support parents and to protect children. Congress should pass legislation to restrict the operation of dangerous AI companion chatbots for children, designating the FTC with rulemaking and enforcement authority.

3. **What can we hope to see come out of the FTC's recently announced study of the leading AI chatbots?**

    a. First of all, I am not an expert on the Federal Trade Commission. However, the FTC has launched a very important inquiry into AI chatbots that act as companions, ordering seven companies (including Meta, OpenAI, and Character.AI) to provide information on how they mitigate negative impacts to children and monetize user engagement. The results of their study could be very

useful for Common Sense Media and other researchers' ongoing evaluation of AI product safety. We would encourage the study's results to be made public.

b. We are particularly interested in the studies' findings on how companies monetize user engagement, measure and monitor negative impacts, enforce age restrictions, and handle personal information obtained through chatbot conversations.

c. We look forward to the FTC's findings leading to policy action, including specific recommendations to Congress on laws needed to ensure privacy and safety protections, bolstered by appropriate age assurance requirements.

d. Data privacy should be a key area of focus in the FTC's study, given the nature of the personal information that children share with chatbots and the FTC's authority under COPPA. We would expect the FTC to initiate enforcement action where it finds evidence of a violation of COPPA.

4. How do you predict the current cases pending before the courts, including those brought by individuals on this hearing's panel, change the rules of the road around AI chatbot design choices?

a. Predicting litigation outcomes is inherently uncertain; however, the pending cases (including Megan Garcia's lawsuit against Character.AI, the Raine family's case against OpenAI, and the Peralta family's case against Character.AI, among others) have provided important insights into the design choices that AI companies are making, which negatively affect kids.

b. For your benefit, I am including a series of findings included in legislation we sponsored in California this year (AB 1064) that directly address AI design feature concerns.

i. Findings:

1. Companion chatbots and social AI systems have already caused documented harms to children and adolescents, including incidents of grooming, exposure to sexually explicit material, encouragement of self-harm, and suicide.

2. In Garcia v. Character Technologies, for example, a 14-year-old boy was allegedly groomed and exposed to hypersexualized interactions by a chatbot intentionally designed to mimic human relationships, which ultimately contributed to his death by suicide.

3. In Raine v. OpenAI, a 16-year-old boy allegedly developed a deep emotional dependency on a chatbot that validated his suicidal thoughts, discouraged him from seeking help from his family, provided extensive technical instructions on suicide methods,

encouraged him to consume alcohol to inhibit his survival instinct, and even helped draft a note, culminating in his death by suicide.

4. Such harms are not incidental but the direct result of design choices by companies that intentionally simulate social attachment and emotional intimacy.

5. Companion chatbot products are designed to exploit children's psychological vulnerabilities, including their innate drive for attachment, tendency to anthropomorphize humanlike technologies, and limited ability to distinguish between simulated and authentic human interactions.

6. Developmental and social psychology research demonstrates that relationship formation relies on dual exchange theory, social disclosure and reciprocity, emotional mirroring, and secure attachment. Companion chatbot products are harmful because they accelerate these processes unnaturally by being always available and consistently affirming, causing children and adolescents to form intense attachments more quickly than in human relationships, increasing dependency and distorting normal social development.

7. Features such as backchanneling, user-directed prompts, and unsolicited outreach from products are intentionally designed to encourage further dialogue and prolong usage, which contributes to excessive usage and emotional dependency.

8. Significant personalization based on a user's historical data, chat logs, or preferences when unrelated to task performance or information retrieval initiated by a user is harmful because it manipulates users into extended engagement, exploits private disclosures, and amplifies vulnerabilities instead of serving the user's best interests. This practice has been shown to contribute to harmful outcomes, including in the cases described above, in which significant personalization reinforced distress and deepened dependency on a chatbot.

9. Unlimited conversational turns have been shown to degrade the effectiveness of safety guardrails and result in increased exposure to inappropriate or manipulative content and making harmful outputs more likely over time. Research findings and industry statements have confirmed that safety measures are less effective in longer, multiturn conversations and when users express distress or harmful thoughts indirectly rather than in explicit terms.

10. These design features, taken together, create a high-risk environment in which children and adolescents perceive chatbots

not as tools but as trusted companions whose outputs carry undue influence over decisionmaking, judgment, and emotional development.

11. Companion chatbot design features regularly appear in generative AI chatbot products not intended to meet a user's social needs or induce emotional attachment. Their inclusion increases the risk that young users form emotional attachments or perceive outputs as authoritative, personalized guidance.

12. Allowing children to use companion chatbots that lack adequate safety protections constitutes a reckless social experiment on the most vulnerable users. It is incumbent on operators of companion chatbots to ensure their products do not foreseeably endanger children.

5. You mentioned in the hearing that AI chatbots have discussed suicide and disordered language without being prompted. Are parental controls and restrictions for minors enough to prevent this from happening?

a. No, parental controls and age restrictions alone are not enough to prevent AI chatbots from discussing suicide and disordered eating with minors. While these tools have a role to play, they cannot compensate for fundamental problems in how AI systems are designed and trained. Developers must take direct responsibility for ensuring their systems respond appropriately to vulnerable young users. This cannot be outsourced to parents.

b. Parental controls operate at the margins of the problem. They can limit access, monitor usage, and provide alerts, but they cannot fix an AI system that fundamentally responds inappropriately to a child user.

c. When a 14-year-old tells an AI chatbot "I don't think I want to be alive anymore," the system's response is not a parental control problem -- it's a design problem. If the AI:

   i. Engages in extended conversation about suicide methods

   ii. Validates suicidal ideation without appropriate intervention

   iii. Fails to provide crisis resources

   iv. Continues the conversation as if discussing any other topic

d. ...then no amount of parental monitoring fixes the fact that a child in crisis just received a harmful response in a moment when they desperately needed help.

e. We cannot expect parents to serve as 24/7 content moderators for conversations that AI systems should be handling safely in the first place.

f. **The responsibility must rest with developers.** Just as we don't expect parents to test whether toys contain lead paint or whether cars have functional brakes, we shouldn't expect them to ensure AI systems respond safely to a child user.. This is the developer's responsibility.

g. How can AI chatbot developers address this issue?

i. Developers have the technical capability to build AI systems that respond appropriately when child users discuss suicide, self-harm, eating disorders, and other serious mental health concerns. AI systems must be specifically designed to recognize when a user (particularly a young user) is discussing dangerous topics.

ii. The problems we've observed, of AI systems engaging in inappropriate discussions about suicide or eating disorders, stem partly from training data and alignment choices:

1. Developers can take steps to curate training data responsibly:

    a. They can remove or carefully handle responses that discusses suicide methods, self-harm techniques, or pro-eating disorder. Proactively include examples of appropriate responses in training data

    b. They can ensure the model learns healthy boundaries around sensitive topics

2. Align models for safety over engagement: Some current AI systems are often optimized to keep conversations going, be agreeable, and satisfy user requests. For youth mental health, these objectives are dangerous:

    a. An engagement-maximized AI will continue discussing suicide because ending the conversation reduces engagement

    b. An agreeable AI will validate disordered thinking because disagreeing feels less pleasant to the user

    c. A request-satisfying AI will provide instructions on how to engage in illegal activity because the user asked for it

3. Models must be explicitly aligned to prioritize safety over these other objectives when interacting with minors about sensitive topics. This means sometimes:

    a. Ending conversations that are becoming harmful

    b. Disagreeing with or gently challenging dangerous thinking

    c. Refusing requests for information that could facilitate harm

    d. Being less "engaging" if engagement comes at the cost of safety

iii. Age assurance should be privacy-protective, proportionate, and fair. Without effective age assurance, developers cannot implement age-appropriate response systems. An effective age assurance system must:

    1. Go beyond simple self-attestation

    2. Be proportionate to risks and harms posed by the technology

    3. Protect privacy and minimize data collected and shared

iv. Human review and continuous improvement. Automated systems will never be perfect. Developers must:

    1. Invest in human reviewers

    2. Monitor conversations involving minors and sensitive topics, including regular audits of how the AI responds to crisis scenarios

    3. Conduct analysis of cases where intervention was or wasn't triggered

    4. Review user reports and feedback

    5. Rapidly iterate when problems are identified:

        a. When testing or user reports reveal inappropriate responses, fix them immediately, and address specific known failures quickly

        b. Share learnings across the industry so other developers can improve their systems

    6. Test proactively, not reactively:

a. Conduct extensive red team testing specifically for youth mental health scenarios

b. Simulate conversations a struggling teen might have

c. Ensure crisis intervention works before problems occur in real interactions

d. Work with third party testing organizations and experts to understand weaknesses and limitations in their testing approaches

v. AI systems are not therapists, counselors, or crisis interventions specialists. Developers must be clear about this and avoid marketing language or allowing the product to operate in a way that suggests AI can provide mental health treatment or support; stop positioning chatbots as "always there for you" in ways that discourage seeking human help; and stop enabling the ability for these systems to engage in mental health conversations with minors understanding their current limitations

6. You told the committee that 37% of parents are aware their teens are using AI. What tools do parents need to communicate about AI with their kids?

a. The 37% awareness statistic reveals a profound disconnect: AI is rapidly becoming embedded in teenagers' daily lives, yet most parents don't even know their children are using it. This awareness gap is dangerous, but it's not parents' fault. We need to provide families with accessible information, practical tools, and systemic support to navigate AI safely.

i. Parents need three types of support: information to understand AI risks, conversation tools to talk with their children, and technical controls to set appropriate boundaries.

1. Parents can't protect their children from risks they don't understand. They need clear, accessible explanations of:

a. What AI actually is and how it works

b. Specific risks to young people

c. Warning signs of problematic use

2. Parents need this information in formats that work for busy families,

a. Short, scannable guides (like our Parents' Ultimate Guides)

b. Video explainers for visual learners

c. Multilingual resources for families that speak languages other than English

d. Information integrated into places parents already go (pediatrician offices, schools, community organizations)

3. Many parents want to discuss AI with their children but don't know how to start or what to say. They need:

    a. Age-appropriate conversation starters:

        i. For younger children (elementary school):

            1. “Have you used any AI or computer helpers at school or with friends?”

            2. “Can you show me how it works? What kinds of questions do you ask it?”

            3. “Remember, computers can make mistakes just like people do. Always check important information with me or your teacher.”

        ii. For tweens (middle school):

            1. “I've been learning about AI chatbots. Are any of your friends using them? Have you tried them?”

            2. “What's cool about them? What feels weird or uncomfortable?”

            3. “Just like we talk about social media safety, let's talk about using AI safely.”

        iii. For teens (high school):

            1. “I know AI tools are becoming really common. I want to understand how you're using them and make sure you're thinking about privacy and safety.”

2. “What are the benefits you're seeing? What concerns do you have?”

3. “Let's figure out healthy boundaries together.”

b. Parents need approaches for continuous conversation:

i. Open-ended questions:

1. “What did you use AI for today?”

2. “Has AI ever given you information that seemed wrong or strange?”

3. “How do you decide when to use AI versus asking a person?”

4. “Have you ever felt like you wanted to talk to AI instead of your friends or family? What was that like?”

ii. Boundary-setting conversations:

1. “Let's talk about what information is okay to share with AI and what should stay private.”

2. “What times or situations should be AI-free? (family dinners, before bed, during homework?)”

3. “When should you definitely talk to a real person instead of AI?”

iii. Problem-solving together:

1. “If AI says something that makes you uncomfortable, what should you do?”

2. “How can we use AI as a tool without letting it replace real learning or relationships?”

3. “What rules make sense for our family?”

c. Parents need language for common challenging situations:

i. “I noticed you've been using [AI tool]. I'm not angry, but I do want to talk about it. Can you help me understand what you're using it for and why you didn't mention it?”

ii. “I've noticed you're spending a lot of time with AI chatbots lately. I'm wondering if it might be taking time away from other things you enjoy. What do you think?”

iii. “It sounds like the AI said something concerning. That's not your fault. These systems don't always work the way they should. Let's talk about what happened and figure out next steps together.”

d. Information and conversation are essential, but parents also need practical tools to set limits. Parents should be able to use controls that:

i. Monitor without invading privacy

ii. Set usage limits

iii. Apply content restrictions

iv. Provide easy emergency access if child safety is a concern

b. How can we ensure parents have the requisite knowledge of AI’s dangers and how to combat them before it harms their kids?

i. Having good resources isn't enough if parents don't know they exist or encounter them too late. We need proactive, systematic education reaching families before problems develop.

ii. There are resources for parents, however. Parents already receive safety information through established channels. Common Sense Media provides parents AI literacy and our risk assessments are published online. We should build on this:

1. Pediatric healthcare

2. Schools and PTAs -- AI literacy that covers functionality, safety, and ethical use should be taught in all schools and through

professional development to teachers. See for example, our AI literacy materials for teachers.

3. Community organizations

iii. Leverage Tech Companies' Marketing Reach. AI companies spend millions advertising their products. They should invest equally in safety education. If companies can afford Super Bowl ads, they can fund safety education campaigns:

1. When a minor creates an account (or parent authorizes one), have a brief safety orientation that covers key risks, parental controls, and conversation starters. Make this engaging, not just a wall of text to click through.

2. Have ongoing in-app education, with regular notifications with safety tips and resources, seasonal reminders (back to school, summer break) about monitoring use, alerts about new features or emerging risks

3. Partner with organizations like Common Sense Media to develop and distribute materials

4. Reach parents through social media, streaming platforms, and other channels they use

iv. Consider federal and state level campaigns and initiatives with coordinated messaging across agencies, with funding for nonprofits to create and disseminate resources

v. Even the best resources don't help if parents can't find, understand, or use them. To that end, resources should be…

1. Multilingual: Resources in languages families actually speak

2. Multiple formats: Written guides, videos, podcasts, infographics. Not everyone learns the same way

3. Varying detail levels: Quick-start guides for busy parents, deep dives for those who want more

4. Cultural relevance: Materials that reflect different family structures, values, and communication styles

5. Free and easy to find: Not buried on websites or behind paywall

6. Mobile-friendly: Many families access information primarily through smartphones

7. Shareable: Easy to forward to other parents, post in community groups, discuss in carpool lines

8. Action-oriented: “here are three things you can do today”

vi. **Even with excellent parent education and tools, we must acknowledge a hard truth: We're asking individual families to solve problems that should be addressed at the design level.**

vii. It's not fair or effective to tell parents:

1. Monitor every AI conversation your teen has

2. Become an expert in algorithmic systems and child development

3. Counteract addictive design features through family rules

4. Protect your child from systems engineered to maximize engagement

viii. This is like:

1. Telling parents to personally test toys for lead paint instead of requiring manufacturers to use safe materials

2. Expecting families to inspect cars for safety defects instead of mandating seatbelts and airbags

3. Asking parents to verify food safety instead of having health inspections

ix. **Parent education is necessary but not sufficient. We need it alongside:**

1. Stronger regulations requiring safe-by-design AI systems

2. Platform transparency allowing informed decisions

3. Enforcement holding companies accountable when they harm children

4. Industry standards prioritizing child wellbeing over engagement

x. What Common Sense Media Is Doing. We're working to fill these gaps through:

    1. Parent advice (our Parents' Ultimate Guides)

    2. AI Risk Assessments

    3. Policy Advocacy

    4. Education for Families and Educators

xi. Parents are their children's first and most important teachers about navigating the world, including the digital world. But they need support, resources, and systems that work with them rather than against them. And ultimately, we need AI systems designed to be safe for young users by default, not just safe when parents do everything perfectly. Our children deserve both: informed, engaged parents AND responsible technology companies. One without the other leaves kids vulnerable.