PUBLIC SUBMISSION

Received: May 28, 2025
Tracking No. mb8-fj6x-3plc
Comments Due: May 28, 2025 **Submission Type:** API

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0157
Comment on FR Doc # 2025-07332

Submitter Information

Organization: Thorn

General Comment

Hello,

My name is Dr. Rebecca Portnoff, Vice President of Data Science at Thorn. I am submitting on behalf of Thorn a comment in response to the NITRD NCO and NSF request for input on the Development of an Artificial Intelligence (AI) Action Plan. We have uploaded the comment as a PDF file.

As noted in the uploaded PDF as well: This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the 2025 National AI R&D Strategic Plan and associated documents without attribution.

I hope this comment is helpful to your work, and please feel free to reach out if you have any follow up questions!

Cheers,
Rebecca

Attachments

Thorn Comment RFI National AI Strategic Plan

THORN

RFI on the Development of a 2025 National AI R&D Strategic Plan

Introduction

Thorn is a nonprofit that builds technology and conducts research to defend children from sexual abuse. Founded in 2012, the organization equips those on the frontlines with the technology and research they need to protect children from sexual abuse and exploitation in the digital age. Thorn's tools have helped the tech industry detect and report millions of child sexual abuse files on the open web, and connected investigators and non-profits with critical information to help them solve cases faster and remove children from harm. Thorn's research has provided the ecosystem with the necessary insights and issue understanding to build robust interventions across the ecosystem.

Thorn has acted as a leader in preventing the misuse of generative AI for furthering child sexual abuse via our Safety by Design for Generative AI initiative[1], both via providing early visibility to the ecosystem on emerging generative AI enabled sexual harms against children, and via galvanizing the ecosystem to move to prevent the harms. These harms include: offenders complicating victim identification by making photorealistic AI-generated child sexual abuse material[2] (AIG-CSAM) at scale, perpetrating re-victimization by fine-tuning open source and open weight models on existing child abuse imagery to generate additional explicit images of these children[3], and scaling their sexual extortion and harassment efforts by using generative AI to sexualize benign imagery of children, accelerating the creation of content necessary to target a child[4].

---

[1] "Thorn and All Tech Is Human Forge Generative AI Principles with AI Leaders to Enact Strong Child Safety Commitments." Thorn, April 2024, https://www.thorn.org/blog/generative-ai-principles

[2] Thiel, Stroebel, and Portnoff. "Generative ML and CSAM: Implications and Mitigations." Stanford Digital Repository, June 2023, https://doi.org/10.25740/jv206yg3793;
"2024 Update: Understanding the Rapid Evolution of AI-Generated Child Abuse Imagery." IWF, July 2024, https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery;
"Generative AI CSAM is CSAM." NCMEC, Mar 2024, https://www.missingkids.org/blog/2024/generative-ai-csam-is-csam

[3] Thiel, Stroebel, and Portnoff. "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/PSA/Archive/2023/PSA230605;
Thorn. (2025). "Deepfake Nudes & Young People: Navigating a new frontier in technology-facilitated nonconsensual sexual abuse and exploitation." https://info.thorn.org/hubfs/Research/Thorn_DeepfakeNudes&YoungPeople_Mar2025.pdf

Beginning in July of 2023, Thorn and All Tech Is Human (ATIH) organized a working group consisting of representatives from leading generative AI companies, to collaboratively define, align on, and commit to a set of Safety by Design principles and mitigations to prevent the misuse of generative artificial intelligence (AI) to further sexual harms against children. The goal of this initiative was to establish a standard such that, if adopted, these technologies are less capable of producing AIG-CSAM and other content that contributes to the sexual exploitation of children, the content that is created gets detected more reliably, and the distribution of the models, apps and services used to create the content is limited. The output of this work is two fold:

1. Our Safety by Design for Generative AI: Preventing Child Sexual Abuse paper[5] that provides fundamental principles, and mitigations to enact those principles, for building generative AI to prevent the misuse of generative AI technologies to perpetrate, proliferate, and further sexual harms against children. These principles and mitigations cover the full lifecycle of machine learning/AI (develop, deploy, maintain), and apply across several key technology players in the ecosystem (AI Developers, first-party and third-party AI Providers, Data Hosting Platforms, Social Platforms, and Search Engines).
2. Commitments[6] from a set of industry stakeholders, to implement the preventative and proactive principles defined in the paper described above, into their generative AI technologies and products. Companies agreed to adopt these principles, and further to transparently share their progress in taking action on these principles. The initial group of companies who joined into these commitments include: Amazon, Anthropic, Civitai, Google, Meta, Metaphysic, Microsoft, Mistral AI, OpenAI, and Stability AI. Since initial launch, Invoke has also joined into these commitments.

Request for Information

Thorn supports OSTP, NITRD NCO and NSF efforts to drive R&D in artificial intelligence that promotes human flourishing. This flourishing definitionally must include the wellbeing of children. Due to the remit of our organization's mission, the recommendations in this submission focus on one category: research needs and

[5] 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-forgenerative-AI.pdf.
[6] "Thorn and All Tech Is Human Forge Generative AI Principles with AI Leaders to Enact Strong Child Safety Commitments." Thorn, April 2024, https://www.thorn.org/blog/generative-ai-principles

development challenges in AI that the federal government should prioritize over the next 3 to 5 years to ensure that the foundation of these vital technologies are established, from the beginning, in such a way that subsequent rapid innovation leads to children's flourishing, rather than children's abuse and harm.

Significant safety incidents, especially when these incidents have real impacts for American children and their families, could have a chilling effect on this new era of innovation. Leading AI and tech organizations have shown their willingness to engage in our Safety by Design for Generative AI initiative, but we need further coordination at scale, ideally facilitated by the government, to allow industry writ large to continue to innovate quickly and effectively.

Invest in technical research to establish stronger safeguards for open source models, open weight models, and model hosting platforms

The accessibility of open source and open weight models allows for significantly easier customization and optimization of models, opening the door to both opportunities and risks[7]. Offenders fine-tune open source and open weight models on existing child abuse imagery to generate additional explicit images of these children[8]. Nudifying apps are built off of open foundation models, and are used today to sexualize benign depictions of children[9]. Similarly, model hosting platforms can act as an accelerant for the distribution of models that facilitate this type of criminal activity[10]. Without prioritizing development and innovation towards robust safeguards baked into the model itself, we are opening the door for criminals to take advantage of American innovation to sexually abuse and harass our children.

There is an urgent need for scalable, reliable model assessments that are not overly reliant on prompt/response strategies. Current strategies for model assessments are inherently manual: at their core, they all involve evaluating using prompts and assessing

---

[7] "Dual-Use Foundation Models with Widely Available Model Weights." NTIA Report, July 2024, https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf

[8] Thiel, Stroebel, and Portnoff. "Generative ML and CSAM: Implications and Mitigations." Stanford Digital Repository, June 2023, https://doi.org/10.25740/jv206yg3793

[9] Thorn. (2025). "Deepfake Nudes & Young People: Navigating a new frontier in technology-facilitated nonconsensual sexual abuse and exploitation." https://info.thorn.org/hubfs/Research/Thorn_DeepfakeNudes&YoungPeople_Mar2025.pdf

[10] Thiel, Stroebel, and Portnoff. "Generative ML and CSAM: Implications and Mitigations." Stanford Digital Repository, June 2023, https://doi.org/10.25740/jv206yg3793; "a16z Funded AI Platform Generated Images That 'Could Be Categorized as Child Pornography,' Leaked Documents Show." 404 Media, Dec 2023, https://www.404media.co/a16z-funded-ai-platform-generated-images-that-could-be-categorized-as-child-pornography-leaked-documents-show; Harris and Willner. "Was an AI Image Generator Taken Down for Making Child Porn?" IEEE Spectrum, Aug 2024, https://spectrum.ieee.org/stable-diffusion;
"An AI companion site is hosting sexually charged conversations with underage celebrity bots." MIT Tech Review, Feb 2025, https://www.technologyreview.com/2025/02/27/1112616/an-ai-companion-site-is-hosting-sexually-charged-conversations-with-underage-celebrity-bots

outputs. Given the pace and scale of newly released models into the ecosystem, and the specific sensitivities of assessing AIG-CSAM related harms, these strategies are not sufficient[11]. There is further a need for model training strategies that prevent and mitigate adversarial fine-tuning, unlearning and other adversarial optimizations downstream[12]. Similarly, there is a need for content provenance solutions that are robust for settings where the models can be directly modified, such that the signals are incorporated into the media as part of the media generation process[13], rather than as an optional, post-processing step.

As a result, we recommend that the government's R&D plan prioritize research to establish stronger safeguards for open source models, open weight models, and model hosting platforms to advance the state of the art in solutions for scalable model assessments, model training strategies for building models which are robust to downstream adversarial manipulation, and content provenance solutions that are robust for settings where the models can be directly modified.

Invest in technical research to establish stronger safeguards for children's imagery, to proactively protect content from unwanted AI-generated manipulation:
We are already observing today the concrete physical, social and financial harms[14] American children and their families are experiencing due to the misuse of generative AI. Offenders are scaling their sexual extortion and harassment efforts by using generative AI to sexualize benign imagery of children, accelerating the creation of content necessary to target a child[15].

There is an urgent need for technology that proactively protects users' content from unwanted AI-generated manipulation. Solutions that add perturbations[16] to content,

---

[11] Bereska and Gavves. "Mechanistic Interpretability for AI Safety: A Review." arxiv.org, Aug 2024, https://arxiv.org/pdf/2404.14082

[12] Pan et al. "Leveraging Catastrophic Forgetting to Develop Safe Diffusion Models against Malicious Finetuning." NeurIps 2024, https://openreview.net/pdf?id=pR37AmwbOt

[13] 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 ;
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. ;
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.

[14] "Criminals Use Generative Artificial Intelligence to Facilitate Financial Fraud. " FBI, Dec 2024. https://www.ic3.gov/PSA/2024/PSA241203

[15] "Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes." FBI, June 2023, https://www.ic3.gov/PSA/Archive/2023/PSA230605;
Thorn. (2025). "Deepfake Nudes & Young People: Navigating a new frontier in technology-facilitated nonconsensual sexual abuse and exploitation." https://info.thorn.org/hubfs/Research/Thorn_DeepfakeNudes&YoungPeople_Mar2025.pdf

[16] 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.

such that the content is more robust to AI-manipulation, are still nascent and not broadly adopted. With this technology in place, users could protect their own benign imagery from being sexualized by "nudifying" models prior to sharing or uploading this imagery more broadly. Similarly, social platforms hosting user-uploaded content could also provide the option for users to protect their data by offering API endpoints to users that add perturbations to their content (or by default for imagery of minors uploaded to their platform).

As a result, we recommend that the government's R&D plan prioritize research to establish stronger safeguards for children's imagery, to proactively protect content from unwanted AI-generated manipulation.

Conclusion

There is real potential for generative AI to promote human flourishing. In order to accomplish this outcome, it will require thorough understanding of how this technology is currently being used to harm and harass the most vulnerable among us: our children, as well as clear commitments from government and industry leaders alike to prevent this abuse. We wish to express our appreciation to the OSTP, NITRD NCO and NSF for their engagement and leadership in ensuring that this formative technology achieves those positive ends. If any questions arise from review of this request for information, please reach out to Dr. Rebecca Portnoff and Lara Gemar.