

GENERATIVE AI AND CHILD SEXUAL ABUSE: SAFETY BY DESIGN

Michael Simpson
Staff Data Scientist at Thorn

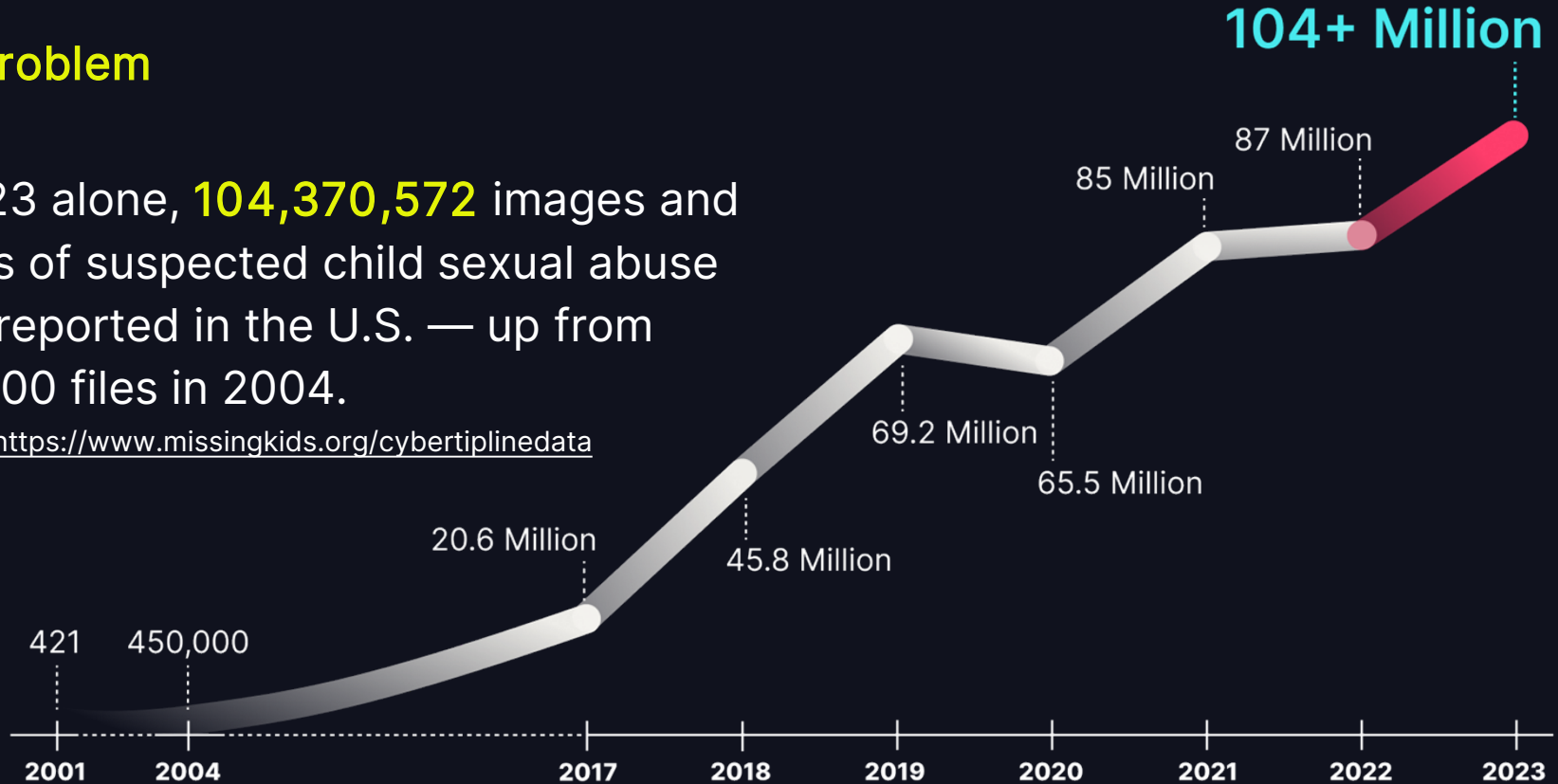
THORN BUILDS
TECHNOLOGY TO
DEFEND CHILDREN
FROM SEXUAL
ABUSE.

SAFETY BY DESIGN

The Problem

In 2023 alone, **104,370,572** images and videos of suspected child sexual abuse were reported in the U.S. — up from 450,000 files in 2004.

Source: <https://www.missingkids.org/cybertiplinedata>



SAFETY BY DESIGN

Emerging Technology

Innovative bursts in technology are a double edged sword. Bad actors misuse technology to harm children; we use the best of tech to defend children.



Improve Existing Solutions

- Update models with newer deep learning architectures



Understand Misuse

- Surveys and analysis
- Tracking trends and measuring prevalence



Respond

- New technology and product solutions
- Safety by Design

SAFETY BY DESIGN

AI for Good



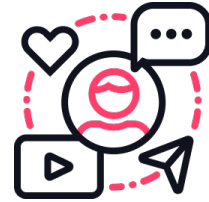
ACCELERATE

victim identification



STOP

revictimization



PREVENT

abuse

SAFETY BY DESIGN

Issue Impact of Generative AI

The child safety ecosystem is already over taxed.

Misuse of generative AI accelerates harms across victim identification, re-victimization and prevention.



Victim Identification

- Models generate photorealistic child sexual abuse material, at scale
- Adding to the haystack: makes victim identification more difficult



Re-Victimization

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



Prevention

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

Source: <https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-ai.pdf>

SAFETY BY DESIGN

IMAGE GENERATION

Figure from:
*Generative ML
and CSAM:
Implications and
Mitigations*



A woman generated with a popular Stable Diffusion model.



The same prompt, but with a LoRA to make the output moderately resemble Audrey Hepburn.



Addition of textual inversion, a learned phrase representing a concept, to make the resulting character appear younger.

SAFETY BY DESIGN

More Than Images

Predators are using AI to master child personas, improve grooming outcomes, and generate graphic descriptions of child sexual abuse

2 in 5

of all kids have been approached online by some who they thought was attempting to “befriend and manipulate” them.



[Online Grooming: Examining risky encounters amid everyday digital socialization](#)

SAFETY BY DESIGN

Now is the moment for **Safety by Design**

Prioritize the child across the entire machine learning/AI lifecycle of **develop, deploy and maintain.**



**Generative AI
Principles to Prevent
Child Sexual Abuse**

THORN ¹ all tech is human amazon ANTHROPIC CIVITAI

Google Meta METAPHYSIC Microsoft

MISTRAL AI_ OpenAI stability.ai Teleperformance

A dark blue banner with white text and logos. The logos are arranged in three rows. The first row contains THORN, all tech is human, amazon, ANTHROPIC, and CIVITAI. The second row contains Google, Meta, METAPHYSIC, and Microsoft. The third row contains MISTRAL AI_, OpenAI, stability.ai, and Teleperformance.

SAFETY BY DESIGN

The Principles

DEVELOP

Develop, build and train generative AI model 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, data source tracking, with adversarial misuse in mind

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


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
















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  	✓		✓		

DEPLOY

 SIGNIFICANT IMPACT


 OPEN SOURCE

 CLOSED SOURCE

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  	✓	✓			










































MAINTAIN

 SIGNIFICANT IMPACT

 OPEN SOURCE

 CLOSED SOURCE

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  					



SAFETY BY DESIGN

IEEE Standardization

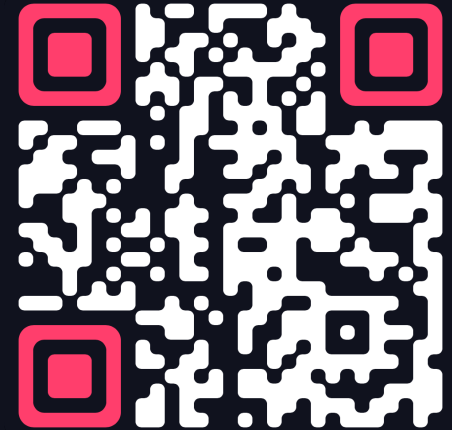
We just formed a working group to standardize the recommended practice with .
If you are interested in this project you should **join!**



SAFETY BY DESIGN

Next Steps

- Put safety at the forefront of your work
- Adopt these principles at your organization
- Get involved with the standardization project
- Follow Thorn's work



Thank You

