Protecting Children in the Age of Generative AI

April 2026

OpenAI

Foreword

“As Co-Chairs of the Attorney General Alliance's AI Task Force, we welcome this blueprint as a meaningful step toward aligning the technology sector's child safety practices with the enforcement realities our offices confront every day. We are particularly encouraged by the framework's recognition that effective GenAI safeguards require layered defenses — not a single technical control, but a combination of detection, refusal mechanisms, human oversight, and continuous adaptation to emerging misuse patterns. This mirrors what we see in practice: the threat evolves constantly, and static solutions are insufficient. Getting the prevention architecture right upstream is the single highest-leverage investment the industry can make in child safety.

Ultimately, the strength of any voluntary framework depends on the specificity of its commitments and the willingness of industry to be held accountable against them. We look forward to continued partnership with OpenAI, NCMEC, and our fellow Attorneys General to ensure these recommendations translate into durable protections for children.”

State Attorneys General Jeff Jackson (North Carolina) and Derek Brown (Utah)
Co-Chairs of the Attorney General Alliance Artificial Intelligence Task Force

“The Attorney General Alliance is leading the way in protecting young people online by bringing together attorneys general, industry leaders, nonprofits, and global partners to advance practical, forward-looking solutions on AI and digital safety. Through collaboration and innovation, AGA is setting a strong standard for how we safeguard youth while responsibly embracing emerging technologies. We applaud OpenAI’s continuing commitment to safety and engagement with AGA and attorneys general in developing a highly valuable blueprint for child safety.”

Karen White
Executive Director of Attorney General Alliance

“Generative AI is accelerating the crime of online child sexual exploitation in deeply troubling ways - lowering barriers, increasing scale, and enabling new forms of harm. But at the same time, the National Center for Missing & Exploited Children (NCMEC) is encouraged to see companies like OpenAI reflect on how these tools can be designed more responsibly, with safeguards built in from the start. No single organization, business or sector can address this alone. We remain committed to working with partners across industry, government, and the child protection community to advance solutions that reduce harm and better support children’s safety.”

Michelle DeLaune
President & CEO, National Center for Missing & Exploited Children

A Policy Blueprint for Preventing and Disrupting Online Child Sexual Exploitation and Abuse Involving Generative AI

Child sexual exploitation remains one of the most urgent public safety challenges of the digital era, and generative AI is reshaping both how these harms emerge and how they can be prevented. AI systems can be misused to create synthetic child sexual abuse material (CSAM), digitally alter existing imagery, and scale grooming activity across platforms and jurisdictions. At the same time, these tools can enable earlier detection of risk, improved prioritization of high-harm cases, and more effective reporting pipelines that support faster action by the National Center for Missing & Exploited Children (NCMEC) and law enforcement.

This blueprint provides a practical roadmap to strengthen U.S. child protection frameworks as statutory and operational approaches adapt to the emerging challenges posed by artificial intelligence. It advances three reinforcing priorities:

- **State legislative modernization** to ensure CSAM statutes cover AI-generated and digitally altered material, clarify attempt liability, and enable a good-faith CSAM prevention safe harbor.

- **Provider reporting and coordination standards** to improve the quality and actionability of CyberTipline submissions, reduce investigative burden, and strengthen collaboration with NCMEC and ICAC task forces.

- **GenAI prevention and detection safeguards** to interrupt attempts to exploit children upstream through safety-by-design controls, human-in-the-loop review, and consistent approaches to classifying suspected synthetic content.

The Evolving Threat Landscape

Digital services have long been used by bad actors to facilitate child sexual abuse, but generative AI introduces new dynamics that challenge traditional legal and investigative models. Synthetic imagery may be created or manipulated without direct access to a victim, and offenders may operate at greater scale and speed across modalities (text, image, video) and jurisdictions. These developments strain investigative resources and expose gaps in statutory definitions, reporting expectations, and prevention mechanisms, while also creating opportunities for earlier detection and disruption when paired with clear safeguards and standards.

Supporting Effective Enforcement

Effective child protection ultimately depends on strong investigative and prosecutorial authority. The recommendations in this blueprint are intended to strengthen, not replace or limit, existing enforcement tools available to state and local prosecutors, state attorneys general, and law enforcement agencies. Modernizing statutory frameworks, improving reporting quality, and encouraging responsible safety practices can enhance investigators' ability to identify victims, disrupt offenders, and pursue accountability as technologies evolve. Nothing in this framework is intended to reduce existing legal obligations or enforcement authority; rather, these recommendations aim to ensure child protection laws remain effective as technology continues to evolve.

Framework Overview

Addressing AI-facilitated child sexual exploitation requires coordinated action across the full harm lifecycle—from prevention and detection to reporting, investigation, and victim identification. No single intervention is sufficient on its own.

The framework advanced in this blueprint is organized around three mutually reinforcing priority areas:

- Priority Area One: State Legislative Modernization enables enforceable action;
- Priority Area Two: Best Practices – Provider Reporting and Coordination Standards strengthens investigative response;
- Priority Area Three: Best Practices – Safety-by-Design GenAI Prevention & Detection Safeguards reduce opportunities for misuse upstream.

Together, these elements are intended to reinforce one another: prevention reduces investigative burden, higher-quality reporting accelerates victim identification, and updated laws ensure enforcement tools remain effective as technologies advance.

Collectively, this framework aligns legal authorities, provider operational practices, and AI system safeguards so that risks can be identified earlier, reports can be acted upon more quickly, and investigators can pursue accountability more effectively.

Priority Area One: State Legislative Modernization

Objective: Ensure state child exploitation statutes remain fully enforceable when applied to AI systems and digital services, while encouraging rigorous and robust safety efforts by responsible providers. As

state law becomes increasingly central to protecting children online, fragmented statutes can create uneven enforcement and uncertainty. Targeted updates can close gaps created by new technologies while strengthening investigators' ability to pursue offenders.

1) Modernize CSAM Definitions to Cover Synthetic and Digitally Altered Material

Many CSAM statutes were drafted before synthetic media. Legislatures should make clear that AI-generated or digitally altered child sexual abuse material is covered under existing prohibitions, and that liability does not turn on the technological form of the content. Updating definitions ensures prosecutors retain clear authority and prevents offenders from exploiting technological or statutory ambiguity.

Recommendation: States should update CSAM statutes to explicitly prohibit:

- AI-generated CSAM
- Digitally altered/computer-edited CSAM
- Knowing and intentional possession, production, and distribution of such material, consistent with existing CSAM frameworks

Definitions should be durable and technologically workable, capturing synthetic depictions without requiring unrealistic proof standards.

According to research by ENOUGH ABUSE®[1], as of August 2025, most states – 45 in total – have enacted laws addressing AI-generated or computer-edited CSAM, leaving only a small number of jurisdictions, including five states and the District of Columbia, without explicit statutory coverage. More than half of these laws were enacted in 2024 and 2025, underscoring the accelerating concern among legislators and advocates about the rapid expansion of AI-generated child exploitation material.

2) Clarify Attempt Liability (Including Intentional Prompt-Based Attempts)

In the GenAI context, safeguards should successfully block prohibited outputs even when users intentionally attempt to generate abusive material. Clarifying attempt liability ensures deliberate efforts to create CSAM remain prosecutable regardless of whether safeguards intervene, and enables earlier intervention.

Recommendation: States should ensure CSAM statutes clearly prohibit:

- attempts to produce CSAM
- attempts to solicit CSAM
- attempts to upload, distribute, or traffic CSAM
- attempts involving digital manipulation or synthetic generation

[1] ENOUGH ABUSE®, *State Laws Criminalizing AI-Generated or Computer-Edited Child Sexual Abuse Material (CSAM)*, accessed March 2026, enoughabuse.org

Attempt provisions can prevent harm before illegal material is produced or disseminated and can support lawful action when offenders repeatedly test safeguards.

3) Establish Good-Faith CSAM Prevention Safe Harbor

Providers conducting responsible detection, moderation, reporting, and safety research should be able to act quickly and proactively without fear of unintended liability. Carefully tailored safe harbor provisions encourage safety investment while preserving accountability. Any safe harbor should apply only to good-faith safety activity and should not shield negligent, reckless, or unlawful conduct.

Recommendation: States should establish safe harbor protections for good-faith efforts to:

- detect child exploitation
- report CSAM to NCMEC and cooperate with law enforcement
- preserve evidence for investigations
- conduct safety research and red-teaming
- develop child protection technologies and best practices

4) Federal Alignment

States cannot solve this alone, and federal alignment can reduce cross-jurisdiction friction. We recommend federal policymakers:

- Support aligned federal proposals that improve reporting quality, evidence preservation, and accountability for online child exploitation
- Enable responsible safety testing through clear protections for good-faith image-based red teaming with the U.S. Department of Justice and cross-sector safety collaboration.
- Encourage harmonization that reduces gaps across jurisdictions and improves operational clarity for providers and law enforcement.

Priority Area Two: Best Practices - Provider Reporting & Coordination Standards

Objective: Improve CyberTipline report quality and operational collaboration so NCMEC and law enforcement can act faster with less friction.

Investigative outcomes depend not only on reporting volume, but on report quality and contextual completeness. When reports are incomplete, duplicative, or missing key identifiers, investigators often must spend scarce time doing follow-up work before they can triage risk or identify victims. There is a critical opportunity to further emphasize high-quality reporting as a core standard, ensuring submissions include clear prioritization indicators to support effective downstream triage. Raising baseline reporting

and coordination standards is one of the highest-impact ways to improve outcomes without waiting for new laws.

1) Improve CyberTipline Report Quality (Structured, Actionable Submissions)

Recommendation: Providers should include structured information in reports and include the following information, where available:

- **Who:** suspected offender identifiers (account IDs, emails/phones where available), relevant user identifiers; victim indicators where available and appropriate
- **What:** the reported content, including all relevant associated information, if available (such as the image, video, prompt, and related metadata); any identifiers or hashes; the file itself; the content modality (for example, image, video, or chat); whether the content appears to be AI-generated, suspected AI-generated, or of unknown origin; clear prioritization indicators (e.g., imminent harm signals, escalation flags, or high-risk behavioral patterns), including designation of high-priority reports or electronic service provider (ESP) escalations involving planned or imminent harms.
- **Where:** jurisdiction indicators (U.S. vs non-U.S.), location signals where lawful and available
- **When:** when the activity occurred, when detected, when reported; sequence of relevant events (e.g., repeated attempts)

High-quality, structured reports reduce investigative back-and-forth, improve triage accuracy, and enable faster routing, prioritization, and victim identification.

2) AI-Assisted Detection and Human-Reviewed Escalation

Recommendation: Providers should use audited AI systems to identify potential indicators of child sexualization or exploitative behavior and route those signals for human review prior to escalation or reporting. AI-assisted triage systems should prioritize cases presenting the highest risk of ongoing exploitation or imminent harm, while maintaining human oversight for reporting decisions.

AI-assisted detection enables providers to surface high-risk activity at scale and focus human review resources on the most urgent cases. When paired with defined escalation procedures, this approach can improve report quality, accelerate response times, and help ensure that situations involving potential immediate harm receive timely attention without increasing investigative burden through automated or low-confidence reporting.

3) Include Sufficient Context in Enticement / Trafficking-Indicator Reports

For messaging-related cases, context is often the difference between an actionable lead and a non-actionable excerpt.

Recommendation: For reports involving potential enticement or trafficking indicators, providers should include sufficient chat context to establish meaning (not only isolated excerpts), while minimizing

unnecessary personal data. Provide what is necessary for investigative triage; avoid broad over-collection.

4) Reduce Investigative Burden Through Bundling and De-Duplication

Recommendation: Where feasible, providers should bundle reports by user or incident, rather than submitting one report per file. Reports should package related files, identifiers, and behavioral patterns associated with the same actor. Bundling reduces ticket volume for both ESPs and NCMEC and helps NCMEC and ICAC task forces connect related activity quickly.

5) Use Technical Identifiers Where Available

Recommendation: Include hashes and technical identifiers, such as IP port numbers and device ID information, where available and lawful, to support cross-case linkage, pattern detection, and de-duplication.

Priority Area Three: Best Practices - Safety-by-Design - GenAI Prevention & Detection Safeguards

Objective: Interrupt exploitation attempts before harm occurs, and generate higher-quality signals when threats emerge.

Generative AI systems create opportunities to intervene earlier in the harm lifecycle by identifying misuse attempts before abusive material is produced or distributed. Strong GenAI safety is not a single control; it is a layered approach combining policy enforcement, technical safeguards, monitoring, child safety operations, and human oversight.

1) Attempt and Intent Detection

Recommendation: AI systems should detect and respond to high-risk prompts and behavioral patterns associated with attempted child exploitation, including repeated probing or iterative refinement intended to bypass safeguards. Attempt detection supports early intervention and improves the quality of signals available for safety operations and, where appropriate, lawful reporting.

2) Generation Refusal and Intervention Controls

Recommendation: Systems should refuse prohibited requests and implement intervention mechanisms (friction, throttling, escalation) when behavior indicates exploitative intent. Preventing prohibited outputs is frontline protection, particularly for synthetic CSAM risk.

3) Human Oversight for High-Risk Cases

Recommendation: Human oversight should be used for high-confidence or high-impact cases to improve classification accuracy and ensure responsible escalation decisions. Human review increases actionability and reduces false positives in high-stakes contexts.

4) Standardized Synthetic Content Classification

Recommendation: To aid prioritization and triaging of cases for investigative purposes, as well as develop metrics on new and emerging trends. Where feasible, providers should classify relevant material and signals as:

- GenAI confirmed or high confidence
- Suspected GenAI
- Unknown

5) Continuous Risk Monitoring and Iteration

Recommendation: Providers should continuously evaluate emerging misuse patterns and adapt safeguards accordingly, including feedback loops informed by trusted partners where appropriate.

Implementation and Collaboration

Protecting children from evolving digital threats requires sustained collaboration among state governments, law enforcement agencies, nonprofit partners, and technology providers. Effective implementation depends on continued coordination across the child protection ecosystem, including information sharing, operational feedback loops, and ongoing evaluation of emerging risks and mitigation strategies.

OpenAI supports these efforts through active participation in cross-sector initiatives focused on strengthening child safety outcomes. OpenAI is a proud industry co-chair of the Attorney General Alliance's AI Task Force, which brings together Attorneys General, law enforcement leaders, and industry partners to address emerging risks associated with artificial intelligence. OpenAI also works closely with leading child protection organizations, including the National Center for Missing & Exploited Children (NCMEC), Thorn, and the TechCoalition, to advance responsible reporting practices, prevention strategies, and collaborative responses to online child exploitation.

Continued engagement across these partnerships helps ensure that policy development, operational practices, and technological safeguards evolve alongside emerging threat patterns. Periodic evaluation, focused on reporting quality, investigative usability, and prevention effectiveness, will remain essential to maintaining durable protections as technologies and risks continue to change.

Conclusion

Generative AI is reshaping both the risks and the tools involved in protecting children online. Meeting this moment requires updated legal frameworks that provide clear enforcement authority, reporting systems that equip investigators with actionable information, and safeguards that limit opportunities for exploitation before harm occurs.

Protecting children online is a shared responsibility across government, law enforcement, nonprofit organizations, and technology providers. Effective coordination across this ecosystem will help ensure innovation supports child safety rather than outpaces existing protections. With sustained collaboration and responsible implementation, AI can become a powerful tool in preventing harm and improving protections for children online.