THORN

Safety by Design for Generative AI: Preventing Child Sexual Abuse

Thorn, All Tech Is Human

AWS AI, Civitai, Hugging Face, Inflection, Metaphysic, Stability AI, Teleperformance

© 2024 Thorn

Table of Contents

© 2024 Thorn

Executive Summary

We are at a crossroads with generative artificial intelligence (AI). Creating content at scale is easier now than ever before. In the same way that offline and online sexual harms against children have been accelerated by the internet [1], misuse of generative AI has profound implications for child safety, across victim identification, victimization, prevention and abuse proliferation. This misuse, and its associated downstream harm, is already occurring, and warrants collective action, today. The need is clear: we must mitigate the misuse of generative AI technologies to perpetrate, proliferate, and further sexual harms against children. This moment requires a proactive response.

Now is the time for Safety by Design [2, 3]. For generative AI, this concept should be expanded to the entire lifecycle of machine learning (ML)/AI from the earliest stages: development, deployment, and maintenance. Each part in the process includes opportunities to prioritize child safety, regardless of data modality (i.e. text, image, video, audio) or if an organization releases its technology as closed source or open source, or some release option between these two. When considering the ecosystem of ML/AI technology players, we further see multiple points of opportunity to prioritize child safety via Safety by Design. Whether you are an AI Developer, AI Provider, Data Hosting Platform, Social Platform or Search Engine, you can minimize the possibility of generative AI being misused to further sexual harms against children.

This paper serves as a resource for a set of co-defined Safety by Design principles and mitigations for generative AI in the context of child sexual abuse. They are written to align and build off of existing guidance and commitments [2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. They are written to be tactical, actionable, and multidisciplinary. These principles and mitigations are written such that a technical, policy, product, or trust and safety team could enact them with minimal friction.

© 2024 Thorn

Issue Overview

Offline and online sexual harms against children have been accelerated by the internet [1]. The child safety ecosystem is already overtaxed. In 2022 reports to the National Center for Missing and Exploited Children (NCMEC) contained over 88 million files of CSAM and other files related to child sexual exploitation; in 2023, reports contained over 100 million such files [12].

The internet has its roots in information sharing. Starting in the 1960s, government researchers explored the use of network computing to share information. The internet as it currently exists grew out of efforts to standardize communication protocols between different networks. Cybersecurity concepts of preventing adversarial misuse of the internet were embedded throughout the ideation and development of the internet. However, similar concepts around preventing adversarial misuse of the internet to scale harms against children did not enter the broader conversation until the 1990s, with debates around the Communications Decency Act [13]. The consequences of this are self apparent: the internet turned the problem of child sexual abuse material (CSAM) from one that was more contained (limited to small networks via the postal service) to where we are today, where platforms, hosting services, internet service providers and more all face the reality that CSAM can - and does - circulate on their services. There are many organizations that have pursued transparent and collaborative governance of their service, by preventing, detecting, removing and reporting CSAM. Yet the fact remains that this work continues to be an uphill battle.

We are at another, similar crossroads with generative AI. Using this technology, human-like text, photorealistic images and videos, music, art and other content can be automatically generated. It is a straightforward task for a person to use these models to create content at scale, irrespective of their technical expertise. This technology unlocks the ability for a single human to easily create and distribute millions of pieces of content.

In this moment, generative AI holds the potential for numerous benefits to consumers in diverse applications. These benefits extend to improving child safety protections: e.g. existing detection technologies can be updated to use new deep learning architectures, while automatic image and text summarization can accelerate prioritization and triage.

However: misuse of this same technology has profound implications across victim identification, victimization, prevention and abuse proliferation.

Looking at each of these separately, misuse of generative AI technologies:

- **Impedes victim identification**

  Bad actors use generative AI to create AI-generated child sexual abuse material (AIG-CSAM) [14, 15, 16]. Models that the actors have access to - broadly shared models that were trained on minimally curated datasets - are misused by bad actors to create AIG-CSAM. Victim identification is already a needle in the haystack problem for law enforcement: sifting through huge amounts of content to find the child in active harm's way. The expanding prevalence of AIG-CSAM is growing that haystack even further, making victim identification more difficult.

© 2024 Thorn

- **Creates new ways to victimize and re-victimize children**
  This same technology is used to newly victimize children, as bad actors can now easily generate new abuse material of children, and/or sexualize benign imagery of a child. Bad actors use this technology to perpetrate re-victimization using primarily broadly shared models and fine-tuning them on existing child abuse imagery to generate additional explicit images of these children [14, 16]. They collaborate to make these images match the exact likeness of a particular child, but produce new poses, acts and egregious content like sexual violence. These images depict both identified and unidentified survivors of child sexual abuse. Bad actors also use this technology to scale their grooming and sexual extortion efforts, using generative AI to scale the creation of content necessary to target a child [17]. This technology is further used in bullying scenarios, where sexually explicit AI-generated imagery of children is being used by children to bully and harass others [18, 19, 20].

- **Reduces social and technical barriers to sexualizing minors**
  The ease of creating AIG-CSAM, and the ability to do so without the victim's involvement or knowledge, may perpetuate the misconception of this content being "harmless". Bad actors use this technology to produce AIG-CSAM and other sexualising content of children, as well as to engage in fantasy sexual role-play with generative AI companions who mimic the voice of children [26]. Research suggests bad actors viewing CSAM may have their fantasies reinforced [21] by viewing abuse imagery and may be at a heightened risk for committing hands on abuse acts [22].

- **Enables information sharing for abuse proliferation**
  Bad actors use generative AI models (particularly text or image editing) in abuse proliferation [14]. Models can support bad actors by providing instruction for hands-on sexual abuse of a child, information on coercive control, details on destroying evidence and manipulating artifacts of abuse, or advice on ensuring victims don't disclose.

This misuse, and its associated downstream harm, is already occurring, and warrants collective action, today. The need is clear: we must mitigate the misuse of generative AI technologies to perpetrate, proliferate, and further sexual harms against children. This moment requires a proactive response. The prevalence of AIG-CSAM is small, but growing [14, 16]. Now is the time to act, and put child safety at the center of this technology as it emerges. Now is the time for Safety by Design [2, 3].

© 2024 Thorn

Response

Rather than retrofitting safeguards after an issue has occurred, Safety by Design requires technology companies to consider how to minimize threats and harms throughout the design, development and deployment process [3]. For generative AI, this concept should be expanded to the entire lifecycle of ML/AI from the earliest stages: develop, deploy, and maintain. Each part in the process includes opportunities to prioritize child safety, regardless of data modality or if an organization releases its technology as closed source or open source.

When considering the ecosystem of ML/AI technology players, we further see multiple points of opportunity to prioritize child safety via Safety by Design. Whether you are an AI Developer, AI Provider, Data Hosting Platform, Social Platform or Search Engine, you can join the effort to minimize the possibility of generative AI being misused to further sexual harms against children.

This paper serves as a resource for a set of co-defined Safety by Design principles for generative AI in the context of child sexual abuse. They are written to align and build off of existing guidance and commitments [2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. They are written to be tactical, actionable, and multidisciplinary. The mitigations are written such that a technical, policy, product, or trust and safety team could enact them with minimal friction.

Notably, while the focus of this document is on mitigating the misuse of generative AI in the context of child sexual abuse, similar misuse may occur in other harm spaces, including terrorism, violence and extremism, mis/disinformation, and adult non-consensual intimate imagery [23, 24, 25]. We encourage readers to read the recommended principles and mitigations with an eye towards how they may be applicable in other harm contexts.

© 2024 Thorn

Limitations

While this document attempts to be comprehensive across several key verticals (data modalities, open and closed source, the ecosystem of ML/AI technology players and the full lifecycle of ML/AI), there are limitations to this work.

First, the resources provided and regulatory understanding outlined in this document lean towards a Global North perspective. We encourage readers to layer on their unique perspective and additional resources when engaging with this document, and acknowledge that some of the resources in the document may not be accessible to readers in certain locations.

Second, while the focus of this document is what the ecosystem of ML/AI technology players can do, the technology ecosystem extends beyond those named in this document, also including players like app stores, domain registrars and GPU providers. We prioritized the ecosystem of ML/AI technology players in this document in order to focus on those organizations who have the most opportunity for upstream, scaled impact at this moment. We encourage readers to ideate on and implement Safety by Design across these other settings.

Finally, the reality is the child safety ecosystem is much broader than just technology players. These principles and mitigations should be understood as one piece of a necessary ecosystem response and holistic approach. To be most effective, this approach will require layered sets of interventions. Stakeholders across non-governmental organizations (NGOs), law enforcement agencies, government bodies, survivor services, and the broader community must coordinate together to have impact, collaboratively developing a victim-centered, preventative approach.

© 2024 Thorn

Definitions

Here, we define several terms referenced in this report.

AI developer

The individuals and organizations that build generative AI technology.

> Examples of organizations that currently develop open source technology: Cerebras, Databricks, Meta, Nomic, OpenAI, Openjourney, Stability AI, Google.

> Examples of organizations that currently develop closed source technology: Anthropic, Inflection, Metaphysic, OpenAI, Meta, Google.

AI-generated child sexual abuse material (AIG-CSAM)

Visual depiction (image/video) of sexually explicit conduct involving a minor, the creation of which has been facilitated by generative AI technologies. This may range from a fully generated image/video, to generated elements applied to a pre-existing image/video.

AI provider

The individuals and organizations that provide a platform for hosting ML/AI models. A "first-party" provider hosts only the models developers within their organization build; a "third-party" provider hosts models built by external developers. Some organizations may act as both a first and third-party provider of models.

> Examples of organizations that currently provide a platform for open source technology: Civitai, Github, Hugging Face, Sourceforge, Google.

> Examples of organizations that currently provide a platform for closed source technology: Anthropic, Inflection AI, Metaphysic, OpenAI, Google.

Adult sexual content

Images, videos, and audio that is pornographic, or primarily depicts explicit sexual acts, containing only adults. ***Additional context:*** the definition of what constitutes pornographic content is highly context-dependent, and content should be assessed in a way that acknowledges this context so as to avoid exacerbating discrimination of already marginalized groups.

Broadly shared models

Models (e.g. the trained model weights, checkpoints, LoRAs, etc.) that have been shared and circulated across the internet. Includes but is not exclusive to open source.

Closed organization

Organizations and institutions that develop, deploy, maintain, or host closed source generative AI technologies.

Closed source

Software where the source code and model weights are not publicly available. The rights to use, modify and distribute the software are restricted by the terms of a license. Access to the underlying code (including model weights, in the ML/AI context) is limited to the software's authors or a select group.

© 2024 Thorn

Content provenance

Facts about the origins of a piece of digital content, such as who created it and how, as well as its history of edits.

Child safety policy

Child safety policies include a portfolio of trust and safety policies (some of which cover illegal behaviors or content) created to mitigate the risk of online harms specific to minors. In this document, it refers to policies covering CSAM, child sexualization and abuse, child grooming, etc.

Child sexual abuse material (CSAM)

Visual depiction (image/video) of sexually explicit conduct involving a minor. Does not require that the material depict a child engaging in sexual activity. Covers lewd and lascivious content, as well as content with a focus on genitalia. N.B. The definition of minor will vary depending on your legal jurisdiction.

Child sexual exploitation material (CSEM)

Throughout this document, we use this phrase as a shorthand for the full list of: image/video/audio content sexualising children, grooming text, sexual extortion text, CSAM advertising, CSAM solicitation, and text promoting sexual interest in children.

CSAM advertising

Noting where child sexual abuse material can be found. This may be a URL, or advertisements of CSAM for sale.

CSAM solicitation

The act of requesting, seeking out or asking for access to, or the location of, child sexual abuse material.

Data hosting platform

The individuals and organizations that provide a platform for hosting data that can be used for training models, and/or provides access to existing datasets.

> Examples: Common Crawl, GitHub, Hugging Face, LAION, Papers with Code, S3, GCS, R2, Azure Blob Storage, Dropbox, Google Drive.

De-aging

Visual effects technique used to make the person depicted in an image or video look younger.

Detect

The method or act of scanning through a larger set of data to attempt to identify the target material (e.g. CSAM or CSEM). Can include both manual and automated methodologies.

Develop

Research and development to build the desired ML/AI model.

Deploy

The method or act of integrating a ML/AI model into a production environment; the method or act of making a ML/AI model available for use.

© 2024 Thorn

Fine-tuning

The method or act of customizing a pre-trained model to perform specific tasks or manifest specific behaviors.

Graphics Processing Unit (GPU) provider

The individuals and organizations that offer cloud computing for ML/AI model training.

> Examples: AWS, Google, IBM, NVIDIA.

Grooming

The act of establishing a trusted relationship with a child to prepare them for abuse and reduce the likelihood of them seeking help.

Machine Learning / Artificial Intelligence (ML/AI)

The field of study that gives computers the ability to learn without explicitly being programmed, and/or imitate intelligent human behavior.

Maintain

The act of maintaining the quality of ML/AI models in the face of data drift and changing landscape.

Model

Software that has been trained to recognize patterns, make predictions, or generate new content.

Model card

A short document that provides information about a ML/AI model.

Open organization

Organizations and institutions that develop, deploy, maintain, distribute, or host open source and broadly shared generative AI technologies.

Open source

Software for which the source code is available to the public. It is accompanied by a license that allows users to view, modify, and redistribute the source code (including model weights, in the ML/AI context).

Platform

Any hardware or software used to host an application or service.

Promotion of sexual interest in children

Content that aims to reduce the societal stigma around abusive and exploitative sexual interactions with children. E.g. forum conversations on sites dedicated to child sexual abuse.

Precision

One metric to measure a model's performance. Refers to the number of true positives divided by the total number of positive predictions. Generally paired with recall (see below).

© 2024 Thorn

Recall

One metric to measure a model's performance. Refers to the number of true positives divided by the total number of all positive instances (regardless of the model's prediction). Generally paired with precision (see above).

Red teaming

The practice of stress testing systems - physical or digital - to find flaws, weaknesses, gaps and edge cases. N.B. In this document, when we refer to "safety assessment" it is distinguished from red teaming vis-à-vis the stage in the ML/AI process (develop, deploy, maintain) in which it occurs, as well as the intended purpose.

Re-victimization

The furthering of trauma experienced by victims of child sexual abuse when a victim faces any sexual abuse or assault subsequent to a first abuse or assault. This can include recirculation of original abuse imagery, development of novel images using the child's likeness, and stalking (online and off), among other experiences.

Safety assessment

Evaluating whether a model has passed a predetermined criteria regarding its propensity to generate images, videos, text and audio that scales sexual harms against children, covering both AIG-CSAM and other CSEM.

Search engine

A software system that searches for and identifies items in a database that corresponds to the terms specified by the user, used for finding particular sites on the World Wide Web.

> Examples: Google, Bing, Yahoo, Yandex, DuckDuckGo.

Sexual extortion

Threatening to distribute private and sensitive material featuring the child unless the child complies with some type of demand. Demands can involve a range of items, including but not limited to sharing additional images of a sexual or intimate nature, sexual favors, or money.

Social platform

A digital service that uses the internet to facilitate interactions (e.g. content sharing) between two or more separate but interdependent users.

> Examples: Facebook, Instagram, Snapchat, Reddit, TikTok, X, YouTube, Google.

Training

The method or act of fitting a combination of weights to a model, such that the model can perform a specific task or generate specific content.

Victim identification

The act of investigating CSAM to work out information about the crime depicted in the content, specifically who the victims depicted in the content are, so that they can be found and recovered.

> Examples of institutions that conduct victim identification efforts: NCMEC, Internet Crimes Against Children task force, Task Force Argos.

Watermarking

The act of incorporating visible or invisible indicators within a piece of digital content to tie that content to the source of the content.

© 2024 Thorn

Principles

In this section, we define foundational principles for building generative AI to prevent the misuse of generative AI technologies to perpetrate, proliferate, and further sexual harms against children.

DEVELOP

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

- **Responsibly source your training datasets, and safeguard them from CSAM and CSEM**
  This is essential to helping prevent generative models from producing AIG-CSAM and CSEM. The presence of CSAM and CSEM in training datasets for generative models is one avenue in which these models are able to reproduce this type of abusive content. For some models, their compositional generalization capabilities further allow them to combine concepts (e.g. adult sexual content and non-sexual depictions of children) to then produce AIG-CSAM. Avoid or mitigate training data with a known risk of containing CSAM and CSEM. Detect and remove CSAM and CSEM from your training data, and report any confirmed CSAM to the relevant authorities. Address the risk of creating AIG-CSAM that is posed by having depictions of children alongside adult sexual content in your video, images and audio generation training datasets.

- **Incorporate feedback loops and iterative stress-testing strategies in your development process**
  Continuous learning and testing to understand a model's capabilities to produce abusive content is key in effectively combating the adversarial misuse of these models downstream. If you don't stress test your models for these capabilities, bad actors will do so regardless. Conduct structured, scalable and consistent stress testing of your models throughout the development process for their capability to produce AIG-CSAM and CSEM within the bounds of law, and integrate these findings back into model training and development to improve safety assurance for your generative AI products and systems.

- **Employ content provenance with adversarial misuse in mind**
  Bad actors use generative AI to create AIG-CSAM. This content is photorealistic, and can be produced at scale. Victim identification is already a needle in the haystack problem for law enforcement: sifting through huge amounts of content to find the child in active harm's way. The expanding prevalence of AIG-CSAM is growing that haystack even further. Content provenance solutions that can be used to reliably discern whether content is AI-generated will be crucial to effectively respond to AIG-CSAM. Develop state of the art media provenance or detection solutions for your tools that generate images and videos. Deploy solutions to address adversarial misuse, such as considering incorporating watermarking or other techniques that embed signals imperceptibly in the content as part of the image and video generation process, as technically feasible.

© 2024 Thorn

DEPLOY

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

- **Safeguard your generative AI products and services from abusive content and conduct**

  Generative AI products and services empower users to create and explore new horizons. These same users deserve to have that space of creation be free from fraud and abuse. Combat and respond to abusive content (CSAM, AIG-CSAM and CSEM) throughout your generative AI systems, and incorporate prevention efforts. Users' voices are key: incorporate user reporting or feedback options to empower these users to build freely on your platforms.

- **Responsibly host your models**

  As models continue to achieve new capabilities and creative heights, a wide variety of deployment mechanisms manifests both opportunity and risk. Safety by design must encompass not just how your model is trained, but how your model is hosted. Responsibly host your first-party generative models, assessing them e.g. via red teaming or phased deployment for their potential to generate AIG-CSAM and CSEM, and implementing mitigations before hosting. Also responsibly host third-party models in a way that minimizes the hosting of models that generate AIG-CSAM. Have clear rules and policies around the prohibition of models that generate child safety violative content.

- **Encourage developer ownership in safety by design**

  Developer creativity is the lifeblood of progress. This progress must come paired with a culture of ownership and responsibility. Encourage developer ownership in safety by design. Endeavor to provide information about your models, including a child safety section detailing steps taken to avoid the downstream misuse of the model to further sexual harms against children. Support the developer ecosystem in their efforts to address child safety risks.

MAINTAIN

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

- **Prevent your services from scaling access to harmful tools**

  Bad actors have built models specifically to produce AIG-CSAM, in some cases targeting specific children to produce AIG-CSAM depicting their likeness. They also have built services that are used to "nudify" content of children, creating new AIG-CSAM. This is a severe violation of children's rights. Remove from your platforms and search results these models and services.

© 2024 Thorn

- **Invest in research and future technology solutions**

  Combating child sexual abuse online is an ever-evolving threat, as bad actors adopt new technologies in their efforts. Effectively combating the misuse of generative AI to further child sexual abuse will require continued research to stay up to date with new harm vectors and threats. For example, new technology to protect user content from AI manipulation will be important to protecting children from online sexual abuse and exploitation. Invest in relevant research and technology development to address the use of generative AI for online child sexual abuse and exploitation. Seek to understand how your platforms, products and models are potentially being abused by bad actors. Maintain the quality of your mitigations to meet and overcome the new avenues of misuse that may materialize.

- **Fight CSAM, AIG-CSAM and CSEM on your platforms**

  Fight CSAM online and prevent your platforms from being used to create, store, solicit or distribute this material. As new threat vectors emerge, meet this moment. Detect and remove child safety violative content on your platforms. Disallow and combat CSAM, AIG-CSAM and CSEM on your platforms, and combat fraudulent uses of generative AI to sexually harm children.

© 2024 Thorn

Recommended Mitigations

N.B. None of the resources provided in this document should be considered an endorsement of that resource. We are including these resources in order to empower teams to quickly take action on these mitigations, rather than be paralyzed by the complexity of the issues at hand.

In this section, we outline a set of recommended mitigations to enact these principles, in order to make it more difficult for bad actors to misuse generative AI technology to create AIG-CSAM and CSEM.

For each recommended mitigation, we outline:

- The intended impact of the mitigation, and an estimate of the expected impact of the mitigation, either "incremental" (a positive step) or "significant" (a game-changing intervention).
- The scope of the mitigation, either "narrow" (can be actioned on independently by a single organization), "medium" or "wide" (may require coordination and collaboration between multiple organizations or stakeholders)
- Variables that can be scaled up or down when actioning on that mitigation
- Whether it is relevant to
    - AI Developers, AI Providers, Data Hosting Platforms, Social Platforms and/or Search Engines
    - Closed and/or Open organizations
- Resources to make actioning on the mitigation more straightforward, either "informational" (intended to provide additional background information or serve as an example of the mitigation) or "implementation" (intended to highlight possible solutions and solution providers)

The relevance for each mitigation is included to support readers in assessing which mitigations are appropriate for their organization (or across their different services). The impact and scope for each mitigation are included to provide guidance in how the relevant mitigations may be ideally sequenced and actioned on. To that end, within each section of develop, deploy and maintain, the recommended mitigations have been listed according to the estimate of expected impact of the mitigation and the scope of the mitigation. This allows for the reader to begin with the most significant and narrow mitigations, then work their way down to the incremental and wide mitigations.

In addition to the recommended mitigations, we also note further opportunities beyond these mitigations. These further opportunities are carved out as distinct from the recommended mitigations, primarily because the coordination, technology and research required to action those sections are still nascent, or out of reach for smaller organizations - representing opportunities for the future, rather than recommendations that are immediately actionable in the near term.

Finally, we recognize that some of the recommended mitigations below have potentially problematic downstream implications, beyond their intended positive impact. We see opportunities to navigate these potential implications, such that the positive impact of the recommendations is assured, and the risk for the implications is minimized. To that end, we include for each set of mitigations some of these potential problematic implications, as well as our suggested response.

© 2024 Thorn

Develop

As defined above, **develop** refers to the research and development necessary to build the desired ML/AI model.

Mitigations

SIGNIFICANT IMPACT | OPEN SOURCE | CLOSED SOURCE

DEVELOP OVERVIEW	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	✓	✓			

DEVELOP: MITIGATION #1

Responsibly source your training data

As a developer, you should know what dataset sources you are using and responsibly source your training data. Avoid ingesting into your training data, any data that have a known risk (as identified by relevant experts in the space) of containing CSAM and CSEM, e.g. by removing from or disallowing in your data collection pipeline sources known for proliferating CSAM. Thoroughly document procedures to substantiate that datasets do not contain CSAM, as well as processes for reporting and preserving such materials, where required and/or permitted by law.[1] Train employees involved in model training on those procedures.

1 Developers which are electronic communication services providers (ECSs) or providers of remote computing services (RCSs) have both preservation and reporting obligations under U.S. federal law, 18 USC § 2258A. Developers who do not have reporting and preservation obligations should consider all of the applicable risks and adopt appropriate policies based on those risks. We would advise companies to consult with legal counsel to determine whether they are RCSs or ECSs, and implement risk-based policies and procedures accordingly, depending on their jurisdiction, risk tolerance, relationship with law enforcement, and other factors.

© 2024 Thorn

IMPACT	Significant. Makes it more difficult for downstream bad actors to either directly misuse or fine-tune and then misuse these models to generate AIG-CSAM and CSEM. Mitigates the potential for a model to inadvertently produce this content through an innocuous prompt.
SCOPE	Narrow
VARIABLES	What data sources are trusted, automated vs. manual verification
RELEVANCE	AI Developers. Closed and Open.

Informational Resource:

- Birhane, Abeba, et al. *Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes.* arXiv:2110.01963, arXiv, 5 Oct. 2021. *arXiv.org*, https://doi.org/10.48550/arXiv.2110.01963.
- Cohen, Neil. *The Ethical Use of Personal Data to Build Artificial Intelligence Technologies: A Case Study on Remote Biometric Identity Verification.* Carr Center for Human Rights Policy Harvard University, https://www.hks.harvard.edu/centers/carr/publications/ethical-use-personal-data-build-artificial-intelligence-technologies-case.
- "Deleting Unethical Data Sets Isn't Good Enough." *MIT Technology Review*, https://www.technologyreview.com/2021/08/13/1031836/ai-ethics-responsible-data-stewardship.
- "MIT, Cohere for AI, Others Launch Platform to Track and Filter Audited AI Datasets." *VentureBeat*, 25 Oct. 2023, https://venturebeat.com/ai/mit-cohere-for-ai-others-launch-platform-to-track-and-filter-audited-ai-datasets.
- Thiel, David. *Identifying and Eliminating CSAM in Generative ML Training Data and Models*. Stanford Digital Repository, Dec. 2023, https://doi.org/10.25740/kh752sm9123.

Implementation Resource:

- *Blocking and Categorizing Content.* https://www.interpol.int/en/Crimes/Crimes-against-children/Blocking-and-categorizing-content.
- *URL List.* Internet Watch Foundation, https://www.iwf.org.uk/our-technology/our-services/url-list.

DEVELOP: MITIGATION #2

Detect, remove and report CSAM and CSEM from your training data

If you cannot determine whether a dataset you are newly ingesting has been audited for CSAM and CSEM, use available tools (e.g. classifiers, hashing/matching technology, etc.) to identify this abuse data in your datasets and ensure that the identified abuse data is excluded prior to preparing the data for training your generative models. Similarly, for already ingested data, where possible do the same. Where applicable, report the content you have found to governing authorities (see the "Reporting CSAM and AIG-CSAM" section in "Additional Resources" below for more details). If the data comes from a pre-existing dataset, notify the curators of the dataset that this content has been identified. Thoroughly document your procedures for detecting and removing CSAM and CSEM from training data, which should include the process for promptly reporting and preserving this content where required and/or permitted by law.[1] Train employees involved in model training on those procedures.

2 See *supra* note 1.

© 2024 Thorn

IMPACT	Significant. Makes it more difficult for downstream bad actors to either directly misuse or fine-tune and then misuse these models to generate AIG-CSAM and CSEM. Mitigates the potential for a model to inadvertently produce this category of content through an innocuous prompt. Mitigates the potential for a model to replicate the specific abuse content.
SCOPE	Narrow/Medium
VARIABLES	Precision/recall of automated detection solutions, in-house vs. external detection models, computational and human resources
RELEVANCE	AI Developers, Data Hosting Platforms. Closed and Open.

Informational Resource:

- Cambridge Consultants. *Use of AI in Online Content Moderation*. UK Office of Communications, https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf.
- Carlini, Nicholas, et al. "Extracting Training Data from Large Language Models." *30th USENIX Security Symposium, Aug. 2021.* https://www.usenix.org/system/files/sec21-carlini-extracting.pdf.
- Carlini, Nicholas, et al. "Extracting Training Data from Diffusion Models." *32nd USENIX Security Symposium, Aug. 2023.* https://www.usenix.org/system/files/usenixsecurity23-carlini.pdf.
- Somepalli, Gowthami, et al. "Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models." *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2023.* https://doi.org/10.48550/arXiv.2212.03860.
- "Using GPT-4 for Content Moderation." *OpenAI*, 15 Aug. 2023, https://openai.com/blog/using-gpt-4-for-content-moderation.

Implementation Resource:

- ActiveFence, https://www.activefence.com.
- Amazon Rekognition, https://aws.amazon.com/rekognition/content-moderation.
- Google Child Safety Toolkit, https://protectingchildren.google/tools-for-partners.
- Hive Moderation, https://hivemoderation.com.
- Haidra Org Horde-Safety, https://github.com/Haidra-Org/horde-safety.
- IWF Services and Technology, https://www.iwf.org.uk/our-technology/our-services.
- Meta HMA, https://github.com/facebook/ThreatExchange/tree/main/hasher-matcher-actioner.
- Microsoft pDNA, https://www.microsoft.com/en-us/photodna.
- NCMEC Hash Sharing API Technical Documentation, https://report.cybertip.org/ws-hashsharing/v2/documentation.
- TaskUs, https://www.taskus.com.
- Teleperformance, https://teleperformance.com.
- Thorn's Safer, https://safer.io.

© 2024 Thorn

DEVELOP: MITIGATION #3

Separate depictions/representations of children from adult sexual content in your image, video or audio generation training datasets

When training an open source model, make best efforts to not include images/videos of children, or audio recordings of children in datasets that contain adult sexual content. Make best efforts to prevent your open source models from having both content of children[3] and adult sexual content in its training data. For models built with the purpose of de-aging image and/or video content, do not include adult sexual content in your training datasets. Note that the definition of adult will vary depending on your legal jurisdiction.

IMPACT	Significant. Makes it more difficult for downstream bad actors to either directly misuse or fine-tune and then misuse these models to generate AIG-CSAM and CSEM. Mitigates the potential for a model to inadvertently produce this content through an innocuous prompt. Makes it more difficult for downstream bad actors to directly misuse de-aging models to alter sexual content of adults to depict AIG-CSAM.
SCOPE	Medium
VARIABLES	Precision/recall of automated detection solutions, in house vs. external detection solutions, computational and human resources
RELEVANCE	AI Developers, Data Hosting Platforms. Open.

Informational Resource:

- Okawa, Maya, et al. "Compositional Abilities Emerge Multiplicatively: Exploring Diffusion Models on a Synthetic Task." *37th Conference on Neural Information Processing Systems, Dec. 2023.* https://doi.org/10.48550/arXiv.2310.09336.
- Ramesh, Rahul, et al. *Compositional Capabilities of Autoregressive Transformers: A Study on Synthetic, Interpretable Tasks*. arXiv:2311.12997, arXiv, 21 Nov. 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2311.12997.

Implementation Resource:

- ActiveFence, https://www.activefence.com.
- Amazon Rekognition, https://aws.amazon.com/rekognition/content-moderation.
- Bumble Private Detector, https://github.com/bumble-tech/private-detector.
- Hive Moderation, https://hivemoderation.com.
- TaskUs, https://www.taskus.com.
- Teleperformance, https://teleperformance.com.
- Yoti, https://www.yoti.com/business/age-verification.
- VeriLook, https://www.neurotechnology.com/verilook.html.

3 Under the U.S. Children's Online Privacy Protection Act, websites and online services that collect some types of personal information through their services from children under 13 for the purpose of training models must obtain parental consent.

© 2024 Thorn

DEVELOP: MITIGATION #4

Conduct red teaming for AIG-CSAM and CSEM

Incorporate structured, scalable, and consistent stress testing of your model for AIG-CSAM and CSEM. This will help your team understand how it can be misused to produce AIG-CSAM and CSEM. Update your model accordingly to mitigate for the issues you discover.

Red teaming should happen throughout the development process, not just in anticipation of model release. If a model is substantively updated or iterated on, such that its capabilities are measurably increased in risk areas connected to the production of AIG-CSAM and CSEM, correspondingly red teaming should also iteratively occur to understand and mitigate for misalignment. Ensure that after each round of red teaming, findings are integrated back into model training and development, such that if a red teaming query produces violative content, the new version of the model is less able to produce that content given the same query.

Attempting to generate AIG-CSAM may implicate local law. Consult with legal counsel on this matter. Regardless, it is possible for red teaming to be carried out such that due regard is given for the regulatory bounds on those carrying out testing.[4]

Thoroughly document compliance procedures for red teaming, which should include (consistent with your legal obligations) instructions on promptly reporting and preserving CSAM and AIG-CSAM.[5] Train employees responsible for red teaming and model evaluation on these procedures.

IMPACT	⚡ Significant. Enables developers to identify misalignment which can be mitigated prior to model release, making it more difficult for downstream bad actors to either directly misuse or fine-tune and then misuse these models to generate AIG-CSAM and CSEM.
SCOPE	Medium
VARIABLES	Cadence (e.g. with each new minor product feature change, with each new major product feature change), in house vs. external service, using known set of prompts that have produced sexually abusive content vs. new ones, manual vs. automated
RELEVANCE	AI Developers. Closed and Open.

4 One option could involve assessing whether:
- the model is capable of producing adult sexual content, including that depicting a specific individual
- the model is capable of producing photo realistic or other representations of children

"Compositional generalization" is a term that is sometimes used to refer to a model's ability to combine attributes seen independently in training. While it is still an open area of research on when and how models are able to do this, some evidence indicates that if both independent factors named above have a high propensity and the model demonstrates strong compositional generalization, this may indicate a corresponding high propensity for a model to be able to produce AIG-CSAM. See the informational resources in the mitigation titled "Separate depictions/representations of children from adult sexual content in your training datasets" for additional references.

5 See *supra* note 1.

© 2024 Thorn

Informational Resource:

- Ganguli, Deep, et al. *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*. arXiv, 22 Nov. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2209.07858.
- "Google's AI Red Team: The Ethical Hackers Making AI Safer." *Google*, 19 July 2023, https://blog.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer.
- "GPT-4 System Card." *OpenAI*, 23 Mar. 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf.
- "GPT-4V System Card." *OpenAI*, 23 Sep. 2023, https://cdn.openai.com/papers/GPTV_System_Card.pdf.
- Henderson and Mitchell, et al. "Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models." *6th AAAI/ACM Conference on AI, Ethics, and Society, Aug. 2023*. https://arxiv.org/abs/2211.14946v2.
- Ouyang, Long, et al. "Training Language Models to Follow Instructions with Human Feedback." *arXiv.org*, 4 Mar. 2022, https://arxiv.org/abs/2203.02155v1.
- "Planning Red Teaming for Large Language Models (LLMs) and Their Applications." *Microsoft*, 6 Nov. 2023, https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming.
- Rando, Javier, et al. *Red-Teaming the Stable Diffusion Safety Filter*. arXiv, 10 Nov. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2210.04610.
- "Red-Teaming Large Language Models." *Hugging Face*, 24 Feb. 2023, https://huggingface.co/blog/red-teaming.
- Wallace, Bram, et al. "Diffusion Model Alignment Using Direct Preference Optimization." *arXiv.org*, 21 Nov. 2023, https://arxiv.org/abs/2311.12908v1.
- Zou, Andy, et al. *Universal and Transferable Adversarial Attacks on Aligned Language Models*. arXiv, 27 July 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2307.15043.

Implementation Resource:

- Scale AI Test & Evaluation, https://scale.com/blog/test-evaluation-vision.
- Thorn Red Teaming, https://www.thorn.org/contact.

DEVELOP: MITIGATION #5

Include content provenance by default

Include indicators of the source of the content in any image or video that your model outputs, or some other ability to detect the model as the source. Where possible, CSAM hotlines (e.g. NCMEC) and relevant law enforcement (e.g. HSI C3, ICAC) should have access to the tool[6] needed to determine the source, if any such tool is needed.

For open source models: include content provenance during the content generation process (e.g. visually imperceptible watermarks in the training data, fine-tuning the decoder that generates images from the latent vectors to natively embed a maximally indelible watermark into images).

IMPACT ⚡ Significant. Makes it easier for law enforcement and NGOs to quickly identify content that depicts an identified or unidentified victim.

6 Note that your choice of content provenance may impact your ability to achieve your desired level of granularity. For example, solutions that add a manifest to the file, like C2PA, allow for including information like the prompt, IPTC digital media type, and input images used to generate the resulting image. In contrast, watermarking decoding typically provides a more binary response about whether a tool was used for generation or not. In both cases, adversarial actors could take advantage of weaknesses in the provenance solution to attempt to strip out the information from the source image. A combination of these methods will likely enable the most robust signal of content provenance that can support decision making.

© 2024 Thorn

SCOPE	Wide
VARIABLES	Robustness to tampering, robustness to benign modification (e.g. resizing of the image), imperceptibility, computational resources
RELEVANCE	AI Developers. Closed and Open.

Informational Resource:

- Fernandez, Pierre, et al. "The Stable Signature: Rooting Watermarks in Latent Diffusion Models." *2023 IEEE/CVF International Conference on Computer Vision, Oct. 2023*. https://doi.org/10.48550/arXiv.2303.15435.
- "From the Darkroom to Generative AI." *Content Authenticity Initiative*, 15 Aug. 2023, https://contentauthenticity.org/blog/from-the-darkroom-to-generative-ai.
- "Meta's AI Watermarking Plan Is Flimsy, at Best." *IEEE Spectrum*, 4 Mar. 2024, https://spectrum.ieee.org/meta-ai-watermarks.
- "SynthID." *Google DeepMind*, 16 Nov. 2023, https://www.deepmind.com/synthid.
- Wen, Yuxin, et al. *Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust*. arXiv, 3 July 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2305.20030.
- Yu, Ning, et al. "Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data." *2021 IEEE/CVF International Conference on Computer Vision, Oct. 2021*. https://doi.org/10.48550/arXiv.2007.08457.

Implementation Resource:

- C2PA, https://c2pa.org.
- "Provenance, Watermarking & Deepfake Detection." *Hugging Face*, https://huggingface.co/collections/society-ethics/provenance-watermarking-and-deepfake-detection-65c6792b0831983147bb7578.
- "Stable Signature." *Meta*, 6 Oct. 2023, https://ai.meta.com/blog/stable-signature-watermarking-generative-ai.

DEVELOP: MITIGATION #6

Define specific training data and model development policies

Document within your organization what type of training data is and is not allowed.[7] Document principles for generative AI development, including core principles on safety and exploitation.

IMPACT	Incremental. Provides clarity and transparency within the organization on ethical lines and boundaries.
SCOPE	Narrow
VARIABLES	In house vs. externally developed policies, regulatory landscape
RELEVANCE	AI Developers. Closed and Open.

7 See *supra* note 3.

© 2024 Thorn

Informational Resource:

- "Google AI Principles." *Google AI*, https://ai.google/responsibility/principles.
- "How Your Data is Used to Improve Model Performance." *OpenAI*, https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance. N.B. This is specific to user data, but can be expanded to all training data.

DEVELOP: MITIGATION #7

Prohibit customer use of your model to further sexual harms against children

Document to customers and users in a clear and accessible manner that using your model to generate AIG-CSAM, CSEM, or any other sexual and violent content involving children is prohibited.

Thoroughly document procedures for detecting potential customer abuses of the model, which should include the process for reporting and preserving CSAM, consistent with the company's legal obligations.[8] Train employees on these procedures. Log abuses so that appropriate enforcement action can be taken against those who violate contractual prohibitions.

IMPACT	Incremental. Provides clarity and transparency outside the organization on ethical lines and boundaries.
SCOPE	Narrow
VARIABLES	Scale of announcement, regulatory requirements
RELEVANCE	AI Developers, AI Providers (first-party). Closed and Open.

Informational Resource:

- "Community Guidelines & Rules." *Snapchat*, https://values.snap.com/privacy/transparency/community-guidelines.
- "Google Generative AI Prohibited Use Policy." *Google*, https://policies.google.com/terms/generative-ai/use-policy.
- "OpenAI Usage Policies." *OpenAI*, https://openai.com/policies/usage-policies.

N.B. If your datasets are in multiple languages, or your technology can generate multiple languages, the mitigations actioned on should also be multilingual (e.g. detect for grooming in English and Spanish, in datasets that contain English and Spanish). Services should not be offered in languages where you cannot protect your users.

8 See *supra* note 1.

© 2024 Thorn

Further Opportunities

Build an open source resource of already cleaned datasets

> Cleaning datasets can be expensive and impact the wellness of the team involved, depending on which strategy is employed for cleaning (e.g. manual moderation, ML/AI models, auto removal of content from blacklisted sites, etc.). The tech ecosystem would benefit from an open source resource for shared, already cleaned datasets so training data for open source models doesn't have to get reviewed more than once. In some cases, this may need to be balanced against the confidential nature of proprietary datasets.

Adopt existing best practices in the computing security field

> There is a wealth of relevant experience in the computing security field, around discovering and sharing known issues, expected response and required mitigations. Organizations could adopt these existing best practices in the context of mitigating the misuse of generative AI for child sexual abuse.

Conduct research on leveraging content provenance metadata to support the detection of sexual harms against children

> As content provenance solutions and standards continue to develop, technologists could shape them in a way that allows provenance metadata (e.g. C2PA metadata) to be used as additional feature inputs for other ML/AI models and algorithmic solutions built to detect CSAM and other sexual harms against children.

Intentionally engage across industry and child safety experts

> These threats are still emerging, and may change shape over the years. Organizations could consult with a diverse user base and engage transparently and regularly with industry peers and child safety experts (e.g. AI Verify Foundation, Center for AI Safety, Center for Democracy and Technology, eSafety, Frontier Model Forum, IWF, NCMEC, Partnership on AI, Responsible AI UK, Thorn) to share best practices, share trends and statistics on how these harms are manifesting, and share evaluations of their models' robustness against those threats.

Develop AI principles or charters

> Good AI governance processes support the mitigation of safety risk in model development. Organizations could develop AI principles or charters (e.g. OpenAI, Google, Anthropic) and AI councils (e.g. Google, Adobe) to help ensure a "safety" north star during model development. Good governance can help guide mitigations against a set of principles and also can be an opportunity to embed child safety expertise.

© 2024 Thorn

Potentially Problematic Downstream Implications

We recognize that some of the recommended mitigations above have potentially problematic downstream implications, beyond their intended positive impact. We include below some of these possibilities, as well as our suggested response.

Model capabilities are reduced

> Overly sanitizing training datasets could result in reducing the model's capabilities for other, non child sexual abuse related outputs. We recommend exploring existing techniques (e.g. dataset transparency, bootstrapped learning) that can reduce these risks.

Informational Resource:

- Flennerhag, Sebastian, et al. "Bootstrapped Meta-Learning." *10th International Conference on Learning Representations, April 2022*. https://doi.org/10.48550/arXiv.2109.04504.

Cultural distinctions in content moderation

> There are cultural distinctions in content moderation that could manifest in certain training datasets being cleaned differently than others. We recommend thoroughly understanding the cultural context of your content moderation solution, whether in house or external, as well as engaging your legal team to understand more deeply the obligations you have in your particular regulatory context.

Informational Resource:

- Gillespie, Tarleton. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press, 2018.
- Jiang, Jialun Aaron, et al. "Understanding International Perceptions of the Severity of Harmful Content Online." *PLOS One*, vol. 16, no. 8, Aug. 2021. https://doi.org/10.1371/journal.pone.0256762.
- Roberts, Sarah T. Behind the Screen: Content Moderation in the Shadows of Social Media. Yale University Press, 2019.

Bias in automated content moderation solutions

> There is the potential for automated content moderation solutions to be biased (e.g. more likely to find CSAM of light skin children than dark skin children). We recommend evaluating your automated solution, whether in house or external, for performance not just of the target category but also of the target category against the full spectrum of demographics represented in the real world. If building in house solutions, we recommend following known best practices for removing bias in your models.

Informational Resource:

- Binns, Reuben, et al. "Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation." *9th International Conference on Social Informatics, Sep. 2017*. https://doi.org/10.48550/arXiv.1707.01477.
- Hacker, Philipp, et al. "Regulating ChatGPT and other Large Generative AI Models." *2023 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, June 2023*. https://doi.org/10.1145/3593013.3594067.
- "Mitigating Unfair Bias in ML Models with the MinDiff Framework." *Google Research*, 16 Nov. 2020, https://blog.research.google/2020/11/mitigating-unfair-bias-in-ml-models.html.
- Vidgen, Bertie, and Leon Derczynski. "Directions in Abusive Language Training Data, a Systematic Review: Garbage In, Garbage Out." *PLOS One*, vol. 15, no. 12, Dec. 2020. https://doi.org/10.1371/journal.pone.0243300.

© 2024 Thorn

Implementation Resource:

- "Responsible AI Toolkit." *Department of Defense*, https://rai.tradewindai.com/tools-list.

Wellness implications in content moderation and red teaming

> There are significant wellness implications that come with content moderation and red teaming. Exposure to CSAM, AIG-CSAM and other forms of child sexual abuse can result in long term trauma, vicarious or secondary trauma and PTSD. We recommend incorporating programmatic wellness support and benefits for your internal content moderation team and red teaming group. If engaging with external content moderation support and red teaming groups, evaluate them for the breadth and depth of their wellness support for their staff. This should also include in role training on child sexual abuse and exploitation, so that teams working with CSAM and AIG-CSAM have the right knowledge to address this content.

Informational Resource:

- "Content Moderators: Superheroes in the Shadows of Social Media." *Middlesex University London*, 6 May 2022, https://www.mdx.ac.uk/our-research/centres/secondarytraumaresearch/blog/content-moderators-the-superheroes-in-the-shadows-of-social-media.
- "Exploring Methods to Improve the Psychological Wellness of Content Moderators." *UT Austin Department of Computer Science*, 7 Mar. 2022, https://www.cs.utexas.edu/news/2022/exploring-methods-improve-psychological-wellness-content-moderators.
- Matias, J. Nathan and Sarah Gilbert. "What Can Companies Do For Moderator Well-Being?" *Citizens and Technology Lab*, 27 Sep. 2022, https://citizensandtech.org/2022/09/moderator-wellbeing-trustcon.
- Mateo, Flora. "Nouvelle édition du Livre Blanc de Point de Contact : la Trust & Safety en mouvement pour protéger celles et ceux qui nous protègent." *Point de Contact*, 18 Oct. 2023, https://www.pointdecontact.net/nouvelle-edition-du-livre-blanc-de-point-de-contact-la-trust-safety-en-mouvement-pour-proteger-celles-et-ceux-qui-nous-protegent.
- "What Helps Content Moderators Cope with the Job?" *Middlesex University London*, 11 Nov. 2022, https://www.mdx.ac.uk/our-research/centres/secondarytraumaresearch/blog/what-helps-content-moderators-cope-with-the-job.

Implementation Resource:

- The Workplace Wellness Project, https://theworkplacewellnessproject.com.
- ZevoHealth, https://www.zevohealth.com/trust-and-safety.

© 2024 Thorn

Deploy

As defined above, **deploy** refers to the method or act of integrating a ML/AI model into a production environment, and/or the method or act of making a ML/AI model available for use. Note that due to the nature of the mitigations in this section, each mitigation (with the exception of the last one in this section) should be understood as describing an ongoing and systemized process.

Mitigations

SIGNIFICANT IMPACT OPEN SOURCE CLOSED SOURCE

DEPLOY OVERVIEW	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	✓	✓			

DEPLOY: MITIGATION #1

Detect abusive content (CSAM, AIG-CSAM, and CSEM) in inputs and outputs

In deployment settings where you have direct access to the inputs and outputs, detect for input prompts intended to produce AIG-CSAM and CSEM. Similarly, detect for CSAM provided at the inputs, and for AIG-CSAM and CSEM that may have been produced at the output. Where it is required or consistent with policy, report CSAM and AIG-CSAM to the proper governing authorities (see the "Reporting CSAM and AIG-CSAM" section in "Additional Resources" below for more details). Set up content moderation flows for outputs. Thoroughly document procedures on detecting abusive content in inputs and outputs, which should include the process

© 2024 Thorn

for reporting and destroying CSAM, consistent with the company's legal obligations.[9] Train employees on such procedures.

IMPACT	Significant. Makes it more difficult for CSAM, AIG-CSAM and CSEM to spread.
SCOPE	Narrow
VARIABLES	Precision/recall of automated detection solutions, in house vs. external detection solutions, computational resources
RELEVANCE	AI Developers, AI Providers (first-party). Closed.

Resources:

- See the resources listed in the mitigation titled "Detect, remove and report CSAM and CSEM from your training data."

DEPLOY: MITIGATION #2

Include user reporting, feedback or flagging options

First-party AI Providers should include a pathway for users to report content the model produces that may violate the model's child safety policies. For third-party AI Providers, include a pathway for users to report models that generate AIG-CSAM and CSEM.

Ensure these pathways allow for in real time reporting and in application flagging/feedback, to reduce user barriers to reporting where applicable. In response to user reports about potential violations of the model's child safety policies, provide links to support services. Provide contact details, so that law enforcement and users can reach out with additional queries or feedback.

Thoroughly document procedures for the trust and safety team to handle user reporting, including the process by which potential CSAM should be reported and preserved, consistent with the company's legal obligations.[10]

IMPACT	Significant. Build up ecosystem knowledge of model's limitations and capacity to generate AIG-CSAM and CSEM. Build up ecosystem knowledge of which models produce violative content.
SCOPE	Narrow
VARIABLES	Ease of reporting mechanism

9 See *supra* note 1.
10 See *supra* note 1.

© 2024 Thorn

RELEVANCE	AI Providers (first and third-party). Closed and Open.

Informational Resource:

- *Child Helplines – Child Helpline International*. https://childhelplineinternational.org/helplines.
- Thorn. (2023). Responding to Online Threats: Minors' Perspectives on Disclosing, Reporting, and Blocking in 2021. https://info.thorn.org/hubfs/Research/Thorn_ROT_Monitoring_2021.pdf.

Implementation Resource:

- "Counselling and Support Services | ESafety Commissioner." *ESafety Commissioner*, https://www.esafety.gov.au/about-us/counselling-support-services.

DEPLOY: MITIGATION #3

Include an enforcement mechanism

Enforcement mechanisms are necessary to address user violations of child safety policies. Traditionally we think of enforcement mechanisms as applying to an individual user or profile. Generative AI developers should also think about what enforcement mechanisms may look like for the model itself.[11] This concept may bridge into model maintenance. Any enforcement of child safety policies should be performed in a manner that allows the company to preserve information sufficient to meet any legal requirements.

IMPACT	Significant. Mitigate the possibility of repeated violations.
SCOPE	Narrow
VARIABLES	Precision/recall of enforcement solutions, level of manual verification
RELEVANCE	AI Developers. Closed.

Informational Resource:

- *DSA Transparency Database*. https://transparency.dsa.ec.europa.eu.
- "On Policy Development at YouTube." *Google*, 1 Dec. 2022, https://blog.google/intl/en-in/products/platforms/on-policy-development-at-youtube.
- "Policy Development." *Trust & Safety Professional Association*, 17 June 2021, https://www.tspa.org/curriculum/ts-fundamentals/policy/policy-development.
- "The Safe Framework." *Digital Trust & Safety Partnership*, Dec. 2021, https://dtspartnership.org/wp-content/uploads/2021/12/DTSP_Safe_Framework.pdf.
- "Trust & Safety Best Practices Framework." *Digital Trust & Safety Partnership*, Apr. 2021, https://dtsp.wpengine.com/wp-content/uploads/2021/04/DTSP_Best_Practices.pdf.

11 For example: user level enforcement may look like a strike or user profile disable, whereas a model enforcement mechanism may look like a new regular expression rule added to an existing safety filter that could flag and block prompts (or variants of a prompt) that have resulted in a model producing AIG-CSAM or CSEM.

© 2024 Thorn

DEPLOY: MITIGATION #4

Assess generative models before access

For first-party AI Providers: assess generative models for their potential to generate AIG-CSAM and CSEM before the models are hosted on your platforms. For models that are assessed and found to be in Category 2a or 2b (as defined in the Safety Assessment Categories in the Additional Resources section), do not host these models until after they have been updated with mitigations in place. If retraining a model, or other mitigations like model editing are impractical or not possible, restrict the model to hosted-generation only. By doing this, you can employ prompt filtering and other measures to prevent abuse, as well as prevent downloads of model weights or use of the model in private, offline settings. Models in Category 2c (as defined in the Safety Assessment Categories in the Additional Resources section) should not be hosted on your platforms.

For third-party AI Providers: where possible, in a similar fashion as described above, directly assess generative models for their potential to generate AIG-CSAM and CSEM before hosting. Where this is currently infeasible, third-party AI Providers should instead require developers to fill out the child safety section of their model card before hosting the model on your platform. Third-party AI Providers should then use this child safety section to assess whether the model satisfies your internal child safety policies, or has high likelihood of being in Category 2a or 2b. Use this information to make a decision on whether to allow the model to be hosted, or require the developer to incorporate mitigations before re-hosting. Models in Category 2c should not be hosted on your platform.

Attempting to generate AIG-CSAM may implicate local law. Consult with legal counsel on this matter. Regardless, it is possible for evaluations to be carried out such that due regard is given for the regulatory bounds on those conducting the evaluations.[12]

IMPACT	⚡ Significant. Mitigates the possibility of bad actors getting access to models that were built and released without first minimizing their potential to generate AIG-CSAM and CSEM. Normalizes assessing and mitigating a model's capability to generate AIG-CSAM and CSEM. Acts as a stopping point for AI Developers to further refine their models before release.
SCOPE	Wide
VARIABLES	Scope of assessment, cadence of assessment (e.g. every new model version, vs. only major releases), computational resources, time allotted for assessment
RELEVANCE	AI Providers (first and third-party). Closed and Open.

For developers with access to the original training dataset and infrastructure, see the resources across Develop.

12 For example, a composite model could be constructed by connecting the output of a generative model directly to the input of a CSAM classification model that performs well on AIG-CSAM. This model would return a CSAM classification score for that particular model when provided with an input prompt to identify potential policy violations without an image being rendered.

© 2024 Thorn

Informational Resource:

- "First of its kind Generative AI Evaluation Sandbox for Trusted AI by AI Verify Foundation and IMDA." *Infocomm Media Development Authority*, 31 Oct. 2023, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox.
- Gandikota, Rohit, et al. *Unified Concept Editing in Diffusion Models*. arXiv, 25 Aug. 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2308.14761.
- Mengyao, Lyu, et al. "One-Dimensional Adapter to Rule Them All: Concepts, Diffusion Models and Erasing Applications." *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2024*. https://doi.org/10.48550/arXiv.2312.16145.
- Mitchell, Eric, et al. "Memory-Based Model Editing at Scale." *39th International Conference on Machine Learning, July 2022*. https://doi.org/10.48550/arXiv.2206.06520.
- Mitchell, Eric, et al. *Fast Model Editing at Scale*. arXiv, 13 June 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2110.11309.
- Röttger, Paul, et al. *SafetyPrompts: a Systematic Review of Open Datasets for Evaluating and Improving Large Language Model Safety*. arXiv, 8 April 2024. *arXiv.org*, https://doi.org/10.48550/arXiv.2404.05399.
- Sociotechnical Safety Evaluation Repository, https://docs.google.com/spreadsheets/u/1/d/e/2PACX-1vQObeTxvXtOs--zd98qG2xBHHuTTJOyNISBJPthZFr3at2LCrs3rcv73d4of1A78JV2eLuxECFXJY43/pubhtml.

Implementation Resource:

- Gandikota, Rohit. "Rohitgandikota/Unified-Concept-Editing." *GitHub*, 15 Apr. 2024, github.com/rohitgandikota/unified-concept-editing.
- Mitchell, Eric. "Eric-Mitchell/Mend." *GitHub*, 16 Apr. 2024, github.com/eric-mitchell/mend.
- MLCommons. "Announcing MLCommons AI Safety v0.5 Proof of Concept." *MLCommons*, 16 Apr. 2024, https://mlcommons.org/2024/04/mlc-aisafety-v0-5-poc.
- See the "Model Safety Assessment" section in "Additional Resources" below for details regarding safety assessment categories, a dataset to support safety assessments, and a dataset containing hashes of Category 2 models.

DEPLOY: MITIGATION #5

Include prevention messaging for CSAM solicitation

Prevention and deterrence responses to CSAM solicitation on search engines or functions are becoming an industry standard. Where generative AI products have text input fields, serve prevention or deterrence messaging[13] (such as prompts, nudges or interstitial warnings which provide users with information).

IMPACT	Incremental. Deterrence of additional CSAM solicitation from a model.
SCOPE	Narrow
VARIABLES	Keyword/query lists, source of deterrence response, automation of deterrence response
RELEVANCE	AI Developers, AI Providers (first and third-party). Closed.

13 Some models may be trained to do this inherently, by highlighting harm and illegality. But a specific triggering program could be built using automated detection mechanisms, and then served as in-product intervention; or a model could potentially be fine-tuned to output a preventative response. These methods may currently be imperfect, but we recommend making best efforts.

© 2024 Thorn

Informational Resource:

- Grant, Harriet. "Pornhub Partners with Child Abuse Charities to Intercept Illegal Activity." *The Guardian*, 28 Sept. 2022. *The Guardian*, https://www.theguardian.com/global-development/2022/sep/28/pornhub-partners-with-child-abuse-charities-to-intercept-activity.
- Hunn, Charlotte, and Paul Watters. *How to Implement Online Warnings to Prevent the Use of Child Sexual Abuse Material*. Australian Institute of Criminology, 2023. https://doi.org/10.52922/ti78894.
- "Preventing Child Exploitation on Our Apps." *Meta*, 23 Feb. 2021, https://about.fb.com/news/2021/02/preventing-child-Exploitation-on-our-apps.
- Prichard, Jeremy et al. 2022. *Online Messages to Reduce Users' Engagement with Child Sexual Abuse Material: A Review of Relevant Literature for the reThink Chatbot*. University Of Tasmania, 1 Jan. 2022. https://hdl.handle.net/102.100.100/490717.
- Steel, Chad M. S. "Web-Based Child Pornography: The Global Impact of Deterrence Efforts and Its Consumption on Mobile Platforms." *Child Abuse & Neglect*, vol. 44, June 2015, pp. 150–58. https://pubmed.ncbi.nlm.nih.gov/25605590.

Implementation Resource:

- Project Intercept - Lucy Faithful Foundation. https://project-intercept--lff.super.site.

DEPLOY: MITIGATION #6

Incorporate phased deployment

Iterate with soft launches with smaller groups of users or external red teamers, monitoring for abuse in early stages rather than launching the first version of the product broadly. Where possible, find statistically representative sets of users for soft launches. Test the assumptions made in the lead up to the launch that might be incorrect. The novelty of your deployment may impact your decisions when considering the variables associated with phased deployment.

IMPACT	Incremental. Increase the chances of discovering issues, before broader release, where your model can generate AIG-CSAM and CSEM.
SCOPE	Narrow
VARIABLES	Number of soft launches, pace of iteration, size of user groups
RELEVANCE	AI Developers. Closed and Open.

Informational Resource:

- *Claude 2*. https://www.anthropic.com/news/claude-2.
- "Guidelines for developing high-quality ML solutions | Cloud Architecture Center." *Google Cloud*, https://cloud.google.com/architecture/guidelines-for-developing-high-quality-ml-solutions.
- "What's Ahead for Bard: More Global, More Visual, More Integrated." *Google*, 10 May 2023, https://blog.google/technology/ai/google-bard-updates-io-2023.

© 2024 Thorn

DEPLOY: MITIGATION #7

Incorporate a child safety section into model cards

Update your model card template to include questions on mitigations implemented for child safety.

IMPACT	Incremental. Normalizes assessing and mitigating for a model's capability to generate AIG-CSAM and CSEM. Acts as a stopping point for AI Developers to further refine their models before release.
SCOPE	Narrow
VARIABLES	Breadth and depth of disclosure, template source
RELEVANCE	AI Developers, AI Providers (first and third-party). Closed and Open.

Informational Resource:

- See the "Model Card: Child Safety" section in "Additional Resources" below for a sample set of questions.

Further Opportunities

For high risk uses, require users to be logged into your service

> Closed organizations could consider different approaches to accountability and preventing recidivism via identity verification. Collecting identifying information on users that use online settings to sexually abuse children allows for significantly more robust information in reports to law enforcement and hotlines. In the generative AI setting, this type of identity verification supports the ability to trace back to the source of misuse of the model for generating AIG-CSAM or CSEM. Different settings and legal requirements may influence what type of identity verification is appropriate in different environments, but there is value in a shared understanding of what information is most valuable to law enforcement when it comes to victim identification. See the "Identity Verification: Settings and Possibilities" section in "Additional Resources" below for more details on some suggested approaches across various low and high risk settings.

Develop a standardized safety assessment; build a shared resource of prompts known to generate AIG-CSAM

> A standardized safety assessment across industry would allow for consistent and transparent evaluation of a model's propensity for generating AIG-CSAM and CSEM. Similarly, a shared dataset of known prompts that generate AIG-CSAM and CSEM could be used to support the development of a standardized safety assessment. See the "Model Safety Assessment" section in "Additional Resources" below for our recommended assessment categories in the AIG-CSAM case, as well as a dataset to support safety assessments.

© 2024 Thorn

For third-party AI Providers, build out automated, scalable assessments of models for their capacity to generate AIG-CSAM

> Building off of the previous opportunity, in the long term scalable model assessments will be necessary to match the scale and speed of model development. Third-party AI Providers could invest in building out pipelines for this type of automated scalable assessment, collaborating with organizations like NIST who are positioned to provide tools for standardized safety assessment.

Potentially Problematic Downstream Implications

We recognize that some of the recommended mitigations above have potentially problematic downstream implications, beyond their intended positive impact. We include below some of these possibilities, as well as our suggested response.

Cultural distinctions in content moderation

> See the previous Develop discussion.

Bias in automated content moderation solutions

> See the previous Develop discussion.

Wellness implications in content moderation

> See the previous Develop discussion.

False positives from automated content moderation solutions

> There is always a precision/recall tradeoff that must be made when deploying automated detection solutions. Ideally, humans should be in the loop when making decisions where automated detection solutions are used to inform those decisions, e.g. when reporting content, actioning on enforcement mechanisms, or responding to policy violations. We further recommend incorporating accessible appeals processes for those who believe a decision has been made in error.

Informational Resource:

- Chowdhury, Nafia, and Daphne Keller. "Automated Content Moderation: A Primer." *Program on Platform Regulation*, Mar. 2022. *cyber.fsi.stanford.edu*, https://cyber.fsi.stanford.edu/publication/automated-content-moderation-primer.
- "The Internet Commission – Advancing Digital Responsibility through Independent Evaluation." The Internet Commission, https://inetco.org/report.

© 2024 Thorn

Maintain

As defined above, **maintain** refers to the act of maintaining the quality of ML/AI models in the face of data drift and changing landscape. Note that due to the nature of the mitigations in this section, they can co-occur simultaneously with the interventions in the Develop and Deploy sections.

Mitigations

SIGNIFICANT IMPACT · OPEN SOURCE · CLOSED SOURCE

MAINTAIN OVERVIEW	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	✓	✓		✓	

© 2024 Thorn

MAINTAIN: MITIGATION #1

Remove services for "nudifying" images of children from search results

Search Engines should delist links to sites that provide services and tutorials for "nudifying" and sexualizing images, where a user can upload an image of a clothed child and have the service output a corresponding image of that same child without clothes.

IMPACT	⚡ Significant. Makes it more difficult for bad actors to generate AIG-CSAM by sexualizing a child's benign imagery.
SCOPE	Narrow
VARIABLES	Trusted sources for sites, automated vs. manual verification
RELEVANCE	Search Engines.

Informational Resource:

- Kristof, Nicholas. "Opinion | The Online Degradation of Women and Girls That We Meet With a Shrug." *The New York Times*, 23 Mar. 2024. *NYTimes.com*, https://www.nytimes.com/2024/03/23/opinion/deepfake-sex-videos.html.
- *#MyImageMyChoice*. https://myimagemychoice.org.

Implementation Resource:

- *URL List*. https://www.iwf.org.uk/our-technology/our-services/url-list.

MAINTAIN: MITIGATION #2

When reporting to NCMEC, use the Generative AI File Annotation

For companies reporting to NCMEC via their API, utilize the "`generativeAi`" file annotation when filing a report including Generative AI content. For other reporting mechanisms, manual annotations indicating the reported content is AI-generated may be necessary.

IMPACT	⚡ Significant. Critical to the workflow of analysts who are reviewing this content.
SCOPE	Narrow
VARIABLES	N/A
RELEVANCE	AI Developers, AI Providers (first and third-party), Data Hosting Platforms, Social Platforms. Closed and Open.

© 2024 Thorn

Implementation Resource:

- *CyberTipline Reporting API Technical Documentation*. https://report.cybertip.org/ispws/documentation/#file-annotations.

MAINTAIN: MITIGATION #3

Detect and remove from your platforms known models that were explicitly built to create AIG-CSAM

There are some models (Category 2c, as defined in the Safety Assessment Categories in the Additional Resources section) that have been trained specifically to create AIG-CSAM. The cryptographic hash of these model files are in some cases known. In those cases, detect and remove from your platforms those models that share the same cryptographic hash. Similarly, search services, as laid out in their CSAM policies, should remove links to Category 2c models.

IMPACT	Significant. Limit the distribution and spread of models that have been built to create AIG-CSAM, and thereby limit the distribution and spread of AIG-CSAM.
SCOPE	Medium
VARIABLES	Trusted sources for hashes
RELEVANCE	AI Providers (third-party), Social Platforms, Search Engines. Open.

Implementation Resource:

- See the "Model Safety Assessment: Known AIG-CSAM Models" section in "Additional Resources" below for more details regarding a dataset containing hashes of Category 2 models.

MAINTAIN: MITIGATION #4

Retroactively assess currently hosted generative models, updating them with mitigations in order to maintain platform access

Some generative models may already have been hosted without undergoing a safety assessment. For first-party AI Providers: assess these currently hosted models for their potential to generate AIG-CSAM and CSEM. For models that are assessed and found to be in Category 2a or 2b, temporarily remove these models from your platforms, restoring them after they have been updated with mitigations in place. Models in Category 2c should be removed from your platforms.

If retraining a model, or other mitigations like model editing are impractical or not possible, restrict the model to hosted-generation only. By doing this, you can employ prompt filtering and other measures to prevent abuse, as well as prevent downloads of model weights or use of the model in private, offline settings.

© 2024 Thorn

For third-party AI Providers: where possible, in a similar fashion as described above, directly assess currently hosted generative models for their potential to generate AIG-CSAM and CSEM before hosting. Where this is currently infeasible, third-party AI Providers should instead require developers to fill out the child safety section of their model card. Third-party AI Providers should then use this child safety section to assess whether the model satisfies your internal child safety policies, or has high likelihood of being in Category 2a or 2b. Use this information to make a decision on whether to allow the model to continue to be hosted, or require the developer to incorporate mitigations before re-hosting. Models in Category 2c should be removed from your platform.

IMPACT	Significant. Limit the distribution and spread of models that can be used to generate AIG-CSAM and CSEM, and thereby limit the distribution and spread of AIG-CSAM and CSEM.
SCOPE	Medium
VARIABLES	Evaluation criteria, evaluation dataset
RELEVANCE	AI Developers, AI Providers (first and third-party). Open.

Resources:

- See the resources listed in the mitigation titled "Assess generative models before access."

MAINTAIN: MITIGATION #5

Detect, report, remove and prevent CSAM, AIG-CSAM and CSEM on your platforms

Regardless of whether users are employing generative AI to accelerate these harms, prioritize detecting, reporting, removing, and preventing sexual harms against children on your platforms. See the "Reporting CSAM and AIG-CSAM" section in "Additional Resources" below for more details. Thoroughly document procedures for detecting, reporting, removing, and preventing these materials, including the process by which potential CSAM should be reported and preserved, consistent with the company's legal obligations.[14]

IMPACT	Significant. Supports victim identification, stops re-victimization, prevents abuse from occurring in the first place.
SCOPE	Medium

14 See *supra* note 1.

© 2024 Thorn

VARIABLES	Precision/recall of automated detection solutions, in house vs. external detection solutions, computational resources
RELEVANCE	Social Platforms.

Resources:

- See the resources listed in the mitigation titled "Detect, remove and report CSAM and CSEM from your training data."

MAINTAIN: MITIGATION #6

Invest in tools to protect content from AI-generated manipulation

There is recent and ongoing research focused on building tools and techniques to protect content from AI-generated manipulation. Organizations should continually invest in this research and new tools, actively collaborating and sharing with others safety-enhancing tools, best practices, processes and technologies.

IMPACT	Significant. Makes it more difficult for bad actors to take benign imagery of children, sexualize it, and use that content to scale their sexual extortion efforts.
SCOPE	Wide
VARIABLES	Computational resources, robustness
RELEVANCE	AI Developers, Social Platforms. Closed and Open.

Informational Resource:

- *Glaze - Protecting Artists from Generative AI*. https://glaze.cs.uchicago.edu.
- Salman, Hadi, et al. *Raising the Cost of Malicious AI-Powered Image Editing*. arXiv:2302.06588, arXiv, 13 Feb. 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2302.06588.

MAINTAIN: MITIGATION #7

Maintain the quality of your mitigations

Whether considering data policies, usage policies, content provenance solutions, etc., ensure that your mitigations are still robust, performant and applicable for new model releases.

IMPACT	Significant. Mitigates the possibility that the mitigations become out of date in comparison to the new model.

© 2024 Thorn

SCOPE	Depending on the mitigation, narrow to wide
VARIABLES	See the variables across Develop and Deploy mitigations
RELEVANCE	AI Developers, AI Providers (first and third-party), Data Hosting Platforms, Social Platforms, Search Engines. Closed and Open.

Resources:

- See the resources across Develop and Deploy mitigations.

MAINTAIN: MITIGATION #8

Disallow the use of generative AI to deceive others for the purpose of sexually harming children. Explicitly ban AIG-CSAM from your platforms.

Document to users on your platforms that using generative AI to deceive other users for the purpose of sexually harming children (including the creation of CSAM and CSEM) is prohibited. Document to users that generating and sharing AIG-CSAM on your platforms is prohibited.

Ensure that this documentation is accessible, easy to find, regularly updated and easy to understand. Advise users of these policies at sign up, and provide periodic reminders, e.g. via educational prompts, warnings, quick tips, shortcuts or summaries.

IMPACT	Incremental. Provides clarity and transparency outside the organization on ethical lines and boundaries.
SCOPE	Narrow
VARIABLES	Scale of announcement
RELEVANCE	Social Platforms.

Informational Resource:

- *Discord's Commitment to Teen and Child Safety*. https://discord.com/safety/commitment-to-teen-child-safety.
- "Integrity and Authenticity." *TikTok*, 8 Mar. 2023, https://www.tiktok.com/community-guidelines/en/integrity-authenticity.
- "Our Approach to Labeling AI-Generated Content and Manipulated Media." *Meta*, 5 Apr. 2024, https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media.
- *Our Synthetic and Manipulated Media Policy | X Help*. https://help.twitter.com/en/rules-and-policies/manipulated-media.
- "PAI's Responsible Practices for Synthetic Media." *Partnership on AI - Synthetic Media*, https://syntheticmedia.partnershiponai.org.

© 2024 Thorn

MAINTAIN: MITIGATION #9

Leverage Open Source Intelligence (OSINT) capabilities

Use OSINT to understand how your platforms, products, and models are potentially being abused by bad actors. For example, understand from dark web chatter how bad actors may be: manipulating your platforms, evading detection, building new models specifically to produce AIG-CSAM, or using a combination of prompts (generic and niche) on existing models. Leverage these insights to maintain the robustness of your mitigations, as well as to source information on new product vulnerabilities.

IMPACT	Incremental. Provides organizations with an ability to find vulnerabilities that may not have been captured during red teaming or other testing but may be found through OSINT efforts.
SCOPE	Medium
VARIABLES	Cadence of monitoring, monitoring conducted by resources in-house vs. partnering with external monitoring services
RELEVANCE	AI Developers, AI Providers (first and third-party), Social Platforms. Closed and Open.

Informational Resource:

- "S1E9 | Where Trust and Safety Meets OSINT." *Authentic8*, https://authentic8.com/needlestack/s1e9-where-trust-and-safety-meets-osint.

Implementation Resource:

- ActiveFence, https://www.activefence.com.
- Resolver, https://www.resolver.com.
- Graphika, https://graphika.com.

Further Opportunities

Continuously engage across industry and child safety experts

> The tech ecosystem could collaborate to keep mitigations up to date as new emerging threats materialize, and build up an understanding of downstream impact in conjunction with public stakeholders (e.g. AI Verify Foundation, Center for AI Safety, Center for Democracy and Technology, eSafety, Frontier Model Forum, IWF, NCMEC, Partnership on AI, Responsible AI UK, Thorn).

© 2024 Thorn

Build a shared resource of Category 2c models

> The tech ecosystem could contribute to shared datasets of known Category 2c models to allow for prompt detection, removal and delisting of these models from third-party AI Providers, Social Platforms and Search Engines.

Use a consistent template across industry for reporting AIG-CSAM

> The tech ecosystem could accelerate hotline and law enforcement workflows by using a consistent template across industry when reporting AIG-CSAM to hotlines and law enforcement. This would support accelerated prioritization and triage of incoming content, as well as open the door for creating hash lists of AIG-CSAM that can then be used by the ecosystem to detect and remove this content.

Proactively protect users' content from unwanted AI-generated manipulation

> As technical research on protecting user content from AI-generated manipulation matures, Social Platforms and AI Developers could proactively protect their users. They could offer API endpoints to users that add perturbations to their content, such that the content is more robust to AI-generated manipulation. Before a minor uploads their content to a Social Platform, they could by default alter images of minors to make it more difficult to use AI-generated manipulation to sexualize their content.

Provide transparency reports and undergo third-party audits

> Tech organizations could provide transparency reports and undergo third-party audits, to demonstrate their commitments and allow for evaluating their efforts to make their platforms and products safe.

Ideate and implement Safety by Design in tooling associated with model development

> AI Developers could consider Safety by Design in not just the models they build, but also in the associated tooling they build and release (e.g. Automatic1111, InvokeAI, ComfyUI). These types of tools are also misused by bad actors to further sexual harms against children. For example, bad actors use tools that allow for the easy merging of multiple models to combine models intended to produce depictions of children, with models intended to produce adult pornography. This merging results in new models they use to produce AIG-CSAM. Incorporating restrictions built into these tools to disallow such merges could be a first step in addressing this issue.

Potentially Problematic Downstream Implications

We recognize that some of the recommended mitigations above have potentially problematic downstream implications, beyond their intended positive impact. We include below some of these possibilities, as well as our suggested response.

Cultural distinctions in content moderation

> See the previous Develop discussion.

© 2024 Thorn

Bias in automated content moderation solutions

> See the previous Develop discussion.

Wellness implications in content moderation

> See the previous Develop discussion.

False positives from automated content moderation solutions

> See the previous Deploy discussion.

© 2024 Thorn

Additional Resources

Reporting CSAM and AIG-CSAM

Depending on where your business is located, there are different governing institutions for reporting CSAM and AIG-CSAM. We list below some of the major institutions, as well as share a resource for hotline referral within the InHope network and the Child Helpline International network list with information on additional local support options for affected children.

- e-Safety Commissioner - AU. https://www.esafety.gov.au/report.
- National Center for Missing and Exploited Children - USA. https://www.missingkids.org/gethelpnow/cybertipline.
- Canadian Centre for Child Protection - Canada. https://www.cybertip.ca/en/report.
- Internet Watch Foundation - UK. https://www.iwf.org.uk/en/uk-report.
- InHope. https://inhope.org/EN#hotlineReferral.
- Child Helpline International. https://childhelplineinternational.org/helplines.

Model Card: Child Safety

While different organizations use different templates for model cards, below are some suggested items to include in your existing template. In whatever existing sections are best suited for these questions, and where relevant, instruct the model developer to include answers to the following questions:

- Is there any CSAM, image/video/audio content sexualizing children, grooming text, sexual extortion text, CSAM advertising, CSAM solicitation, and text promoting sexual interest in children, in the training data, the testing data or the render? If yes, please remove and report.
- For open source models, is your data a mix of adult and child imagery? If yes, why have you chosen to combine these two?
- What detection measures have been run on the data to ensure CSAM and CSEM do not appear in your dataset? Link to resources used for detection.
- Did you use a content provenance mechanism in your model? If yes, please describe and add the necessary links/documentation. If not, please explain why.
- Did you add an inbuilt detection layer or any safety filter in your model? If yes, please describe and add the necessary links/documentation. If not, please explain why.
- What is your model capable of producing? Please describe and add the necessary links/documentation.
- Have you gone through red teaming to understand what your model is capable of producing? If yes please describe and add the necessary links/documentation. If not, please explain why.

© 2024 Thorn

- Is your model incapable of producing AIG-CSAM and CSEM? Please add the necessary links/documentation.
- Has your model been through a safety assessment? If yes, please describe and add the necessary links/documentation. If not, please explain why.

Model Safety Assessment

Safety Assessment Categories

N.B. currently scoped to the image/video modality

When assessing a model for its potential to generate AIG-CSAM, we recommend categorizing the model into the following hierarchy detailing categories of propensity:

1

Model is **incapable** of generating AIG-CSAM

2

Model is **capable** of generating AIG-CSAM

2a. When **explicitly** prompted to do so

2b. **Inadvertently** without explicit prompting

2c. Has been **optimized** specifically for the purpose of generating AIG-CSAM

Safety Assessment Dataset

Thorn has curated a dataset of prompts and configuration parameters that, in combination with the specified Category 2 model, are known to generate AIG-CSAM. This dataset can be used to help inform whether a model is in Category 2 of the Safety Assessment Categories outlined above. This dataset should not be treated as providing assurance that a model is **not** in Category 2, but rather can be used to help assess whether a model **is** in Category 2. For more information and access to this dataset, please reach out to tech-standards@wearethorn.org.

Known AIG-CSAM Models

Thorn has curated a dataset of hashes of models that are known to be in Category 2 of the Safety Assessment Categories outlined above. This dataset should not be treated as representing the entire set of models that exist in these categories, but as a subset of the full set of models that exist in these categories. For more information and access to this dataset, please reach out to tech-standards@wearethorn.org.

© 2024 Thorn

Identity Verification: Settings and Possibilities

N.B. This is primarily applicable for closed organizations.

We provide below some suggested approaches across various low and high risk settings.

Search-Based AI Tools

> Search-based AI tools could return harmful or inappropriate content, potentially even CSAM, during a query. Covers both basic searches and advanced searches (e.g. searching within a specialized database of images or videos). Low risk.

Generative Capabilities

> Generative AI tools can create content that might be used for harm, including AIG-CSAM, or manipulate existing content in harmful ways. High risk.

Ability to Upload Content

> Users might upload existing harmful content, including CSAM, or innocent images that could be manipulated. High risk.

User Information Possibilities

CAPTCHA, Email Verification, Name, Terms of Use Agreements, Multi-factor Authentication, IP logs for product use, Device ID and Cookies, Phone number, ID verification

© 2024 Thorn

Authors

Thorn - Dr. Rebecca Portnoff, Michael Simpson, Rob Wang, Tim O'Gorman

All Tech Is Human - David Polgar, Rebekah Tweed, Professor Renee Cummings

AWS AI - Mikaela C. Myers, Peter W. Hallinan

Civitai - Micah Schaffer

Hugging Face

Inflection - Solianna Herrera

Metaphysic - Beni Beeri Issembert

Stability AI - Ben Brooks

Teleperformance - Farah Lalani

Preferred Citation

Thorn & ATIH. (2024). *Safety by Design for Generative AI: Preventing Child Sexual Abuse.* Thorn Repository. Available at https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf.

© 2024 Thorn

Acknowledgements

We acknowledge the contributions made by many experts in providing their review and insights on this document, including the following:

Akash Pugalia, Global President, Media, Entertainment, Gaming and Trust & Safety, Teleperformance

Alicia Hurst, Staff Product Manager, Truepic

Alison Moyer, Corporate Counsel, Teleperformance

Amanda Towler, Founder, SUDOGirl Consulting

Anthropic

Australian eSafety Commissioner

Bindu Sharma, Vice President Global Policy & Industry Alliances, Managing Director Asia Pacific, International Centre for Missing & Exploited Children

Christian Cardona, Program and Research Lead, Artificial Intelligence and Media Integrity, Partnership on AI

Claire Leibowicz, Head of AI and Media Integrity, Partnership on AI

Dan Sexton, CTO, Internet Watch Foundation

Daniel Fried, Assistant Professor, Carnegie Mellon University

David Thiel, Chief Technologist, Stanford Internet Observatory

Demian Ahn, Maneesha Mithal, and Clinton Oxford of Wilson Sonsini Goodrich & Rosati

Matthew Daggett, Humanitarian Assistance and Disaster Relief Systems Group, Massachusetts Institute of Technology Lincoln Laboratory

Melissa Stroebel, Head of Research & Insights, Thorn

NAVER Z Global Affairs

Shailey Hingorani, Head of Policy, Advocacy and Research, WeProtect Global Alliance

Dr. Shaunagh Downing, Research and Development Engineer, CameraForensics

Tom Thorley, Director of Technology, The Global Internet Forum to Counter Terrorism

Dr. Yacine Jernite, Machine Learning & Society Lead, Hugging Face

© 2024 Thorn

References

1. "Issue and Response | End Violence." *End Violence Against Children*, https://www.end-violence.org/node/7939.
2. *Generative AI – Position Statement | eSafety Commissioner*. https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai.
3. "Safety by Design." *ESafety Commissioner*, https://www.esafety.gov.au/industry/safety-by-design.
4. "Blueprint for an AI Bill of Rights | OSTP." *The White House*, https://www.whitehouse.gov/ostp/ai-bill-of-rights.
5. "Evaluating social and ethical risks from generative AI." *Google DeepMind*, 19 Oct. 2023, https://www.deepmind.com/blog/evaluating-social-and-ethical-risks-from-generative-ai.
6. "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." *The White House*, 30 Oct. 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.
7. "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI." *The White House*, 12 Sept. 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai.
8. United Nations. *Guiding Principles on Business and Human Rights Implementing the United Nations "Protect, Respect and Remedy" Framework*. United Nations, 2011, https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.
9. "PAI's Responsible Practices for Synthetic Media." *Partnership on AI - Synthetic Media*, https://syntheticmedia.partnershiponai.org.
10. "PAI's Guidance for Safe Foundation Model Deployment." *Partnership on AI*, https://partnershiponai.org/modeldeployment.
11. Schuett, Jonas, et al. *Towards Best Practices in AGI Safety and Governance: A Survey of Expert Opinion*. arXiv:2305.07153, arXiv, 11 May 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2305.07153.
12. "CyberTipline Data." *National Center for Missing & Exploited Children*, https://www.missingkids.org/content/ncmec/en/cybertiplinedata.html.
13. "Communications Decency Act of 1996 (CDA)." *Glossary | Practical Law*, https://content.next.westlaw.com/Glossary/PracticalLaw/I0f9fea42ef0811e28578f7ccc38dcbee.
14. "How AI is Being Abused to Create Child Sexual Abuse Imagery." *IWF*, Oct. 2023, https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf.
15. Paltieli, Guy. "How Predators Are Abusing Generative AI." *ActiveFence*, 18 Apr. 2023, https://www.activefence.com/blog/predators-abusing-generative-ai.

© 2024 Thorn

16. Thiel, D., Stroebel, M., and Portnoff, R. *Generative ML and CSAM: Implications and Mitigations*. Stanford Digital Repository, June 2023, https://doi.org/10.25740/jv206yg3793.

17. *Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes*. https://www.ic3.gov/Media/Y2023/PSA230605.

18. "Children Are Using AI to Bully Their Peers Using Sexually Explicit Generated Images, eSafety Commissioner Says." *ABC News*, 15 Aug. 2023. *www.abc.net.au*, https://www.abc.net.au/news/2023-08-16/esafety-commisioner-warns-ai-safety-must-improve/102733628.

19. Jargon, Julie. "Fake Nudes of Real Students Cause an Uproar at a New Jersey High School." *WSJ*, 2 Nov. 2023, https://www.wsj.com/tech/fake-nudes-of-real-students-cause-an-uproar-at-a-new-jersey-high-school-df10f1bb.

20. "AI-generated naked child images shock Spanish town of Almendralejo." *BBC News*, 23 Sep. 2023. *www.bbc.co.uk*, https://www.bbc.co.uk/news/world-europe-66877718.

21. *Child Molesters: A Behavioral Analysis*. https://www.missingkids.org/content/dam/missingkids/pdfs/publications/nc70.pdf.

22. Insoll, Tegan, et al. *CSAM Users in the Dark Web: Protecting Children Through Prevention*. Suojellaan Lapsia ry. ReDirection Survey Report, 2021.

23. Engler, Maggie. *Considerations of the Impacts of Generative AI on Online Terrorism and Extremism*. Sep. 2023, https://gifct.org/wp-content/uploads/2023/09/GIFCT-23WG-0823-GenerativeAI-1.1.pdf.

24. Funk, Allie, et al. *Freedom On The Net 2023: The Repressive Power of Artificial Intelligence*. Oct. 2023, https://freedomhouse.org/sites/default/files/2023-10/Freedom-on-the-net-2023-Digital-Booklet.pdf.

25. "No Laws Protect People From Deepfake Porn. Here's How Some Victims Fought Back." *Bloomberg.Com*, 29 Nov. 2023. *www.bloomberg.com*, https://www.bloomberg.com/news/features/2023-11-29/deepfake-porn-victims-learn-us-has-no-federal-laws-to-fight-it.

26. "Meta and OpenAI Have Spawned a Wave of AI Sex Companions—and Some of Them Are Children." Fortune, https://fortune.com/longform/meta-openai-uncensored-ai-companions-child-pornography/.

© 2024 Thorn