Reducing the Risk of Synthetic Content: Preventing generative AI from producing child sexual abuse material

Summary

Beginning in July of 2023, Thorn and All Tech Is Human organized a working group consisting of representatives from leading generative AI companies, to collaboratively define, align on, and endorse a set of Safety by Design mitigations to prevent the misuse of generative artificial intelligence (AI) to further sexual harms against children. The output of this work is a paper that is nearing completion, and an associated set of commitments that are currently being secured, from the involved companies. In this comment to NIST, we provide:

- an issue overview
- the framework for our response
- definitions[1] of various relevant terms
- some of the key mitigations that are recommended in the paper

Issue Overview

We are at a crossroads with generative 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 [15], misuse of generative AI has profound implications for child safety, 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, 21, 25]. 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

[1] Note that the companies listed in the definitions section are not necessarily representative of the companies involved with the working group

active harm's way. The expanding prevalence of AIG-CSAM is growing that haystack even further, making victim identification more difficult.

- Creates new ways to victimize and re-victimize children: This same technology is used to newly victimize children, as bad actors can now easily 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, 24]. 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 is being used by children to bully and harass others [4, 16, 23].
- Generates more demand: The growing frequency of AIG-CSAM generates more demand, desensitizing society to the sexualization of children and growing the appetite for CSAM [1]. Bad actors use this technology to produce AIG-CSAM and other sexualizing content of wholly fictional minors that depict egregious abuse; this material is found collocated with CSAM. Research indicates a link between engaging in this type of material and contact offending, where normalization of this material also contributes to other harmful outcomes for children [3].
- 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, 25]. 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 [12, 22].

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 [22]. 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.

The mitigations below are written to align and build off of existing guidance and commitments [2, 8, 9, 10, 12, 13, 19, 20, 22, 24]. They are written to be tactical, actionable, and multidisciplinary. They are written such that a technical, policy, product, or trust and safety team could enact them with minimal friction.

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.

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.

Category 1 model: A model that is incapable of generating AIG-CSAM.

Category 2 model: A model that is capable of generating AIG-CSAM.

Category 2a model: A model that is capable of generating AIG-CSAM when explicitly prompted to do so

Category 2b model: A model that inadvertently generates AIG-CSAM without explicit prompting

Category 2c model: A model that has been optimized specifically to generate AIG-CSAM

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.

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 created to mitigate the risk of online harms specific to minors. Policies may

include, but are not limited to: CSAM, child sexualization and abuse, child grooming, child endangerment, 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, and/or provide 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.

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.

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

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

Hosted-generation: A form of deployment such that a model cannot be downloaded into a private, offline setting, but users can (within the confines of the hosting platform) still use the model to generate content.

Machine Learning / Artificial Intelligence (ML/AI): The field of study that gives computers the ability to learn without explicitly being programmed, and/or imitating 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.

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).

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.

Red teaming: The practice of stress testing systems - physical or digital - to find flaws, weaknesses, gaps, and edge cases. This can also be referred to as "adversarial testing" or "safety evaluations". N.B. In this document, when we refer to "safety assessment" it is distinguished from red teaming/adversarial testing/safety evaluation 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 engines: 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 creator, or source, of the content.

Mitigations

Develop

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 data that have a known risk of containing CSAM and CSEM, e.g. by blacklisting in your data collection pipeline sites known for proliferating CSAM.

Thoroughly document procedures for verifying that datasets do not contain CSAM, as well as processes for reporting and preserving such materials, where required and/or permitted by law.[2] Train employees involved in model training on those procedures.
Relevance: Closed and Open. AI Developers.

2. **Scan for, remove, and report CSAM from your training data**: If you cannot determine whether a dataset has been audited for CSAM and CSEM, use available tools to identify this abuse data in your datasets and ensure that it is excluded prior to training your models. Report the content you have found to governing authorities where applicable. Thoroughly document your procedures for scanning and removing CSAM from training data, which should include the process for promptly reporting and preserving CSAM where required and/or permitted by law.[3] Train employees involved in model training on those procedures.
Relevance: Closed and Open. AI Developers, Data Hosting Platforms.

3. **Separate depictions/representations of children from adult sexual content in your training datasets**: When training a model, do not include images/videos of children, or audio recordings of children in datasets that contain adult sexual content. Your objective should be eliminating the possibility of adversarial actors creating depictions of sexual content that also contain children. Make best efforts to prevent your models from having both content of children[4] 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.
Relevance: Closed and Open. AI Developers, Data Hosting Platforms.

4. **Conduct red teaming**: Incorporate structured, scalable, and consistent stress testing of your model for AIG-CSAM and CSEM. Update your model accordingly

[2] 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.
[3] *See supra* note 1.
[4] Note that you cannot train your model on personal information provided by children under 13 without parental consent under the U.S. Children's Online Privacy Protection Act, and that different countries have different restrictions on use of data involving children.

to mitigate for the issues you discover. Red teaming should happen throughout the development process, not just in anticipation of model release. As the model is iteratively built to increase its capabilities, it should also be iteratively red-teamed to understand and mitigate for misalignment. Ensure that after each iterative 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 no longer has the capabilities 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.[5]

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[6]. Train employees responsible for red teaming and model evaluation on these procedures.
Relevance: Closed and Open. AI Developers.

5. **Include content provenance**: Include indicators of provenance by default in any image or video that your model outputs. If possible, and where relevant, include details such as which portions of the content were generated vs. not, and other characteristics of the model[7]. Ensure that the provenance[8] of your

[5] One option is for "propensity testing". This could involve testing for a model's likelihood to produce AIG-CSAM, by assessing:

- is the model capable of producing adult sexual content, including that depicting a specific individual
- is the model 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, 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.

[6] *See supra* note 1.

[7] E.g. model version, settings, hashed weights

[8] 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.

model's generated image or video content can be identified by CSAM hotlines (e.g. NCMEC) and relevant law enforcement (e.g. HSI C3, ICAC).

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 watermark into all generated images) such that it is more difficult for the content provenance to be disabled.
Relevance: Closed and Open. AI Developers

Deploy

1. **Scan for abusive content in inputs and outputs:** Scan for input prompts intended to produce AIG-CSAM and CSEM. Similarly, scan 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. Set up content moderation flows for outputs. Thoroughly document procedures on detecting abusive content in inputs and outputs, which should include the process for reporting and destroying CSAM, consistent with the company's legal obligations.[9] Train employees on such procedures.
Relevance: Closed. AI Developers, AI Providers.

2. **Assess models before access:** Assess models for their potential to generate AIG-CSAM and CSEM before the models are hosted on your platform. For models that are assessed and found to be in Category 2a or 2b, 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 should not be hosted on your platform[10].

[9] *See supra* note 1.
[10] Thorn has curated a dataset of hashes of models that are known to be in either Category 2a, 2b or 2c. 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.

Attempting to generate AIG-CSAM may implicate local law. Consult with legal counsel on this matter. Regardless, it is possible for safety assessments to be carried out such that due regard is given for the regulatory bounds on those conducting the evaluations.[11]
Relevance: Closed and Open. AI Providers.

3. **Include user reporting, feedback or flagging options**: Include a pathway for users to report content the model produces that may violate the model's child safety policies, to the organization's internal trust and safety team, or equivalent team. Include a pathway for users to report models that generate AIG-CSAM and CSEM to the organization's internal trust and safety team, or equivalent team.

   Ensure these pathways allow for in real time reporting and in application flagging/feedback, to reduce user barriers to reporting. In response to user reports about potential violations of the model's child safety policies, provide links to support services, breaking support services down by regional location. Provide the option for timely feedback loops, such that users can be informed of steps taken post the report. 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.[12]
   Relevance: Closed and Open. AI Developers, AI Providers.

4. **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. [13] This concept may bridge into model maintenance. Any enforcement of child safety policies–such as account

[11] 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.
[12] *See supra* note 1.
[13] 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.

suspension or blocking–should be performed in a manner that allows the company to preserve information sufficient to meet any legal requirements.
Relevance: Closed. AI Developers.

Maintain

1. **Scan for, and remove from platform known models that were explicitly built to create AIG-CSAM**: There are some models (Category 2c) that have been trained specifically to create AIG-CSAM; the cryptographic hash of these model files are in some cases known[14]. In those cases, scan and remove from your platform those models that share the same cryptographic hash. Similarly, search engines should remove links to Category 2c models.
Relevance: Open. AI Providers, Social Platforms, Search Engines.

2. **Retroactively assess currently hosted models, updating them with mitigations in order to maintain platform access**: Some models may already have been hosted without undergoing a safety assessment. Evaluate 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 platform, restoring them after they have been updated with mitigations in place. Models in Category 2c should be removed from your platform[15].

   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.
   Relevance: Open. AI Developers, AI Providers.

3. **Remove services for "nudifying" images of children from search results:** Search engines should delist links to sites that provide services and tutorials for "nudifying" 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.
   Relevance: Search Engines.

[14] *See supra* note 10.
[15] *See supra* note 10.

4. **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.
   Relevance: Closed and Open. AI Developers, AI Providers, Data Hosting Platforms, Social Platforms, Search Engines.

References

1. *Artificial Intelligence and the Exploitation of Children*. Sep. 2023, https://www.naag.org/wp-content/uploads/2023/09/54-State-AGs-Urge-Study-of-AI-and-Harmful-Impacts-on-Children.pdf.
2. "Blueprint for an AI Bill of Rights | OSTP." *The White House*, Oct. 2022, https://www.whitehouse.gov/ostp/ai-bill-of-rights.
3. Christensen, Larissa S., and Noah Vickery. "The Characteristics of Virtual Child Sexual Abuse Material Offenders and the Harms of Offending: A Qualitative Content Analysis of Print Media." *Sexuality & Culture*, May 2023, pp. 1–15. *PubMed Central*, https://doi.org/10.1007/s12119-023-10091-1.
4. "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.
5. "Communications Decency Act of 1996 (CDA)." *Glossary | Practical Law*, https://content.next.westlaw.com/Glossary/PracticalLaw/I0f9fea42ef0811e28578f7ccc38dcbee.
6. "CyberTipline Data." *National Center for Missing & Exploited Children*, http://www.missingkids.org/content/ncmec/en/cybertiplinedata.html.
7. 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.
8. "Evaluating Social and Ethical Risks from Generative AI." *Google DeepMind*, Oct. 2023, https://www.deepmind.com/blog/evaluating-social-and-ethical-risks-from-generative-ai.
9. "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." *The White House*, 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.
10. "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI." *The White House*, 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.

11. 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.
12. "Generative AI – Position Statement." *eSafety Commissioner*, https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai.
13. *Guiding Principles on Business and Human Rights*. United Nations, Apr. 2011, https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.
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. "Issue and Response | End Violence." *End Violence Against Children*, https://www.end-violence.org/node/7939.
16. Jargon, Julie. "Fake Nudes of Real Students Cause an Uproar at a New Jersey High School." *WSJ*, https://www.wsj.com/tech/fake-nudes-of-real-students-cause-an-uproar-at-a-new-jersey-high-school-df10f1bb.
17. "Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes." *FBI*, June 2023, https://www.ic3.gov/Media/Y2023/PSA230605.
18. "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.
19. "PAI's Responsible Practices for Synthetic Media." *Partnership on AI - Synthetic Media*, Feb. 2023, https://syntheticmedia.partnershiponai.org.
20. "PAI's Guidance for Safe Foundation Model Deployment." *Partnership on AI,* Oct. 2023, https://partnershiponai.org/modeldeployment.
21. Paltieli, Guy. "How Predators Are Abusing Generative AI." *ActiveFence*, Apr. 2023, https://www.activefence.com/blog/predators-abusing-generative-ai.
22. "Safety by Design." *eSafety Commissioner*, https://www.esafety.gov.au/industry/safety-by-design.
23. "AI-generated naked child images shock Spanish town of Almendralejo." *BBC*, Sep. 2023, https://www.bbc.co.uk/news/world-europe-66877718.
24. 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.
25. 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.

Additional Resources

1. 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.
2. Carlini, Nicholas, et al. *Extracting Training Data from Large Language Models*. arXiv:2012.07805, arXiv, 15 June 2021. *arXiv.org*, https://doi.org/10.48550/arXiv.2012.07805.
3. Carlini, Nicholas, et al. *Extracting Training Data from Diffusion Models*. arXiv:2301.13188, arXiv, 30 Jan. 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2301.13188.
4. 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://carrcenter.hks.harvard.edu/files/cchr/files/200228_ccdp_neal_cohen.pdf.
5. Dodge, Jesse, et al. "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2021, pp. 1286–305. *ACLWeb*, https://doi.org/10.18653/v1/2021.emnlp-main.98.
6. Fernandez, Pierre, et al. *The Stable Signature: Rooting Watermarks in Latent Diffusion Models*. arXiv, 26 July 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2303.15435.
7. Gandikota, Rohit, et al. Unified Concept Editing in Diffusion Models. arXiv, 2023. *arXiv.org*, https://arxiv.org/pdf/2308.14761.pdf.
8. 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.
9. Mitchell, Eric, et al. *Memory-Based Model Editing at Scale*. arXiv, 13 June 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2206.06520.
10. Mitchell, Eric, et al. *Fast Model Editing at Scale*. arXiv, 13 June 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2110.11309.
11. Okawa, Maya, et al. *Compositional Abilities Emerge Multiplicatively: Exploring Diffusion Models on a Synthetic Task*. arXiv:2310.09336, arXiv, 29 Dec. 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2310.09336.
12. Ramesh, Rahul, et al. *How Capable Can a Transformer Become? A Study on Synthetic, Interpretable Tasks*. arXiv:2311.12997, arXiv, 21 Nov. 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2311.12997.
13. 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.
14. 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.

15. Somepalli, Gowthami, et al. *Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models. arXiv:2212.03860, arXiv, 12 Dec. 2022. arXiv.org, https://doi.org/10.48550/arXiv.2212.03860*.
16. Thiel, D. "Identifying and Eliminating CSAM in Generative ML Training Data and Models". *Stanford Digital Repository*. Dec 2023, https://doi.org/10.25740/kh752sm9123.
17. 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.
18. Yu, Ning, et al. *Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data*. arXiv, 17 Mar. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2007.08457.
19. 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.