THORN

Openness in AI Request for Comment

Introduction

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 NTIA, in response to a subset of the questions enclosed in the request for comment, Thorn provides:

- Our perspective, in the intersection of machine learning/AI and child safety
- Select mitigations lifted from the paper written by the working group

Questions

2) a) What, if any, are the risks associated with widely available model weights? How do these risks change, if at all, when the training data or source code associated with fine tuning, pretraining, or deploying a model is simultaneously widely available?

One concrete risk that is already manifesting as a harm occurring today, is the misuse of broadly shared and open source foundation models to make AI-generated child sexual abuse material (AIG-CSAM) [1, 2, 3]. This technology is used to newly victimize children, as bad actors can now easily sexualize benign imagery of a child to scale their sexual extortion efforts, using generative AI to scale the creation of content necessary to target a child [4]. This technology is further used in bullying scenarios, where sexually explicit AI-generated imagery is being used by children to bully and harass others [5, 6, 7].

This risk is exacerbated when the source code associated with fine-tuning is simultaneously widely available. Bad actors use this technology to perpetrate re-victimization by fine-tuning these broadly shared and open source models on existing child abuse imagery to generate additional explicit images of these children [1, 3]. 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.

3) a) What benefits do open model weights offer for competition and innovation, both in the AI marketplace and in other areas of the economy? In what ways can open dual-use foundation models enable or enhance scientific research, as well as education/training in computer science and related fields?

In his book The Cathedral & the Bazaar, Eric S. Raymond defines open-source software as "the process of systematically harnessing open development and decentralized peer review to lower costs and improve software quality." Open source is not new, and open source is not niche: in fact, it has been estimated that up to 98% of codebases include free and open source software [8].

In the context of machine learning/AI, the open sharing of algorithms, datasets, machine learning frameworks, tools and pre-trained models contributes to the acceleration of development, unblocking what might otherwise be significant bottlenecks in development. Open source also allows for a broader net for peer review, ultimately resulting in more independent review, collaborative development, and accelerated innovation.

For example: at Thorn, we build machine learning/AI technologies to accelerate victim identification, stop re-victimization (the viral spread of child sexual abuse material, or CSAM) and prevent abuse from occurring in the first place. In our mission, we have concretely benefited from access to open source model architectures, code and pre-trained model weights. Having access to those open source resources accelerates our development, making it possible for us to have a positive impact faster.

5) a) What model evaluations, if any, can help determine the risks or benefits associated with making weights of a foundation model widely available?

Model evaluations that focus on a model's capabilities or propensity to generate abuse material can help determine the risks associated with making weights of a foundation model widely available. In regards to AIG-CSAM, we recommend the following categorizations for models:

- 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

Standardized safety assessment across industry would allow for consistent and transparent evaluation of a model's propensity for generating AIG-CSAM.

Further, model evaluations focused on the training dataset of the model can also help determine the risks. Evaluating:

- The sources of the data: did the developer avoid ingesting into training data, any data that has a known risk of containing CSAM?
- The content of the data: did the developer audit the data for CSAM? What combination of technologies and human review was used in this auditing process?
- The content of the data: did the developer combine adult sexual content[1] with images/videos of children in the training of their model?

5) d) Are there ways to regain control over and/or restrict access to and/or limit use of weights of an open foundation model that, either inadvertently or purposely, have already become widely available? What are the approximate costs of these methods today? How reliable are they?

Currently, the child safety ecosystem heavily relies on perceptual hashing and cryptographic hashing as mechanisms to detect known CSAM at scale, allowing for accelerated removal of this content from online platforms. In the same way that images and videos can be hashed and matched against a list of verified CSAM, open foundation models in Category 2b, and variants of open foundation models in Category 2c (e.g. models that have been built by fine-tuning open foundation models with CSAM) can also be cryptographically hashed and matched. These models can then be detected and removed from platforms where they are hosted, shared and circulated.

The same limitations that exist in cryptographic hashing for images and videos persist in this arena: these solutions are brittle to minor modifications and changes to the file. However, they allow for detection of unmodified files (the "low hanging fruit"), and are reliable in that modern cryptographic hashing algorithms are generally collision resistant [9].

A similar but distinct intervention opportunity also exists around the services and applications that make use of these open foundation models to provide services for "nudifying" and sexualizing images of children and adults [10]. By delisting links to sites that provide services and tutorials for "nudifying" and sexualizing images of children and adults, search engines can act as a bottleneck, restricting the access and limiting the misuse of these open foundation models.

7) e) What should the role of model hosting services (e.g. HuggingFace, GitHub, etc.) be in making dual-use models with open weights more or less available? Should hosting services host models that do not meet certain safety standards? By whom should those standards be prescribed?

Model hosting platforms play a critical role in the ecosystem, both in terms of benefits and risks. They provide critical resources that accelerate innovation and development. However, they also act as an accelerant for the sharing of models that are misused to generate AIG-CSAM and other sexual harms against children [3].

[1] "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.

With these benefits and risks both in mind, there are multiple opportunities for model hosting platforms to mitigate these risks, while still providing those critical resources that accelerate innovation. Below, we list some of these opportunities:

1. **Include user reporting, feedback or flagging options**: Include a pathway for users to report models that generate AIG-CSAM. 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, provide links to support services. Provide contact details, so that law enforcement and users can reach out with additional queries or feedback.
2. **Assess generative models before access**: Assess generative models for their potential to generate AIG-CSAM and CSEM before the models are hosted. For models that are assessed and found to be in Category 2a or 2b (as defined previously in this comment), 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, model hosting platforms 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 previously in this comment) should not be hosted on these platforms.

   Where this type of scaled assessment is currently infeasible, model hosting platforms should instead require developers to fill out a child safety section of their model card before hosting the model. Model hosting platforms should then use this child safety section to assess whether the model satisfies their internal child safety policies, and/or has a high likelihood of being in Category 2a or 2b. They should 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.

3. **Include prevention messaging for CSAM solicitation**: Prevention and deterrence responses to CSAM solicitation on search engines or functionality is becoming an industry standard. Where model hosting platforms provide text input fields as part of their resources, serve prevention or deterrence messaging (such as prompts, nudges or interstitial warnings which provide users with information).
4. **Incorporate a child safety section into model cards**: Model hosting platforms should update their model card template to include questions on mitigations the developer implemented for child safety.
5. **Detect and remove from platforms known models that were explicitly built to create AIG-CSAM**: There are some models (Category 2c, as defined previously in this comment) that have been trained specifically to create AIG-CSAM. The cryptographic hash of these model files are in some cases known. In those cases, model hosting platforms should detect and remove from their platforms those models that share the same cryptographic hash.

In regards to model assessment in particular, scalable model assessments will be necessary to match the scale and speed of model development. Model hosting platforms should 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. Safety standards should not be prescribed by individual technology companies, but instead be established by existing, third party standard setting institutions, such as NIST and IEEE, in collaboration with civil society organizations.

References

1. 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.

2. Paltieli, Guy. “How Predators Are Abusing Generative AI.” ActiveFence, Apr. 2023, https://www.activefence.com/blog/predators-abusing-generative-ai.

3. 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.

4. “Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes.” FBI, June 2023, https://www.ic3.gov/Media/Y2023/PSA230605.

5. “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.

6. 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.

7. “AI-generated naked child images shock Spanish town of Almendralejo.” BBC, Sep. 2023, https://www.bbc.co.uk/news/world-europe-66877718.

8. Census II of Free and Open Source Software — Application Libraries. The Linux Foundation and The Laboratory for Innovation Science at Harvard, March 2022, https://8112310.fs1.hubspotusercontent-na1.net/hubfs/8112310/LF%20Research/Harvard%20Census%20II%20of%20Free%20and%20Open%20Source%20Software%20-%20Report.pdf.

9. Secure Hash Algorithms | Practical Cryptography for Developers. 19 June 2019, https://cryptobook.nakov.com/cryptographic-hash-functions/secure-hash-algorithms.

10. 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.