PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA CASE STUDY

Mitigating the risk of generative AI models creating Child Sexual Abuse Materials

An analysis by child safety nonprofit Thorn

THORN

This is Thorn's Case Submission as a Supporter of PAI's Synthetic Media Framework.
Learn more about the Framework

1 Organizational Background

A contextual introduction to the case study.

Thorn is a nonprofit that builds technology 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 nonprofits with critical information to help them solve cases faster and remove children from harm. Thorn's research has provided the necessary insights and issue understanding to build robust interventions across the ecosystem.

Thorn is a civil society supporter of Partnership on AI's (PAI) Synthetic Media Framework and PAI's Risk Mitigation Strategies for the Open Foundation Model Value Chain. While we build machine learning (ML)/AI technology as part of our mission to defend children from sexual abuse, we focus our efforts on predictive, not generative AI. We do not directly create, host, or distribute synthetic media or its underlying technologies.

As a result, Thorn is submitting a case study on a third-party example. Thorn was not affiliated with the **Builders** (as defined in PAI's Framework) of the generative technology and infrastructure assessed in this case study. It also was not involved in the development, deployment, and distribution of the technology (Stable Diffusion 1.5), nor in the abuse material that was, and still is, generated using this model. This case study was selected because of the opportunity that remains with this particular model and its derivative models to prevent further misuse, abuse, and revictimization related to synthetic media and sexual harms against children.

2 Issue

Elaborate on the issue your organization addressed in the case study, i.e. the issue to which your organization decided to apply the Framework's principles.

Generative AI is being misused today to further sexual harm against children. This ongoing misuse has serious implications for child safety across several key verticals. While this misuse occurs across several data modalities, we focus on the image modality for this case study.

In brief, the impact of the misuse of generative AI on the issue of child sexual abuse (when focusing on the image modality) is three-fold:

1. **Impedes victim identification:** Bad actors primarily misuse broadly shared and open-source generative AI models to produce photorealistic AI-generated child sexual abuse material (AIG-CSAM) [1, 2, 3]. This content is increasingly difficult to visually distinguish from CSAM, which depict a child in an active abuse scenario, making victim identification work more difficult.

2. **Increases revictimization:** Bad actors use this technology to perpetrate

revictimization 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].

3. **Reduces barriers to harm:** Bad actors use this technology to sexualize benign imagery of a child and, in some cases, to scale their sexual extortion efforts (using generative AI to scale the content creation necessary to target a child) [4, 5]. Children also use this technology to bully and harass other children via the use of "nudifying" apps and services [6, 7, 8, 9, 10].

For this case study, we focus on the initial release of Stable Diffusion 1.5 (by well intentioned **Builders**) and the subsequent misuse of this same model (by malicious **Creators**) to produce AIG-CSAM. As noted in the impact overview above, there are also generative AI technologies developed by malicious **Builders** (e.g., models fine-tuned with CSAM and "nudifying" services). For this case study, we focus on what remains the most popular base model used by offenders dedicated to the sexual abuse of children — Stable Diffusion 1.5.

There are several dimensions we use to understand the impact of generative AI on the issue of child sexual abuse. For this report, we highlight three core elements:

1. **Prevalence of photorealistic AIG-CSAM (in relation to CSAM):** In [3], we reported our internal findings (as of June 2023) that within a sample of communities dedicated to child sexual abuse, less than one percent of CSAM files shared within that community are photorealistic AIG-CSAM. When focusing on just the AIG-SCAM images in the sample, we found that approximately 66 percent appeared photorealistic. An update to this analysis found that the prevalence of photorealistic AIG-CSAM in these communities has remained relatively stable, hovering around one percent; however, photorealism has increased. Compared to the original analysis, in which 66 percent appeared photorealistic, the updated analysis found 82.8 percent of sampled AIG-CSAM images now appear photorealistic. These findings on photorealism match those from the analysis conducted in [11] by the Internet Watch Foundation.

2. **Distribution of models used to generate AIG-CSAM:** As we previously reported in [3], Stable Diffusion 1.5 has been widely used for producing this content. Since the reporting, additional organizations have found similar trends within their investigations [12]. While additional models have been identified as having the potential to output CSAM, Stable Diffusion 1.5 is a uniquely high risk due to a few factors, including the presence of CSAM in its training data [13] and its training on sexually explicit images of adults alongside benign images of minors. Stable Diffusion 1.5 based models (that have been further trained on adult sexually explicit content) continue to be the most popular mechanism for generating AIG-CSAM in dark web communities dedicated to the sexual abuse of children. Such models are, in some cases, fine-tuned with CSAM by offenders who make use of Low-rank Adaptations (LoRAs) as a method to optimize model weights [3, 14].

3. **Scale of misuse of generative AI to sexualize benign imagery of children:** Reports of cases of this nature are increasing, reflecting the capability and will to produce explicit imagery from non-abuse material [11, 14, 15, 16]. The recent investigation of Justin Culmo shows these source images are being derived not only from online forums, such as social media profiles, but can be captured while victims go about daily business in the offline world [17]. As noted above, instances of misuse of generative AI to sexualize benign imagery of children has moved beyond adults looking to sexually abuse minors to also include minors generating images of their peers, in some cases resulting in arrests of minors [18].

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. (2023) Generative ML and CSAM: Implications and Mitigations. Stanford Digital Repository. Available at 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/2023/psa230605.

5 Thorn and National Center for Missing and Exploited Children (NCMEC). (2024). Trends in Financial Sextortion: An investigation of sextortion reports in NCMEC CyberTipline data. Available at https://info.thorn.org/hubfs/Research/Thorn_TrendsInFinancialSextortion_June2024.pdf.

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

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

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

9 Thorn. (2024). Youth Perspectives on Online Safety, 2023. Available at: https://info.thorn.org/hubfs/Research/Thorn_23_YouthMonitoring_Report.pdf.

10 "Florida Middle Schoolers Arrested for Allegedly Creating Deepfake Nudes of Classmates." Wired, March 2024, https://www.wired.com/story/florida-teens-arrested-deepfake-nudes-classmates/.

11 What Has Changed in the AI CSAM Landscape?. IWF, July. 2024, https://www.iwf.org.uk/media/drufozvi/iwf-ai-csam-report_update-public-jul24v12.pdf.

12 The dark reality of Stable Diffusion. CameraForensics, Aug. 2024, https://www.cameraforensics.com/blog/2024/02/08/the-dark-reality-of-stable-diffusion/.

13 Thiel, D. (2023) Identifying and Eliminating CSAM in Generative ML Training Data and Models. Stanford Digital Repository. Available at https://doi.org/10.25740/kh752sm9123.

14 IPPPRI Insights. (2024). Artificial intelligence-produced child sexual abuse material: Insights from Dark Web forum posts. Available at: https://www.cambridgenetwork.co.uk/sites/default/files/IPPPRI-Insight-No-1-AI-CSAM.pdf.

15 Office of Public Affairs | Man Arrested for Producing, Distributing, and Possessing AI-Generated Images of Minors Engaged in Sexually Explicit Conduct | United States Department of Justice. 20 May 2024, https://www.justice.gov/opa/pr/man-arrested-producing-distributing-and-possessing-ai-generated-images-minors-engaged.

16 Western District of North Carolina | Charlotte Child Psychiatrist Is Sentenced To 40 Years In Prison For Sexual Exploitation of A Minor And Using Artificial Intelligence To Create Child Pornography Images Of Minors | United States Department of Justice. 8 Nov. 2023, https://www.justice.gov/usao-wdnc/pr/charlotte-child-psychiatrist-sentenced-40-years-prison-sexual-exploitation-minor-and.

17 Brewster, Thomas. "A Pedophile Filmed Kids At Disney World To Make AI Child Abuse Images, Cops Say." Forbes, https://www.forbes.com/sites/thomasbrewster/2024/08/30/pedophile-filmed-kids-at-disney-world-to-make-ai-child-abuse-images-cops-say/.

18 Haskins, Caroline. "Florida Middle Schoolers Arrested for Allegedly Creating Deepfake Nudes of Classmates." Wired. www.wired.com, https://www.wired.com/story/florida-teens-arrested-deepfake-nudes-classmates/.

3 Objective

Describe what your organization is attempting to accomplish with its chosen strategy.

Thorn has conducted research, technology development, and advocacy in collaboration with the broader ecosystem — academics, nonprofits, policymakers, and industry — to prevent the misuse of generative AI from furthering sexual abuse. Our strategy to date has focused on aligning the ecosystem around principles and mitigations to: 1) reduce generative AI models' capability to produce AIG-CSAM and other abusive content, 2) ensure that the abusive content that is still produced is detected more reliably, and 3) minimize the spread of the underlying technologies used to produce the abusive content. Our primary vehicle for this effort has been Thorn and All Tech Is Human's initiative, Safety by Design for Generative AI: Preventing Child Sexual Abuse [1, 2]. At launch, ten companies (Amazon, Anthropic, Civitai, Google, Meta, Metaphysic, Microsoft, Mistral AI, OpenAI, and Stability AI) agreed to commit to these principles; since then, Invoke has also joined the commitments. Several companies also (Amazon AWS AI, Civitai, Hugging Face, Inflection, Metaphysic, Stability AI, and Teleperformance) co-authored a paper outlining principles, mitigations, and more for preventing generative AI misuse from furthering child sexual abuse [3].

With these commitments secured, we are currently pursuing accountability via three main avenues: 1) publishing reports with insights from the committed companies, sharing their progress in taking action on these principles and mitigations; 2) collaborating with standard-setting institutions such as IEEE and NIST to scale the reach of these principles and mitigations (opening the door for third-party auditing); and 3) engaging with policymakers to help them understand what is technically feasible and impactful in this space, according to the experts and sectors represented across our initiative.

All three themes in PAI's Framework (consent, disclosure, and transparency) are relevant to this issue space. For this case study on Stable Diffusion 1.5, we focus on consent. For readers interested in an analysis on the merits and failures of indirect disclosure in this issue space, feel free to reach out to PAI or Thorn for access to Thorn's analysis.

Regarding consent, it is obvious but nonetheless important to note that in this issue space, consent is not possible nor meaningful. Sexual activity between an adult and a child never constitutes consent, it is always abuse. The documentation of this abuse in the form of CSAM is likewise abuse and illegal. Furthermore, consent is irrelevant when it comes to AIG-CSAM, as (like CSAM) this media is illegal. This has serious implications for Stable Diffusion 1.5, which — however unintentionally — was trained with CSAM. We analyze the implications this underlying reality has for both the responsible building of generative AI models and infrastructure.

We seek to extend the assumptions of the Framework to explore guidelines for well-intentioned **Builders**, such that they are taking into account these critical circumstances where abuse has occurred or is occurring.

REFERENCES

1. "Thorn and All Tech Is Human Forge Generative AI Principles with AI Leaders to Enact Strong Child Safety Commitments." Thorn, 16 July 2024, https://www.thorn.org/blog/generative-ai-principles/.
2. "OpenAI, Meta and Google Sign On to New Child Exploitation Safety Measures." WSJ, April. 2024, https://www.wsj.com/tech/ai/ai-developers-agree-to-new-safety-measures-to-fight-child-exploitation-2a58129c?
3. Portnoff, et al. (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.

4 Framework Scope and Application

Identify which Framework principle(s) was used to help address the issue, how it was chosen, and describe how it was applied.

Consent

Stable Diffusion 1.5 was released in October 2022 [1]. The model was released by Runway ML [2], building off of prior development effort done in collaboration with the CompVis research group at the Ludwig Maximilian University of Munich [3]. The project was supported by Stability AI, which provided computational resources for the development of the model [4]. The training dataset was sourced from LAION-5B [5], a dataset established by a number of researchers from various institutions and curated by the non-profit LAION [6]. Hugging Face and Stability AI funded the work to establish the LAION-5B dataset [7]. In June of 2023, Thorn and SIO released research [8] that found Stable Diffusion 1.5 was being misused by bad actors dedicated to the sexual abuse of children to generate AIG-CSAM. In December of 2023, SIO further uncovered [9] that Stable Diffusion 1.5's training dataset contained CSAM. Following the notification of that finding, LAION announced their decision to temporarily take down the dataset [10], "in an abundance of caution, we are temporarily taking down the LAION datasets to ensure they are safe before republishing them."[1] When inspecting LAION's datasets hosted in Hugging Face [11], no datasets referencing the original LAION-5B collection seemed present.[2]

In May 2024, Thorn was contacted by representatives from CITED (the California Initiative for Technology and Democracy) to discuss potential collaborations. At the time, David Evan Harris and David Willner intended to co-author an essay focused on the misuse of Stable Diffusion 1.5 to enable AIG-CSAM and adult synthetic non-consensual intimate imagery (NCII), with a specific focus on the role model hosting platforms play in the distribution of this model. Thorn advised this effort for several months, providing details into our Safety by Design initiative and strategy for accountability, as well as highlighting other efforts across

1. On August 30, 2024, LAION announced that they have re-uploaded a version of LAION-5B with "additional safety fixes." For an analysis of these fixes, see David Thiel's thread on the topic.

2. When searching Hugging Face's Datasets more generally, one resource that appeared to be the LAION-5B dataset was still available. Accessing this dataset required being logged in as a user to Hugging Face, and confirming your email address. Upon notifying Hugging Face of the presence of this dataset, they filed a report on the repository, offering guidance to the user on next steps to stay in compliance with Hugging Face's content policy. After following Hugging Face's moderation process, access to this dataset was disabled.

the child safety ecosystem to combat AIG-CSAM and other generative AI-enabled sexual harms against children. In August 2024, with PAI's permission, we provided Harris and Willner with a preprint of this report to give them a complete picture of the different organizations (Runway ML, the CompVis research group at the Ludwig Maximilian University of Munich, Stability AI, LAION, and Hugging Face) involved in the development of Stable Diffusion 1.5 and the LAION-5B dataset.

In September 2024, Harris announced [12] that after personal outreach to some of the organizations listed above, Stable Diffusion 1.5 was taken down from Hugging Face. In their subsequent essay [13], Harris and Willner shared that a representative of Hugging Face indicated that Hugging Face did not themselves take the model down. It seems likely that Runway ML chose to take the model down as a result, but this has not been confirmed.

No similar action to remove Stable Diffusion 1.5 and make it unavailable for download appears to have been taken by other platforms that hosted third-party models (e.g., Github and Civitai), by the platforms themselves or third-party developers who uploaded the model.

While PAI's Risk Mitigation Strategies for the Open Foundation Model Value Chain offers possible avenues to ensure model safety for actors within the foundation model ecosystem, PAI's Framework does not contain any guidance for **Builders of Technology and Infrastructure** regarding consent generally or specifically around training datasets. There is an ongoing crucial public conversation on consent expectations for training datasets used to build AI models [14, 15]. For this case study, that conversation is out of scope — although we note again that guidance on this point (e.g., guidance found in the ethics guidelines [16] for NeurIPS, a leading computer vision conference) is missing from the current Framework.

As we stated previously, sexual activity between an adult and a child never constitutes consent, it is always abuse. CSAM is likewise abuse and is illegal. Responsible **Builders** must take this into account. These images depict some of the worst moments of a child's life. Detecting, removing, and reporting CSAM from training datasets to avoid furthering the revictimization of these children and the co-opting of these moments to generate (related or unrelated) synthetic content is an ethical imperative.

There are currently two known avenues in which a general-purpose model is capable of producing AIG-CSAM. The model may have sufficient CSAM in its training data such that it has learned this concept and can reproduce the material. The model may also exhibit compositional generalization capabilities, having been trained on a combination of adult sexual content and benign depictions of children, allowing it to combine these two concepts to produce AIG-CSAM. Detecting, removing, and reporting CSAM from training datasets is a necessary (but certainly not sufficient on its own) mitigation in the former scenario to reduce the model's capabilities to produce AIG-CSAM. Similarly, refraining from co-training models on both benign depictions of children and adult sexual content can mitigate the risk of compositional generalization noted above. These interventions are important in supporting the efforts of law enforcement (LE) and other victim identification specialists.

From both victim identification and revictimization perspectives, we recommend incorporating guidance into PAI's Framework on synthetic media for **Builders**, so they make the best efforts to detect, remove, and report CSAM from training datasets when developing datasets and training models.

We note above that Stable Diffusion 1.5 is still available for download on platforms that fit the PAI Framework's definition of "generative AI infrastructure." It is clear that a model's weights are concretely shaped by the training data used to build that model's weights. Recent research indicates that diffusion models in particular can memorize individual images from their training data [17]. There is ongoing debate on whether this type of memorization indicates that the model, or the model weights, can be understood as a "copy" of the model's training data [18]. Regardless of one's stance on this, it is clear that a model trained on data that includes CSAM will have a different set of weights than a model trained on the equivalent dataset with the CSAM removed. Synthetic content produced using a model trained on CSAM, regardless of whether the synthetic content is abuse material, has also been shaped accordingly.

Furthermore, Stable Diffusion 1.5 is definitively capable of producing AIG-CSAM. Opinions vary on whether tools should have safeguards built into them to avoid their downstream misuse. Yet, at this moment, there is consensus across several leading technology companies and child safety advocates that, at a minimum, models capable of producing AIG-CSAM should be temporarily removed from access until they are updated with mitigations [19]. Stable Diffusion 1.5 fits this definition.

Again, from both victim identification and revictimization perspectives, we recommend incorporating guidance into PAI's Framework on synthetic media for **Builders** to ensure that, when developing infrastructure for synthetic media, they do their best to remove models trained with CSAM — such as Stable Diffusion 1.5 and its derivatives — and only replace the models once they have been rebuilt without the CSAM in their training data.

Finally, it is worth noting that these guidelines will be most impactful if adopted not just by large industry players, but also by hobbyists/individuals and groups with fewer resources. As noted in the issue overview, Stable Diffusion 1.5-based models that have been further trained on adult sexually explicit content continue to be the most popular mechanism for generating AIG-CSAM in communities dedicated to abuse. These derivative models are generally not developed by corporations, but rather by hobbyists/individuals and informal groups. Incentives for adopting these types of guidelines include avoiding reputational risk and legal liability. Currently, there is no legal liability for failing to comply with these types of guidelines and reputational risk may be more significant for corporations, which can be impacted by negative press surrounding the misuse of their technology for harming children.

While it is difficult to predict "what would have happened," it seems likely — based on current usage patterns — that even if the original **Builders** of Stable Diffusion 1.5 followed data curation, cleaning, and other mitigation best practices to develop and release their

model, the impact of those efforts would have been limited unless the hobbyists/individuals who further trained and fine-tuned the model on additional data followed the same best practices. This is not to say that those interventions by the original **Builders** of Stable Diffusion 1.5 are any less a critical or foundational component of responsible development. As discussed above, from both victim identification and revictimization perspectives, those interventions are crucial — but not sufficient on their own. This is more true for combatting child sexual abuse, which requires a whole-of-society approach for maximum impact.

We recognize that the resources, incentives, and oversight for complying with such best practices can vary quite widely between hobbyists/individuals and corporations. Nonetheless, a culture of responsibility across **Builders** in various settings will be necessary to have a fuller impact.

REFERENCES

1 Runwayml/Stable-Diffusion-v1-5, Hugging Face. https://huggingface.co/runwayml/stable-diffusion-v1-5
2 Runway | Tools for Human Imagination. https://runwayml.com/.
3 "High-Resolution Image Synthesis with Latent Diffusion Models." Computer Vision & Learning Group, https://ommer-lab.com/research/latent-diffusion-models/.
4 Runwayml/Stable-diffusion, Github. https://github.com/runwayml/stable-diffusion
5 https://openreview.net/forum?id=M3Y74vmsMcY
6 https://laion.ai/
7 https://openreview.net/attachment?id=M3Y74vmsMcY&name=supplementary_material
8 Thiel, D., Stroebel, M., and Portnoff, R. (2023) Generative ML and CSAM: Implications and Mitigations. Stanford Digital Repository. Available at https://doi.org/10.25740/jv206yg3793.
9 Thiel, D. (2023). Identifying and Eliminating CSAM in Generative ML Training Data and Models. Stanford Digital Repository. Available at https://doi.org/10.25740/kh752sm9123.
10 https://laion.ai/notes/laion-maintenance/
11 https://www.linkedin.com/posts/davidevanharris_ai-csam-childsafety-activity-7235030045528891392-FwWE/?
12 Harris, David and Willner, David. "AI Image Generators Make Child Sexual Abuse Material (CSAM)." IEEE Spectrum. https://spectrum.ieee.org/stable-diffusion.
13 https://huggingface.co/laion
14 Gilbertson, Annie. "Apple, Nvidia, Anthropic Used Thousands of Swiped YouTube Videos to Train AI." Wired. https://www.wired.com/story/youtube-training-data-apple-nvidia-anthropic
15 "AI Is Being Trained on Images of Real Kids Without Consent." Futurism, 12 June 2024, https://futurism.com/ai-trained-images-kids.
16 https://neurips.cc/Conferences/2023/EthicsGuidelines
17 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.
18 Cooper and Grimmelman. "The Files are in the Computer: On Copyright, Memorization, and Generative AI." Chicago-Kent Law Review. https://james.grimmelmann.net/files/articles/the-files-are-in-the-computer.pdf
19 Portnoff, et al. (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.

5 Approaches to Direct Disclosure, in Policy and Practice

Elaborate on any internal or external obstacles intrinsic to the Framework or the role your organization played in addressing the issue that were overcome.

The efficacy of PAI's Framework is heavily dependent on interest and commitment to combat the abuse of generative AI technologies and unintended harms resulting from their abuse. Those abusing and exploiting minors via this technology have no such interest and, in fact, will actively attempt to circumvent such safeguards in service of their abuse.

As a result, it is paramount that those in positions to reduce abuse of the technology are made fully aware of how that abuse may show up and what specific actions they can take to mitigate these risks. **Builders** — not just in a corporate setting, but hobbyists/individuals as well — will play a prominent role in addressing this risk.

We have highlighted three interventions for **Builders** in this report: removing generative AI models trained on CSAM from their platforms, removing generative AI models that can produce AIG-CSAM from their platforms; and detecting, reporting, and removing CSAM from the training datasets of their generative AI models.

The primary obstacles we have heard from **Builders** in a corporate setting that prevent them from engaging in these interventions include the lack of awareness about which models fit these categories and lack of access to technology capable of detecting CSAM at scale. It is worth noting that some of this information is discoverable and we include several details and options for these interventions in [1].

The other obstacle worth noting is the lack of financial incentive. Hosting-platform businesses predicated on providing access to models and creating generative AI models have a financial incentive to continue doing both of these. There's also a cost associated with putting in the effort to follow these interventions. As noted earlier in this report, this can be balanced by reputational risk and legal liability.

We have not engaged directly with **Builders** not associated with a corporation (such as the hobbyists/individuals) and, therefore, have not heard their perspectives on what prevents them from engaging in these interventions. Based on anecdotal observation, the same obstacles described above also seem to apply here. Furthermore, there appears to be a lack of consensus in these communities on the harm that comes from AIG-CSAM, which we believe is likely one significant blocker in the broader adoption of these interventions by hobbyists/individuals. Another blocker is general skepticism about the effectiveness of these types of interventions.

REFERENCES

1 Portnoff, et al. (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.

6 Benefits

Identify the ways in which the Framework can support decreased harm and increased benefits from synthetic media.

It is worth reinforcing that voluntary commitments and guidelines will not be sufficient to bring everyone to the table. Some organizations will not voluntarily commit, others will commit but will not follow through on their commitments, and others will commit and follow through but not at a sufficient pace to have maximal impact. The same is true for **Builders** who are hobbyists/individuals. As indicated earlier in this report, accountability will require legislation. Policy interventions that lean into safety by design (to prevent harm) and recognize that these generative AI models have already been built, deployed, and are causing harm will both be necessary. PAI's Framework offers clear and accessible guidance addressing many opportunities across the generative AI ecosystem to deliver safe and responsible generative AI media. It increases visibility of potential misuses and unintended harms, and can be paired with other frameworks more focused on preventing intentional misuse (e.g., Thorn's Safety by Design work), for positive impact. Together, these efforts can serve as the foundation for potential future policy interventions.

In addition to voluntary commitments and legislation, we believe the role of education and awareness is critical, particularly to reach those **Builders** who are hobbyists/individuals. As noted earlier, there is not yet consensus in those communities on either the harm that manifests from this type of abuse material or the efficacy of proposed strategies to mitigate that harm. Finding the right mechanisms for providing that education and building that awareness are necessary first steps. Furthermore, in relation to the efficacy of particular interventions and mitigations, we believe reporting concrete metrics (e.g., around prevalence of this type of abuse material before and after mitigations have been put in place) will be another important aspect to building alignment and understanding within these communities.

Finally, **Builders** may need access to tools (e.g., Microsoft's PhotoDNA service) to implement some of these mitigations and interventions. For corporations, getting access to these tools may be less of a challenge than for hobbyists/individuals. One way to navigate this reality could be for larger institutions to collaborate on releasing curated datasets (e.g., hashed/matched against lists of known CSAM and NCII, filtered using CSAM classifiers, etc.) These datasets should be further audited by third-party institutions before release, which is a critical step based on the LAION dataset's history. Once released, they could open the door for hobbyists/individuals to follow these mitigations and interventions more easily.

7 Conclusion/Key Takeaways

Identify the ways in which the Framework can support decreased harm and increased benefits from synthetic media.

Online child sexual exploitation and abuse involve highly driven offenders, often technically skilled and interested in novel ways to produce imagery, groom victims, and avoid detection. As a result, additional efforts are needed to not just develop a framework for those committed to the safe and responsible development and deployment of generative AI technologies, but also implement protections against the adversarial actions of bad actors.

To this end, **Builders** — not just in a corporate setting, but hobbyists/individuals as well — will play a significant role in ensuring that available tools can stand up to adversarial attacks. Continuing to raise awareness about the risks of generative AI misuse, how it may manifest, and how it can be mitigated, will serve as a foundation to collaboration among the ecosystem to deliver effective and relevant solutions. Implementing and enforcing these solutions will require both voluntary commitments and legislation.