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.

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.

Deepfake nudes are a harmful reality for youth

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

AI MODEL STAGE: DEVELOPMENT

Child safety considerations checklist

SIGNIFICANT IMPACT | TECHNOLOGY SOLUTION | CONSULTING SERVICE

Detection & removal

- [ ] Use purpose-built CSAM detection solutions to hash files and match against hash sets of verified CSAM and predictive AI to identify suspected CSAM and flag it for review
- [ ] Remove harmful content from training data
- [ ] Report any 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.

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 UK AI Security Institute, 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.

TRUSTED BY

stability.ai OpenAI Reve lenso.ai

How Thorn can help

Safer by Thorn

With a 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, INPUTS AND OUTPUTS

- **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 suspected novel CSAM in image and video training datasets, inputs and outputs.
- **Safer Predict text classifier** can help identify textual training data that can result in violative model behavior.

> **DID YOU KNOW...**
>
> The LAION-5B and NudeNet open-source datasets, used to train AI image generators, were 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

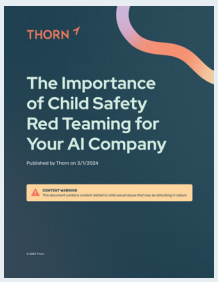
- Comprehensive model stress testing focused on child sexual abuse and exploitation, and offender typologies
- Expert-guided vulnerability assessments
- Actionable results and mitigation recommendations

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



TRUST & SAFETY INSIGHTS

The Importance of Child Safety Red Teaming for Your AI Company

Learn about Thorn's child safety red teaming service, which helps mitigate the misuse of generative AI technologies.

AI MODEL STAGE: DEPLOYMENT

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

> *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.*

ANTHROP\C

AI MODEL STAGE: MAINTENANCE

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

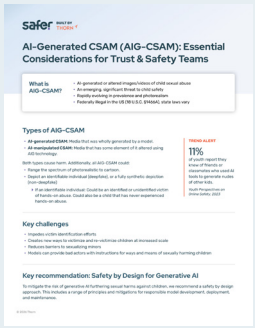
- [ ] Establish clear violation reporting channels
- [ ] Create comprehensive and actionable reporting templates
- [ ] Include prompt/input documentation
- [ ] Maintain efficient escalation paths
- [ ] Enable direct authority notification

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

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.

OpenAI



TRUST & SAFETY INSIGHTS

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

Learn more about AIG-CSAM and key recommendations for mitigation.

Get a copy of this guide, scan here.

THORN