



Protect your AI model from misuse

Online sexual harms against children are on the rise. AI-generated and AI-manipulated child sexual abuse material (CSAM) are now counted among those harms. Without child safety mitigations in place, bad actors can and do misuse generative AI technologies to harm and exploit children.

Thorn equips generative AI companies with solutions and expert guidance to mitigate misuse of their models to help prevent these online harms.

What you'll find in this guide

You'll find child safety considerations and solutions for the entire lifecycle of machine learning (ML)/AI. We've organized these by AI model stage: Development, deployment and maintenance.

Why work with Thorn?

✓ Child safety expertise

Thorn is singularly focused on protecting children from sexual abuse and exploitation in the digital age. We conduct original research and develop technology solutions to create safer online environments.

✓ Leaders in Safety by Design for generative AI

Our team convened the world's top AI companies to develop Safety by Design principles. Thorn has also collaborated with National Institute of Standards and Technology and Institute of Electrical and Electronics Engineers to integrate the principles and recommended mitigations into new and existing industry standards.

✓ Partner through the entire lifecycle

Each AI model stage presents opportunities to prioritize child safety, regardless of data modality (i.e. text, image, video, audio) or if it's released as closed or open source. Thorn can meet you where you are with expert-backed solutions and guidance.

1 in 8

teens personally know someone who has been targeted by deepfake nudes.

Source: Deepfake Nudes & Young People, Thorn

1 in 17

teens report they have been the victim of deepfake nudes.

Source: Deepfake Nudes & Young People, Thorn

TRUSTED BY

stability.ai

 OpenAI

 cantina

 Reve AI

 lenso.ai

Child safety considerations checklist

⚡ SIGNIFICANT IMPACT

🔒 TECHNOLOGY SOLUTION

🚩 CONSULTING SERVICE

Detection & removal

- ☐ Use purpose-built CSAM detection solutions for hashing and matching the data against hash sets of verified CSAM and predictive AI to identify potential CSAM and flag it for manual review ⚡🔒
- ☐ Remove harmful content from training data ⚡
- ☐ Report any verified CSAM to a reporting agency 🔒

Content segregation

- ☐ Separate child-related content from adult sexual material in training datasets
- ☐ Apply across all modalities (image, video, audio)

Safety testing

- ☐ Conduct thorough red teaming sessions ⚡🚩
- ☐ Test for CSAM-generating prompts ⚡🚩
- ☐ Identify potential safety gaps and edge cases 🚩

Technical safeguards

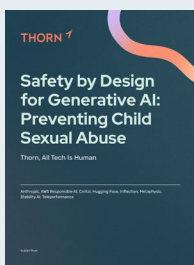
- ☐ Implement model biases against child exploitation and other harmful content (e.g., biasing the model against outputting child nudity or sexual content involving children)
- ☐ Create barriers to harmful content generation 🚩

Transparency & accountability

- ☐ Maintain transparent training set documentation (especially for open source models)
- ☐ Enable independent safety audits 🚩
- ☐ Publish safety benchmarks and evaluations

Policy development

- ☐ Establish clear training data guidelines ⚡🚩
- ☐ Define training data and model development policies with child safety focus ⚡🚩
- ☐ Document child safety-first protocols 🚩



WHITE PAPER

Safety by Design for Generative AI: Preventing Child Sexual Abuse

To mitigate the risk of your generative AI model furthering sexual harms against children, we recommend a Safety by Design approach.

[Download the white paper to learn more](#) →

How Thorn can help

Safer by Thorn

With a sole focus on child sexual abuse and exploitation, Safer by Thorn is a purpose-built solution to help generative AI companies mitigate the risks of generating CSAM or being misused to sexually exploit children.

PROTECT YOUR TRAINING DATA

- **Safer Match**'s multiple hashing methods and expansive database of verified CSAM hashes offers broad detection of known CSAM.
- **Safer Enterprise** combines multiple hashing methods along with predictive AI to provide comprehensive detection of both known and novel CSAM.
- **Safer Predict CSAM classifier** can detect potential novel CSAM in image and video training datasets.
- **Safer Predict** text classifier can help identify textual training data that can result in violative model behavior.

DID YOU KNOW...

The open-source LAION-5B dataset used to train AI image generators was [found to contain thousands of images of child sexual abuse?](#) Corrupted training data puts AI models at risk of being misused by bad actors.

Expert consultation

BUILD SAFER AI SYSTEMS WITH STRATEGIC GUIDANCE

Child safety red teaming

- Comprehensive model stress testing focused on child sexual abuse and exploitation, and offender typologies
- Expert-guided vulnerability assessments
- Actionable results and mitigation recommendations

Safety-by-Design strategy

- Early-stage safety integration
- Risk assessment frameworks
- Proactive protection measures
- Feature safety evaluation
- Team wellness planning and staff protection protocols

Policy development

- User safety policy creation
- Enforcement strategy design
- Implementation guidelines
- Best practices integration



As an organization that deeply prioritizes safety, Thorn was a natural partner to help ensure that we're building safe and responsible AI models. After three months of red teaming with their ML/AI consulting team, they provided us with invaluable insights and identified model limitations and policy gaps related to child safety.

ANTHROPIC







Child safety considerations checklist

 SIGNIFICANT IMPACT


 TECHNOLOGY SOLUTION

 CONSULTING SERVICE






Cloud-based systems

- ☐ Monitor input prompts for CSAM solicitation attempts and input image uploads for CSAM   
- ☐ Implement output scanning for CSAM  
- ☐ Deploy real-time content filtering systems 



Open-source management

- ☐ Evaluate hosting platform safety standards
- ☐ Screen for platforms known to host harmful content 



User agreements & policies

- ☐ Require explicit child safety compliance  
- ☐ Include clear terms against CSAM generation  
- ☐ Document consequences for violations 


Platform hosting standards

- ☐ Vet hosted models for safety compliance prior to hosting 
- ☐ Establish model safety requirements 

Content authentication

- ☐ Implement built-in provenance (e.g. watermarking) systems  
- ☐ Deploy synthetic content detection

Prevention & deterrence

- ☐ Display warning messages for suspicious prompts
- ☐ Implement interstitial safety notices
- ☐ Provide prevention resources and reporting options 




Child safety considerations checklist

 SIGNIFICANT IMPACT


 TECHNOLOGY SOLUTION

 CONSULTING SERVICE


Legacy model assessment

- ☐ Audit existing models for safety gaps, where necessary implement additional safety mitigations  
- ☐ Remove non-compliant legacy models (e.g. models that were explicitly built to create AIG-CSAM, and models, services/apps that are used to “nudify” images of children) 




Detection systems updates

- ☐ Maintain synthetic content detection accuracy
- ☐ Conduct regular testing against new model outputs (e.g. including details of the inputs that produced the harmful content) 
- ☐ Detect and remove from your platforms known models that were explicitly built to create AIG-CSAM, and those models, services and apps that are used to “nudify” images of children



Threat monitoring

- ☐ Partner with safety organizations 
- ☐ Track emerging misuse patterns
- ☐ Document new exploitation methods
- ☐ Share insights with the AI safety community

Reporting infrastructure

- ☐ Establish clear violation reporting channels  
- ☐ Create comprehensive and actionable reporting templates 
- ☐ Include prompt/input documentation
- ☐ Maintain efficient escalation paths
- ☐ Enable direct authority notification

Continuous improvement

- ☐ Maintain ongoing safety audits 
- ☐ Conduct regular safety performance reviews 
- ☐ Update safety protocols based on findings
- ☐ Document and share best practices
- ☐ Implement emerging safety standards

How Thorn can help

Safer by Thorn

With a singular focus on child sexual abuse and exploitation, Safer by Thorn is a purpose-built solution to help generative AI companies mitigate the risks of generating CSAM or being misused to sexually exploit children.

SAFEGUARD INPUTS AND OUTPUTS

- **Safer Predict CSAM classifier** identifies potentially novel CSAM in both inputs and outputs—creating multiple layers of defense for AI platforms and their users.
- **Safer Match** employs multiple hashing methods to detect known CSAM in image inputs to help prevent misuse of AI systems.
- **Safer Enterprise** offers comprehensive CSAM detection with a content review tool that helps moderation teams efficiently prioritize and report detected CSAM—and has wellness features built in.

Classifier specialization

Available as an add-on service for existing Safer Predict CSE text classifier implementations. We'll fine-tune our classification model to your specific needs.

CUSTOM MODEL ADAPTATION

- Optimize detection for your unique user behaviors and content patterns
- Train on your private datasets to improve accuracy
- Target platform-specific child safety concerns

Expert consultation

MAINTAIN SAFER AI SYSTEMS WITH STRATEGIC GUIDANCE

Technical implementation support

- Classifier integration and optimization
- GenAI watermarking strategy and setup
- Safety feature architecture review

Policy & process development

- Child safety policy development and enforcement planning
- Moderation workflow design
- Response protocol creation for safety incidents

Reactive support

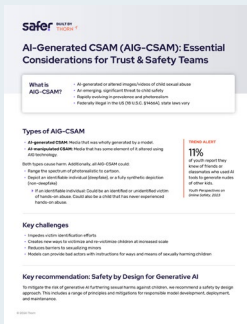
- Incident response planning
- Policy enforcement optimization
- Case-specific consultation for emerging threats
- Problem-solving for platform-specific challenges

Proactive strategy

- Regular safety policy reviews and updates
- Enforcement strategy refinement
- Trend analysis and prevention planning
- Best practice recommendations



Thorn is unique in its depth of expertise in both child safety and AI technology. The combination makes them an exceptionally powerful partner in our work to assess and ensure the safety of our models.



TRUST & SAFETY INSIGHTS

Essential Considerations: AI-Generated CSAM (AIG-CSAM)

Learn the types of challenges related to and recommendations to mitigate AIG-CSAM.

[Download](#) →



**Ready to explore expert-backed
child safety services and solutions?**

[Contact Thorn today](#) →