

Safety by Design for Generative AI

Preventing Child Sexual Abuse

Dr. Rebecca Portnoff | VP Data Science & AI, Thorn
rebecca@wearethorn.org | <https://www.linkedin.com/in/dr-rsportnoff/>

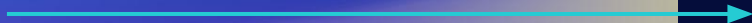
THE SCALE

In just two decades,
reports of suspected
CSAM files have
exploded by **over**
13,400%



2004

450,000 files



2024

61 million+ files

New abuse imagery; more extreme abuse imagery

40% minors

experienced an immediate outreach from a stranger soliciting nude images¹

22% increase

in category A content reported to IWF from 2022 to 2023²

1. Thorn. (2022). Online Grooming: Examining risky encounters amid everyday digital socialization, 2021. https://info.thorn.org/hubfs/Research/2022_Online_Grooming_Report.pdf

2. IWF. (2024). Annual Report, 2023. <https://www.iwf.org.uk/news-media/news/iwf-declares-a-record-year-for-online-child-sexual-abuse-reports-at-annual-report-2023-launch/>

More than images

812 reports

of sexual extortion submitted on average per week to NCMEC in the last year¹

140% increase

in the number of cases reported to NCMEC involving a child in imminent danger since 2021²

1. Thorn and National Center for Missing and Exploited Children (NCMEC). (2024). Trends in Financial Sextortion: An investigation of sextortion reports in NCMEC CyberTipline data. https://info.thorn.org/hubfs/Research/Thorn_TrendsInFinancialSextortion_June2024.pdf

2. NCMEC. (2024). 2023 CyberTipline Report. <https://www.missingkids.org/content/dam/missingkids/pdfs/2023-CyberTipline-Report.pdf>

Harms of Generative AI

We know that these technologies are being misused today to create sexual content depicting children, and further sexual harms against children.

Complicate Victim Identification:

- Models can generate photorealistic CSAM, at scale
- AIG-CSAM adds to the already massive number of reports, making victim identification more difficult

Increase Re-Victimization:

- Models fine-tuned on existing CSAM can generate more abuse material
- For survivors, distribution of their abuse content exacerbates trauma

Reduce barriers to harm:

- Models generate sexual imagery from benign content
- Models can be used to scale sexual extortion, bully/harass peers

New abuse imagery; more extreme abuse imagery

1 in 10

minors know of cases where their peers use generative AI to create explicit images of other kids¹



10% increase

in category A content assessed by IWF from 2023 to 2024²

1. Thorn. (2024). Youth Perspectives on Online Safety, 2023. https://info.thorn.org/hubfs/Research/Thorn_23_YouthMonitoring_Report.pdf

2. IWF. (2024). AI and the production of child sexual abuse imagery. <https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery/>

More than images

11% of reports

of sexual extortion to NCMEC in the last year (in which tactics were apparent), included threatening children with fake sexual imagery¹

50% of LE

encountered AI-generated CSAM used for online grooming of minors²

1. Thorn and National Center for Missing and Exploited Children (NCMEC). (2024). Trends in Financial Sextortion: An investigation of sextortion reports in NCMEC CyberTipline data. https://info.thorn.org/hubfs/Research/Thorn_TrendsInFinancialSextortion_June2024.pdf

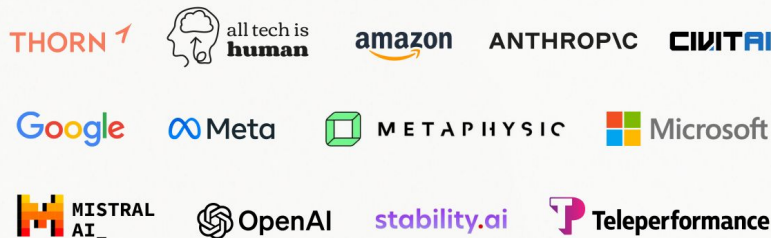
2. Centre for AI and Robotics at UNICRI. Generative AI: A New Threat for Online Child Sexual Exploitation and Abuse, 2024.

Safety by Design

Safety by Design means that companies must prioritize child safety across the entire ML/AI life cycle of development, deployment, and maintenance.

<https://www.thorn.org/blog/generative-ai-principles/>

Generative AI Principles to Prevent Child Sexual Abuse



Safety by Design Principles

DEVELOP

Develop, build and train generative AI models that proactively address child safety risks.

- ✓ Responsibly source our training datasets, and safeguard them from child sexual abuse material (CSAM) and child sexual exploitation material (CSEM)
- ✓ Incorporate feedback loops and iterative stress-testing strategies in our development process
- ✓ Employ content provenance with adversarial misuse in mind

THORN 



DEPLOY

Release and distribute generative AI models after they have been trained and evaluated for child safety, providing protections throughout the process.

- ✓ Safeguard our generative AI products and services from abusive content and conduct
- ✓ Responsibly host models
- ✓ Encourage developer ownership in safety by design

THORN 



MAINTAIN

Maintain model and platform safety by continuing to actively understand and respond to child safety risks.

- ✓ Prevent our services from scaling access to harmful tools
- ✓ Invest in research and future technology solutions
- ✓ Fight CSAM, AIG-CSAM and CSEM on our platforms

THORN 





















Safety by Design White Paper
















What's in the paper:

1. Safety by Design principles
2. Recommended mitigations to enact the principles
3. An assessment of each mitigation
4. Further opportunities & potential downstream implications


















Mitigations

		AI Developers	AI Providers	Data Hosting Platforms	Social Platforms	Search Engines
1	Responsibly source your training data   	✓				
2	Detect, remove and report CSAM and CSEM from your training data   	✓		✓		
3	Separate depictions/representations of children from adult sexual content in your image, video or audio generation training datasets  	✓		✓		
4	Conduct red teaming for AIG-CSAM and CSEM   	✓				
5	Include content provenance by default   	✓				
6	Define specific training data and model development policies  	✓				
7	Prohibit customer use of your model to further sexual harms against children  	✓	✓			

Mitigations

		AI Developers	AI Providers	Data Hosting Platforms	Social Platforms	Search Engines
1	Detect abusive content (CSAM, AIG-CSAM, and CSEM) in inputs and outputs  	✓	✓			
2	Include user reporting, feedback or flagging options   		✓			
3	Include an enforcement mechanism  	✓				
4	Assess generative models before access   		✓			
5	Include prevention messaging for CSAM solicitation 	✓	✓			
6	Incorporate phased deployment  	✓				
7	Incorporate a child safety section into model cards  	✓	✓			

Mitigations

		AI Developers	AI Providers	Data Hosting Platforms	Social Platforms	Search Engines
1	Remove services for “nudifying” images of children from search results 					✓
2	When reporting to NCMEC, use the Generative AI File Annotation   	✓	✓	✓	✓	
3	Detect and remove from your platforms known models that were explicitly built to create AIG-CSAM  		✓		✓	✓
4	Retroactively assess currently hosted generative models, updating them with mitigations in order to maintain platform access  	✓	✓			
5	Detect, report, remove and prevent CSAM, AIG-CSAM and CSEM on your platforms 				✓	
6	Invest in tools to protect content from AI-generated manipulation   	✓			✓	
7	Maintain the quality of your mitigations   	✓	✓	✓	✓	✓
8	Disallow the use of generative AI to deceive others for the purpose of sexually harming children. Explicitly ban AIG-CSAM from your platforms.				✓	
9	Leverage Open Source Intelligence (OSINT) capabilities  	✓	✓		✓	

Deep Dive: Challenges of Assessing Models



Missing scalable solutions



Lack of standardized evaluation data sets



Legal ambiguity with red teaming for images/videos




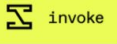

Reliance on prompting strategies

Phase 2

Adoption and Accountability

1. Progress Reports
2. Standards
3. Policy

Progress Reports

		 CIVITAI	 invoke	 METAPHYSIC
DEVELOP	Sub-principle 1	Not applicable	Not applicable	No current gaps observed
	Sub-principle 2	Not applicable	Not applicable	Not started
	Sub-principle 3	Not started	Some progress	Some progress
DEPLOY	Sub-principle 1	No current gaps observed	Some progress	No current gaps observed
	Sub-principle 2	Some progress	Not applicable	Some progress
	Sub-principle 3	Not started	Not applicable	No current gaps observed
MAINTAIN	Sub-principle 1	Some progress	No current gaps observed	Not applicable
	Sub-principle 2	No current gaps observed	No current gaps observed	No current gaps observed
	Sub-principle 3	Some progress	Some progress	Some progress

Thorn's blog:

<https://www.thorn.org/blog/safety-by-design-for-generative-ai-3-month-progress-report/>

Full report:

<https://info.thorn.org/hubfs/Thorn-SafetybyDesign-ThreeMonthProgressReport-3.pdf?>

Standards

IEEE SA

**STANDARDS
ASSOCIATION**

NIST

**National Institute of
Standards and Technology**
U.S. Department of Commerce

IEEE Recommended Practice for Using
Safety by Design in Generative Models to
Prioritize Child Safety:

<https://standards.ieee.org/ieee/3462/11584/>

NIST AI 100-4: Reducing Risks Posed by
Synthetic Content:

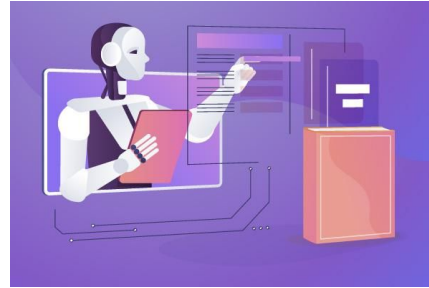
<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf>

Policy Engagement



General-Purpose
AI Code of
Practice

Transparency Chapter



Safety & Security Chapter



To have lasting change, we must take ownership of what we build, and recognize that we are developing that double-edged sword.

Thank You!

Dr. Rebecca Portnoff | VP Data Science & AI, Thorn
rebecca@wearethorn.org | <https://www.linkedin.com/in/dr-rsportnoff/>