

NIST Proposed Zero Draft for a Standard on Documentation of AI Datasets and AI Models RFI

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 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 initiative¹. These harms include: offenders complicating victim identification by making photorealistic AI-generated child sexual abuse material² (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 children³, 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 child⁴.

Thorn has established foundational principles, adopted by major technology companies, to prevent the misuse of generative AI to further sexual harms against children. Thorn has further contributed to and authored other relevant standards and recommended practices. These principles, standards and practices include insight into necessary documentation on both datasets and models, to support efforts to prevent sexual harms against children.

Request for Information

Thorn supports NIST's efforts to provide guidance on documentation of AI datasets and models for transparency between AI stakeholders. Due to the remit of our organization's mission, the

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

recommendations in this submission focuses on the role that this type of documentation can have to support efforts to prevent sexual harms against children. Below, we provide our input on the templates in clause 6.

6.2 Dataset Documentation Template

Identifier 4: Dataset Composition and Provenance

Regarding identifier 4, we recommend that descriptions around how and by whom the data has been acquired and processed should explicitly include efforts to ascertain or review the extent to which the data contains illegal material, including CSAM. This would bring the documentation template in line with existing global guidance⁵.

Identifier 5: Dataset Evaluation

Regarding identifier 5, we recommend that assessment around dataset minimization also be included in the evaluation process. Data minimization is a broadly agreed upon principle reflected across global regulation for ensuring privacy of individuals, including children. The presence of children's training data in generative models has serious implications⁶ with respect to potential downstream misuse of those models. Including this assessment also reflects the quality of "judiciousness" highlighted by NIST in this draft.

Identifier 6: Maintenance and Monitoring

Regarding identifier 6, we recommend that descriptions around how the dataset is updated, shared and maintained also include descriptions on how these decisions have been made in light of identifiers 4 and 5. For example, if datasets were not reviewed or filtered for CSAM, and yet were subsequently shared with third parties, this creates a risk to both the original data holder and those third parties. Decisions around datasets should reflect appropriate risk assessments that include the chain of decisions made previously.

6.3 AI Model Documentation Template

Identifier 6: Model Evaluation

Regarding identifier 6, we recommend that the descriptions include not just protocols and datasets used to assess the model, but also whether and how they engaged with third-parties as part of model assessment. While third-party assessment is not yet an industry norm, it nonetheless provides significant value⁷, as external testing discovers issues missed by internal teams. Transparency and visibility into which models have undergone this type of additional

⁵ This includes the Model Documentation Form for the Transparency Chapter of the [EU AI Act Code of Practice](#)

⁶ "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 adult sexual content and non-sexual depictions of children both are present in the training data, and the model demonstrates strong compositional generalization, this may indicate a corresponding high propensity for a model to be able to produce AIG-CSAM.

⁷ Longpre, et al. "In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI." ICML, 2025.

evaluation will be valuable to stakeholders making decisions on what models to use, in what contexts.

Identifier 8: Governance

Regarding identifier 8, we recommend that in addition to descriptions on frameworks and practices followed in model development, that the descriptions include as well what enforcement and auditing mechanisms were put in place to ensure that these frameworks and practices were consistently applied. This reflects the quality of “trust” highlighted by NIST in this draft, as it gives stakeholders visibility into not just what the ideal process is intended to be, but how it is executed in practice.

Conclusion

We wish to express our appreciation to NIST for their engagement and leadership in this space. If any questions arise from review of this request for information, please reach out to Dr. Rebecca Portnoff (rebecca@wearethorn.org) and Lara Gemar (lara.gemar@wearethorn.org).