Institute for Youth in Policy

Policy Media Department | Brief

5214F Diamond Heights Blvd #3055
San Francisco, CA 94109
+1 (509) 565-0545
eic@yipinstitute.org
www.yipinstitute.org

AI Content Safeguards for Minors

Policy Leads: Nishna Makala[#1]

Policy Analysts: Queen-Aset Blissett[#2], Lola Agboola[#3], Lachlan Feichthaler[#4], Bhavana Rupakula[#5]

Technology Policy, Institute for Youth in Policy

[1]nishna@yipinstitute.org

***Abstract* — With the growing prevalence of AI-generated content on the internet, children become more vulnerable to manipulation through disinformation, inappropriate, and misleading media. This policy brief aims at introducing measures to guarantee that online platforms will be obliged to reveal the use of AI-generated content, mitigate its negative algorithmic impact, conduct human review, and be subject to independent oversight.**

***Keywords* — Artificial Intelligence (AI); online safety; misinformation; deepfakes; content moderation; algorithmic recommendations; transparency; Federal Trade Commission (FTC); digital platforms.**

I. Executive Summary

The following bill develops safeguards for AI-generated content shared on online platforms, where minors are likely to encounter them. Responding to the growing use of AI-generated text, images, videos in online media, this proposal aims to reduce children's exposure to misleading, manipulative, or age-inappropriate materials.

Key findings from UNICEF indicate that minors are especially vulnerable to misleading AI-generated content as they are less likely to distinguish synthetic media from authentic content and may encounter these content from platform recommendation systems. Thus, the bill recommends the mandatory disclosure of AI-generated content, independent auditing, and enforceable oversight mechanisms that prioritize transparency, age-appropriate content design, children's safety, privacy, and platform compliance.

II. Overview

A. Relevance

The increasing prevalence of realistic AI-generated content in digital services prompts major implications for the livelihoods of children online. Not only can AI-generated images, videos, and text fuel the mass dissemination of explicit material or disinformation online, recommendation algorithms may also amplify children's exposure to this content by prioritizing engagement over user safety. Prolonged exposure to misleading AI-generated content, in turn, can negatively affect children's mental health by contributing to anxiety, low self-esteem, unhealthy social comparisons, and distrust.

Holding online platforms such as social media and search engine platforms accountable for proliferating AI-generated disinformation is essential to promoting a safer digital ecosystem for children. By initiating legislative efforts that require transparency and age-appropriate safeguards, policymakers can promote the well-being of minors online.

B. Background

Children are being exposed to misinformation at an unprecedented pace.

Despite the myriad benefits of AI, it has repeatedly been misused to spread disinformation, generate inappropriate photos, and amplify these content by increasing engagement through internal algorithms. As such, safeguards on AI will significantly contribute to online child safety.

C. Notable Stakeholders

Children are the primary beneficiaries of safeguards for AI content as policy change directly affects their online safety and educational experiences. In addition, educators, parents, caregivers, and other community members responsible for supporting children's well-being arise as another key stakeholder in identifying the effects of harmful, misleading, or age-inappropriate AI content. Schools and child advocacy organizations are similarly affected as they play an important role in promoting safe technology use among youth. Lastly, digital platforms and technology companies hosting AI-generated content are also critical in implementing age-appropriate design features and content moderation practices.

III. Impact

The rising prevalence of artificial intelligence (AI) adversely impacts minors by making it more difficult for them to distinguish AI-generated content from authentic media. As AI becomes more deeply integrated into social media platforms and company advertisements, minors grow more susceptible to disinformation, increasingly unable to discern genuine content.

A notable example involves the AI impersonation of Mr. Beast, a YouTuber whose content targets younger audiences. The perpetrators created a deepfake video in which he appeared to promote an "iPhone 15 giveaway" via a phishing link. Many of his younger fans failed to recognize the video as AI-generated and fell victim to the scam.

Inaccurate AI-generated videos targeting young children have also proliferated online, particularly nursery rhymes for educational purposes. For example, the YouTube video "Vroom Vroom! Car Ride Song," which has gained thousands of views, features children singing "Red means stop, and green means go right," to teach how to cross the road. This message does not accurately represent safe street-crossing practices, yet it is watched by minors who are unaware of the misinformation in this video. As children often internalize information from such content, the spread of misinformation may pose dire consequences for broader public safety.

Access to social media often begins at a very young age, sometimes even before children have developed the cognitive skills necessary to distinguish disinformation from factual information. To encourage minors to consume authentic content, it is essential to implement safeguards protecting them from accepting AI-generated misinformation.

IV. Regulatory Provisions

Online platforms with algorithmic recommendations and a reasonable probability of being accessed by minors must implement appropriate mechanisms to detect, disclose, and mitigate the spread of AI-generated content. In addition to using commercially reasonable methods to detect AI-generated images, videos, audio, and text from uploaded content, any undetected AI content is to be labeled with an appropriate notice before it is viewed, shared, or recommended. The recommendation algorithms must also be designed to mitigate the spread of AI-generated content, especially when the content may be accessed by minors.

In order to prevent automated systems from taking independent final decisions on reports, covered platforms are required to create a Human-in-the-Loop Review System for reports regarding AI content, including deepfakes, manipulation, and algorithmic amplification of such content which could cause harm to minors. Such reports will be reviewed by human moderators pre-trained in issues related to digital safety and protection of children from misinformation. These moderators may issue requests for corrective disclosures, limit the promotion of AI content algorithms, remove the content that violates this Act, or restore wrongly removed content. Platforms must provide an easily accessible mechanism for appeals and produce yearly transparency reports on the number of AI content reports, their outcomes, the response time, and disclosure compliance rate.

V. Policy Enforcement

The Federal Trade Commission (FTC) should serve as the primary federal enforcement agency as they currently enforce privacy protection efforts for children under COPPA and have offered critical views of AI products affecting children and teenagers. The Department of Justice (DOJ) may support litigation efforts in fitting federal cases, while state attorneys generals can bring parallel enforcement under state privacy, consumer protection, and child-safety laws.

Platforms that fall under compliance measures must be required to implement the bill's safeguards within 120 days before undergoing recurring audits by qualified third-party assessors approved by the FTC. Audit findings should disclose compliance alignment, the detection and removal of harmful AI-generated content, retention practices to ensure long-term impact, and examine the effectiveness of these age-appropriate protections. This approach aligns with UNICEF's recommendations for transparency, safety-testing, and continuous monitoring to improve mechanisms for child-safety efforts.

Violations, once reported, should trigger civil penalties and corrective action orders with escalation based on repeated violations and noncompliance. Annual public reporting to Congress can also be enforced, focusing on easily-verifiable and measurable violations including the number of complaints submitted, the associated enforcement actions, and proactive detection rates.

Implementation costs will be borne primarily by the covered platforms, including but not limited to the expenses associated with moderation systems, human review, independent audits, and disclosure tools. Federal oversight will be carried out by existing agency authority to lessen taxpayer burden.

VI. Conclusion

Following the passing of this bill, the FTC would enforce all major web browsers, social media websites, and any other internet service that falls under its framework to add these safeguards against AI with a 120-day grace period. The FTC will then heavily monitor the services for the next six months to ensure compliance. After the six-month period, oversight will be reduced but will ensure the requirements are upheld.

VII. Acknowledgement

The Institute for Youth in Policy wishes to acknowledge Donna Kim for editing this policy brief.

VIII. References

Aerps.com. *Laptop Displays "The AI Code Editor" Website*. Photograph. Published May 3, 2025. Unsplash. https://unsplash.com/photos/laptop-displays-the-ai-code-editor-website-wjFOjA2zXy8.

Dewey, J. (n.d.). *Media Manipulation | Social Sciences and Humanities | Research Starters*. EBSCO. Retrieved July 12, 2026, from https://www.ebsco.com/research-starters/social-sciences-and-humanities/media-manipulation.

Federal Trade Commission. (2026, February 25). *FTC Issues COPPA Policy Statement to Incentivize the Use of Age Verification Technologies to Protect Children Online*. Retrieved July 12, 2026, from https://www.ftc.gov/news-events/news/press-releases/2026/02/ftc-issues-coppa-policy-statement-incentivize-use-age-verification-technologies-protect-children.

Guidance on AI and Children Draft 3.0. (2025, December). Retrieved July 12, 2026, from https://www.unicef.org/innocenti/media/11991/file/UNICEF-Innocenti-Guidance-on-AI-and-Children-3-2025.pdf.

Howard, P. N., Neudert, L.-M., Prakash, N., & Vosloo, S. (2021, August). *Digital misinformation / disinformation and children* (Global Insight Policy Brief). UNICEF Office of Global Insight and Policy. Retrieved July 12, 2026, from https://www.unicef.org/innocenti/media/856/file/UNICEF-Global-Insight-Digital-Mis-Disinformation-and-Children-2021.pdf.

Policy guidance on AI for children Draft 1.0. UNICEF. (2020, September). Retrieved July 12, 2026, from https://www.unicef.org/innocenti/media/1326/file/UNICEF-Global-Insight-policy-guidance-AI-children-draft-1.0-2020.pdf.

Policy guidance on AI for children Draft 2.0. UNICEF. (2021, November). Retrieved July 12, 2026, from https://www.unicef.org/innocenti/media/1341/file/UNICEF-Global-Insight-policy-guidance-AI-children-2.0-2021.pdf.

Policy guidance on AI for children Draft 3.0. UNICEF. (2025, December). Retrieved July 12, 2026, from https://www.unicef.org/innocenti/media/11991/file/UNICEF-Innocenti-Guidance-on-AI-and-Children-3-2025.pdf.

Synthetic media and its identification and detection | ICO. (n.d.). Information Commissioner's Office. Retrieved July 12, 2026, from https://ico.org.uk/about-the-ico/research-reports-impact-and-evaluation/research-and-reports/technology-and-innovation/tech-horizons-and-ico-tech-futures/tech-horizons-report-2025/synthetic-media-and-its-identification-and-detection/.

UNICEF. (2025, December 4). *Guidance on AI and children | Office of Strategy and Evidence Innocenti*. UNICEF. Retrieved July 12, 2026, from https://www.unicef.org/innocenti/reports/policy-guidance-ai-children.