Safer BUILT BY THORN

AI-Generated CSAM (AIG-CSAM): Essential Considerations for Trust & Safety Teams

What is AIG-CSAM?

- AI-generated or altered images/videos of child sexual abuse
- An emerging, significant threat to child safety
- Rapidly evolving in prevalence and photorealism
- Federally illegal in the US (18 U.S.C. §1466A), state laws vary

Types of AIG-CSAM

- **AI-generated CSAM:** Media that was wholly generated by a model.
- **AI-manipulated CSAM:** Media that has some element of it altered using AIG technology.

Both types cause harm. Additionally, all AIG-CSAM could:

- Range the spectrum of photorealistic to cartoon.
- Depict an identifiable individual (deepfake), or a fully synthetic depiction (non-deepfake)
  - If an identifiable individual: Could be an identified or unidentified victim of hands-on abuse. Could also be a child that has never experienced hands-on abuse.

TREND ALERT

11%

of youth report they knew of friends or classmates who used AI tools to generate nudes of other kids.

Youth Perspectives on Online Safety, 2023

Key challenges

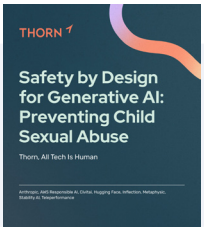
- Impedes victim identification efforts
- Creates new ways to victimize and re-victimize children at increased scale
- Reduces barriers to sexualizing minors
- Models can provide bad actors with instructions for ways and means of sexually harming children

Key recommendation: Safety by Design for Generative AI

To mitigate the risk of generative AI furthering sexual harms against children, we recommend a safety by design approach. This includes a range of principles and mitigations for responsible model development, deployment, and maintenance.

Principles Summary

- Develop responsible model development, deployment, and maintenance.
- Release and distribute generative AI models after they have been trained and evaluated for child safety, providing protections throughout the process.
- Maintain model and platform safety by continuing to actively understand and respond to child safety risks.



RESOURCE

"Safety by Design for Generative AI: Preventing Child Sexual Abuse"

Get the White Paper ⟶

Additional considerations for mitigating risks

Advanced detection methods for CSAM and text-based child sexual exploitation (CSE)

- Use hash matching to identify known CSAM in training datasets and at inputs
- Consider solutions like Thorn's Safer Predict to analyze outputs for novel CSAM or text-based CSE

Collaboration

- Participate in child safety red teaming for AI models
- Engage in standard-setting efforts (IEEE, NIST, AISI)
- Share actionable intelligence on effective interventions

User education and policy

- Raise awareness about AIG-CSAM harms and illegality
- Implement clear policies prohibiting AIG-CSAM creation/distribution
- Develop resources for users (e.g., how to respond if victimized)

Looking ahead

- Stay informed on AIG-CSAM research and trends
- Be prepared to adapt strategies as technology evolves
- Support research on deepfake victimization and related issues

Get the child safety expertise your team needs to mitigate AIG-CSAM risks.

Contact Us Today ⟶