PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA CASE STUDY

# Direct disclosure has limited impact on AI-generated Child Sexual Abuse Material

An analysis by researchers at Stanford HAI

This is a case submission by researchers Riana Pfefferkorn and Caroline Meinhardt of Stanford HAI as a Supporter of PAI's Synthetic Media Framework.
Learn more about the Framework

# 1 Organizational Background

1. Provide some background on your organization.

The Stanford Institute for Human-Centered Artificial Intelligence (HAI) is a nonpartisan academic research institute that aims to advance AI research, education, policy, and practice to improve the human condition. Founded in 2019, HAI plays a leading role not only in producing critical scholarship on AI governance, but also in engaging, educating, and convening government officials, civil society, and industry organizations around cutting-edge AI policy issues.

As an interdisciplinary institute focused on producing evidence-based research, Stanford HAI represents a diverse range of voices and perspectives. The views expressed in this case study reflect the perspectives of the authors and do not represent the official position of Stanford HAI. Since Stanford HAI itself does not directly create, host, or distribute synthetic media or its underlying technologies, we submit this case study to contribute a third-party analysis of some of the urgent challenges posed by synthetic media. By analyzing how direct disclosure practices may be effective — or ineffective — in real-life examples of AI-generated child sexual abuse material (AIG-CSAM), we hope to demonstrate the complexities of applying PAI's Synthetic Media Framework in certain cases of synthetic media misuse.

# 2 Framing Direct Disclosure at your Organization

1. Please elaborate on how your organization recommends providing direct disclosure (as defined in our Glossary for Synthetic Media Transparency Methods) to users/ audiences.

Stanford HAI does not directly provide direct disclosure or have an institutional position on best practices for it. However, the authors recommend that in the context of AIG-CSAM, images should always be labeled directly and prominently if they are AI-generated or -modified, using means (to the degree technically possible) that are difficult to remove from the image and to insert falsely into a real image.

CSAM, by its nature, is circulated and consumed by bad actors, who are generally not incentivized to apply direct disclosure techniques. If labels or watermarks are present, the bad actors may have more incentive to make alterations compared to synthetic media stakeholders.

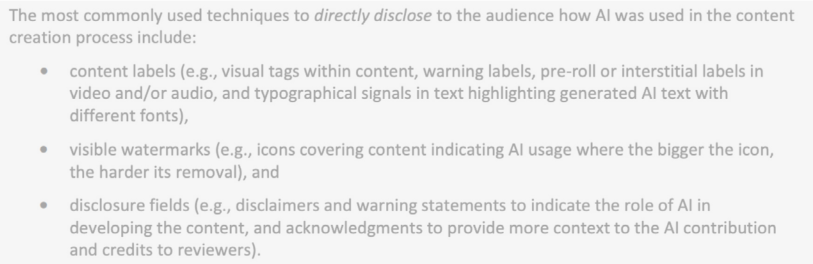
2. Does your organization understand the goal of direct disclosure as specified in the PAI Framework: “to mitigate speculation about content, support resilience to manipulation or forgery, be accurately applied, and communicate uncertainty without furthering speculation” or does it have a different understanding?

Stanford HAI does not have an organizational viewpoint on this. While the authors generally agree with the goals of direct disclosure outlined in the PAI Framework, we believe that it is important to consider a wider range of goals that take into account the different audiences of direct disclosure.

Many primarily think of the consumers of synthetic content (e.g., the users of social media platforms or messaging services) as the audience for labels or watermarks, in which case raising awareness and educating users are key goals. However, direct disclosure mechanisms can also have a variety of other audiences. For example, various stakeholders who monitor, analyze, and work to mitigate the risks of synthetic content (including Trust and Safety teams at platforms, independent researchers, civil society groups, or even law enforcement) can also benefit from direct disclosure. For these audiences, the goal is not merely to inform, but also to help streamline the detection, review, and — when needed — the reporting of synthetic content in order to efficiently manage resources. Direct disclosure may even play a role in determining the legality of certain types of synthetic content, though labels and watermarks cannot convey legality definitively. (We elaborate on this in the context of AIG-CSAM in Sections 4.2. and 4.6.) These other audiences illuminate how the Framework, as a voluntary self-regulatory set of practices, can intersect with other regulatory mechanisms such as criminal laws and platforms’ terms of service.

3. What, if anything, does your organization believe is missing from this NIST taxonomy? Should it be added to a taxonomy of direct disclosure? If so, why?

*From NIST’s Reducing Risks Posed by Synthetic Content:*



The most commonly used techniques to *directly disclose* to the audience how AI was used in the content creation process include:

- content labels (e.g., visual tags within content, warning labels, pre-roll or interstitial labels in video and/or audio, and typographical signals in text highlighting generated AI text with different fonts),
- visible watermarks (e.g., icons covering content indicating AI usage where the bigger the icon, the harder its removal), and
- disclosure fields (e.g., disclaimers and warning statements to indicate the role of AI in developing the content, and acknowledgments to provide more context to the AI contribution and credits to reviewers).

Stanford HAI does not have an institutional position on the NIST taxonomy. The individual authors find the methods described in this taxonomy to be necessary but not sufficient to mitigate the risks of AIG-CSAM. As noted throughout this document, e.g., in Sections 3.3 and 3.4, the harms of AIG-CSAM are inherent to its very creation. Due to this, labels, watermarks, and disclaimers can reduce only some risks; fully negating the risks requires not creating the content at all or even refraining from assembling a training data set that could support AIG-CSAM generation in the first place.

4. What criteria should [Builders, Creators, and/or Distributors] of synthetic media use to determine whether content should be Directly Disclosed?

Since there are no legitimate use cases for AIG-CSAM, good-faith **Builders** should take measures to prevent their models from being used to generate AIG-CSAM at all. As said, however, AIG-CSAM should always be directly disclosed, as disclosure is helpful for any victims depicted as well as other good-faith stakeholders (such as Trust & Safety teams at platforms). **Builders** attempting to prevent misuse of their models for AIG-CSAM should consider building direct (and indirect) disclosure into images generated with their model, in case their preventive efforts fail — at minimum, into any images involving children, even where the prompt is not sexually explicit (since models may generate CSAM even when they are not asked to). For more, see Sections 3.4, 4.2, and 4.4.

5. Per the Framework, PAI recommends disclosing "visual, auditory, or multimodal content that has been generated or modified (commonly via artificial intelligence). Such outputs are often highly realistic, would not be identifiable as synthetic to the average person, and may simulate artifacts, persons, or events." How does your organization's thinking align with, or diverge from, this recommendation?

While Stanford HAI does not have an institutional position on this recommendation, it aligns with the individual authors' thinking in the specific context of CSAM that is AI-generated or -modified, which we believe should always be disclosed for reasons discussed throughout Sections 3 and 4.

## 3 Real World, Complex Direct Disclosure Example

1. Provide a real-world example in which either: a) direct disclosure should have been applied, or b) direct disclosure was applied to a piece, or category, of content for which it was challenging to evaluate whether it warranted a disclosure. This could be because the threshold for disclosing was uncertain, the impact of such content was debatable, understanding of how it was manipulated was unclear, etc. Be sure to explain *why* it is challenging.

In recent years, many different AI tools have emerged that enable the creation of sexually explicit synthetic content, including explicit content of minors. These tools generally fall into two categories:

First, there are AI-powered apps or sites that allow users to create highly realistic nude images of real individuals that are based on real images of those same clothed individuals. These tools either "undress" victims by replacing the pixels that represent clothing in the original picture with an AI-generated nude body or "face swap" the image of a victim's face onto an image of another real person's nude body.

Second, there are AI models and associated tools that allow the creation of entirely new explicit images. These include predominantly open-source general-purpose image generation models such as Stable Diffusion, which were not designed to generate explicit content but can be prompted to produce such content, as well as models fine-tuned with the specific purpose of creating explicit content.

Below, we present three real-world examples of AIG-CSAM to highlight the different contexts in which CSAM is generated. They present different sets of motivations and intended audiences that are crucial to understanding the incentives for and implications of direct disclosure (or the lack thereof). We do not name the specific tools that were used in these examples to avoid inadvertently contributing to increasing their prominence. We also narrowly focus on synthetic images, even though synthetic video content is increasingly emerging.

**1. Teen "Deepnudes": AIG-CSAM created by minors for minors**

There has been a vast uptick in minors using AI tools to create and then disseminate nonconsensual AI-generated sexual images of other minors (primarily girls). In February 2024, a group of Beverly Hills eighth-graders created and shared AI-generated nude pictures of 16 classmates. The images were briefly circulated among the middle school's student body before being reported to school officials, who contained their spread and later expelled five students they found to be "most egregiously involved" in creating and sharing the images.

The school did not specify exactly how the images were made. However, an announcement by the Beverly Hills Unified School District indicates that the creators used an app that "undresses" victims. According to media reports, the images were shared through messaging apps, though it is unclear which platforms were used.

There is no indication that the creators of the images directly disclosed that they were generated using AI tools. Given that their motives will likely have been to humiliate, bully, or harass their victims, they were not incentivized to disclose that the images were not real. On the contrary, their goal was likely to generate images that can't readily be distinguished from real images. Since the app was not named, we do not know whether that app implements direct (or indirect) disclosure mechanisms.

**2. Supercharging grooming: AIG-CSAM created by adults for minors**

Generative AI tools are also already being adopted by adults to produce highly realistic CSAM. Among other things, AIG-CSAM is being used by bad actors to scale their grooming and sexual extortion efforts. In May 2024, a man was arrested in Wisconsin for allegedly producing, distributing, and possessing thousands of realistic AI-generated images of nude or partially clothed minors engaged in sexually explicit conduct. Law enforcement recovered the evidence from his electronic devices after the National Center for Missing & Exploited Children (NCMEC) flagged messages containing the images, which were allegedly sent in October 2023.

The defendant generated the images by feeding specific, sexually explicit text prompts related to minors into Stable Diffusion, a text-to-image generative AI model created by Stability AI. He used "negative prompts" (i.e., prompts that specified what he did not want to see in the generated images) and add-ons to ensure the resulting images would not depict adults and create realistic images of genitalia. Crucially, the defendant had sent a 15-year-old boy descriptions of how to use Stable Diffusion to convert text prompts into sexually explicit images of minors and sent several such images to the boy using Instagram's direct message. According to information law enforcement obtained from Instagram, he also indicated that he widely shared AIG-CSAM via Telegram.

In this context, the images' creator verbally disclosed that the images were generated using AI tools when communicating with the minor. According to prosecutors, the goal of the defendant, who had acknowledged having a sexual interest in children, was to sexually entice the minor. In addition, the images may have contained invisible watermarks, which Stable Diffusion embeds in graphical content image files by default. However, removing the relevant code for the watermark is easy and common in the AIG-CSAM community. It's therefore conceivable that the defendant, a software engineer, had the technical skills to remove the watermarks, though it's unclear whether he did so.

**3. Scaling child sexual exploitation: AIG-CSAM created by adults for adults**

As the above example indicates, AIG-CSAM is also being created for and shared with adults. The Wisconsin defendant's creation of thousands of AIG-CSAM images, which he advertised on Instagram and Telegram, indicates that generative AI enables the scaling of CSAM content creation.

In another notable real-life example, a man was sentenced to 40 years in prison in North Carolina in November 2023 for producing, transporting, and possessing CSAM, including AIG-CSAM that was based on real minors. The perpetrator used a web-based AI application to alter real images of clothed minors he had obtained from a website into highly realistic, nude images. While some of the images were recent, others were photographs taken several decades ago of minors at the time who are now adults.

There is no indication that the creator of the images distributed them via any platforms or messaging services. Similarly, there is no indication that he directly disclosed that they were generated using AI tools. In this instance, the perpetrator may have generated the images only for his own sexual gratification and may have never intended for others to see the images. As such, he was not incentivized to disclose that the images were not real. There is also no indication that the AI image generation tool he used implements direct disclosure mechanisms.

2. How was this piece/kind of content identified?

In most cases, AIG-CSAM that depicts real people is identified as synthetic content by the victims themselves (or people who know the victims) once the content has already been circulated, as well as by law enforcement or other institutions of authority. In the first example presented above, the content was initially identified by students at the school and their parents who recognized the victims and brought the content to the attention of school officials. In the third example, the content was discovered by the perpetrator's wife, after which the Federal Bureau of Investigation uncovered the victims' identities.

In rare cases, AIG-CSAM is declared as synthetic content by the perpetrators themselves. In the second example described above, the creator of the content clearly communicated that it was AI-generated.

3. Was there any potential for reputational (e.g., negative impact on the organization's brand, products, etc.), societal (e.g., negative impact on the economy, etc.), or any other kind of harm from such content?

The creation and distribution of AIG-CSAM causes grave harm to victims. Victims who are affected by nonconsensual, sexualized AI-generated images based on real images of their faces and/or bodies suffer substantial emotional, psychological, and reputational harm due to their privacy, agency, and dignity being violated. Victims have described how they have suffered "substantial emotional distress, mental anguish, anxiety, embarrassment, shame, [and] humiliation," and how such sexual exploitation of images elicits "nausea, fear, and overwhelming discomfort and distrust."

Additionally, victims can suffer from the long-term impact of the content, which may continue to be circulated indefinitely. Victims have to live with the never-ending threat of their continued exploitation. Additionally, as demonstrated in our third example, victims may suffer from sexual exploitation through AI-generated images decades after the original photos are taken.

Crucially, there are also a variety of upstream and downstream harms. Upstream, the very process of training models to produce AIG-CSAM typically involves using photographic abuse imagery as training data, which revictimizes the children depicted. Even the use of non-sexual images of children for training purposes can lead to various privacy harms, as those images are used without consent to train models to create AIG-CSAM. Downstream, AIG-CSAM — which can be created easily at scale — puts a vast additional strain on law enforcement resources, which are already overburdened [PDF]. As platforms flood with AIG-CSAM, it becomes even more challenging for law enforcement and other investigative teams at NCMEC or messaging platforms to identify real victims who need help.

4. What was the impact of implementing, or not implementing, this direct disclosure? How would your organization assess such impact (studying users, via the press, other civil society, community reactions, etc.)? Did the disclosure mechanism mitigate the harm described in the previous question (3.3)?

Stanford HAI does not have an organizational viewpoint on these questions; this is the viewpoint of the authors.

In the case of AIG-CSAM, implementing direct disclosure does not **prevent** harm. AIG-CSAM is not like most types of AI-generated content because its harms are inherent, not contingent on downstream beliefs, behaviors, decisions, or uses that flow from disclosure or nondisclosure of the image's AI-generated origin. The creation, existence, and distribution of the content all cause substantial harm.

Direct disclosure may have some effect on the extent of the harm to victims. For example, the reputational harm may be less severe if people know that a nude or otherwise sexualized image of a minor is not real. Yet even when this is known, the victims will suffer reputational harm, including being blamed for their own victimization if they had posted clothed photos of themselves online (as hundreds of millions of people have been doing for decades).

Particular direct disclosure methods such as visible watermarks may also help mitigate demand for AIG-CSAM if they are applied strategically to impede the consumption of the image, e.g., by covering the child's face and/or body with a watermark in a way that is prominent and hard to remove. However, **Creators** of AIG-CSAM and **Builders** of CSAM-specific AI models have little incentive to apply watermarks to their images, the very creation of which may be illegal in the first place. Measuring the impact of watermarking or other disclosure methods on demand is difficult due to research ethics. However, organizations such as Thorn or the UK's Internet Watch Foundation, which monitor online AIG-CSAM trading forums, may be able to research forum users' sentiment toward disclosed images.

Direct disclosure may have a more substantial impact on mitigating the downstream harm of AIG-CSAM. Knowing whether CSAM is AI-generated or not allows NCMEC, platform Trust and Safety teams, and law enforcement to prioritize their resources appropriately. We expand on the impact of direct disclosure on harms in Sections 4.2 and 5.3 below.

5. Is there anything your organization believes either the Builder, Creator, or Distributor of the content should have done differently to support direct disclosure?

Stanford HAI does not have an institutional position on this question, but the individual authors consider AIG-CSAM a unique type of synthetic content where direct disclosure has limited impact. Since there are no legitimate use cases for CSAM, direct disclosure that CSAM is AI-generated or -modified serves few legitimate goals. As discussed above, the **Creators** and **Distributors** of AIG-CSAM are rarely incentivized to disclose that their content is AI-generated. Even in the rare instances when direct disclosure is implemented, this practice doesn't mitigate much of the harm.

Instead, the content creation itself must become the point of intervention. The **Builders** of models that could allow the creation of various types of AIG-CSAM must take measures to ensure that their models cannot be misused for these purposes.

One key intervention point is the training data. Ensuring that AI training datasets don't contain known CSAM (for example, by detecting and removing image links that match with hashed images in online safety organizations' databases) is the absolute minimum. Beyond that, **Builders** may need to ask themselves whether non-explicit images of children should even be included in training datasets (see Section 4.4 for a more detailed discussion of this question).

Other interventions apply at the model development stage: **Builders** should deploy classifiers to detect, flag, and ultimately prevent the creation of imagery of children that may be sexually explicit. In addition, rigorous internal and third-party red teaming is necessary to test for any accidental creation of sexualized imagery of minors, given image generation models' propensity for producing sexualized images even when not prompted. Such testing may implicate legal concerns, which should be addressed before commencing testing.

Finally, platforms that host models should consider requiring **Builders** to take certain AIG-CSAM prevention measures as a prerequisite for model hosting and distribution. Detailed best practices for **Builders** and **Distributors** are beyond the scope of this case study, but recommendations are available from the child safety tooling organization Thorn (and in their complementary case study as a PAI Framework supporter).

6. In retrospect, what, if anything, does your organization believe should have been done differently by the stakeholders identified in the previous question?

Please see Section 3.5 for a deeper dive into this topic.

7. Were there any other policy instruments that should have been relied upon in deciding whether to, and how, to disclose the content? What external policies may have been helpful to supplement internal policies?

**Builders** need to anticipate that their AIG-CSAM prevention efforts may fail or be circumvented. As mentioned above, **Creators** of AIG-CSAM may make efforts to remove watermarks or other direct disclosure techniques, especially if they are highly intrusive. Conversely, bad actors may exploit direct disclosure for their own gains. For example, they may apply fake labels or watermarks to actual photographic CSAM, passing off real imagery as AI-generated. Viewers mistaking fake images for real is one risk of synthetic content, but mistaking real images for fake is also a major concern that false disclosures can exacerbate. That is particularly true in the unique context of CSAM, where real images are highly illegal worldwide, thereby providing an external policy basis for guiding internal disclosure policies. Mislabeling real imagery as “AI-generated” can harm the children depicted, for example, by inducing platforms, NCMEC, or law enforcement to deprioritize removal, reporting, and investigation. This could also delay or prevent identification and rescue efforts that would have been promptly undertaken had the image accurately been understood as real from the start.

To address the concerns of real imagery being passed off as AI (and vice versa), direct disclosure should be accompanied by indirect disclosure, both of which should be tamper-proof or tamper-evident to the degree feasible so that information cannot be inadvertently or intentionally altered or removed. Ideally, there should be a simple way for a viewer of an image to, for example, check file metadata to verify that an image disclosed as AI is what it claims to be. This may require additional technical investment.

8. What might industry practitioners or policymakers learn from this example? How might this case inform best practices for direct disclosure across those Building, Distributing, and/or Creating synthetic media?

We talk more about this in Section 3.5 and in Section 4 below.

# 4 How Organizations Understand Direct Disclosure

1. What research and/or analysis has contributed to your organization's understanding of direct disclosure (both internal and external)?

To date, the authors' understanding has been largely informed by Stanford scholars' prior research into the CyberTipline reporting system, the legality of AIG-CSAM, the misuse of diffusion models to create CSAM, and CSAM in the LAION-5B dataset. Our understanding is also informed by a review of ongoing reporting and court documentation on known, real-life cases of AIG-CSAM creation and distribution, as well as resulting harms and criminal prosecution and sentencing. We also consider news coverage and policy analysis of the related issue of AIG image-based sex abuse of adults (aka nonconsensual intimate imagery or NCII).

2. Does your organization believe there are any risks associated with either OVER or UNDER disclosing synthetic media to audiences? How does your organization recommend navigating these tensions?

Stanford HAI does not have an institutional position on this question, but the authors find that a threshold question is: "Disclosure for whom?" There is a distinction between disclosure to "consumers" of AIG-CSAM (which may include adults seeking out such material, children being groomed by adults, and children viewing a "nudified" image one child made of another child) vs. Trust & Safety teams at online platforms, vs. the justice system (law enforcement, prosecutors, judges and juries, plaintiff's counsel).

Overdisclosure has little applicability in the context of AIG-CSAM. Disclosure that an item of CSAM is AI-generated likely doesn't prevent harm (as recognized by one deepfake and NCII bill that forecloses the use of disclaimers as a defense to liability). But disclosure also doesn't create additional harm. In some online forums, disclosure may be unnecessary because it is widely understood that any image shared on that forum is AI-generated; in that context, any disclosure might be viewed as superfluous. However, if the image spreads beyond that forum (as seems likely), that context and shared understanding would get lost, so disclosure is not superfluous.

With underdisclosure, the primary risk is the burden on investigations and harm to triaging efforts. With the CyberTipline receiving tens of millions [PDF] of reports of suspected CSAM per year, disclosure that an image is AI-generated is useful for platform Trust & Safety teams (who can flag the material as AI-generated when reporting it to the CyberTipline), NCMEC (which processes those reports and distributes them to law enforcement), and law enforcement (which has to triage incoming reports). By contrast, underdisclosure risks wasting resources by sending NCMEC and investigators on a wild goose chase to identify and locate a child who doesn't exist.

Underdisclosure also poses additional risks to children victimized by "nudified" images. These images harm the victim whether or not they are disclosed as AI-generated. However, there can be further harm if an image is believed to be real (e.g., by peers or adults with disciplinary power over the child). (N.B. all of this is also true where the "nudify" victim is an adult.) In the grooming context, by contrast, over/underdisclosure is less pertinent, as CSAM can be used to groom a child, whether the content is real or AI-generated.

3. What conditions or evidence would prompt your organization to re-calibrate your answer to the previous question (4.2)? E.g., in an election year with high stakes events, your organization may recommend over labeling.

The authors believe there are no conditions under which AIG-CSAM's existence is legitimate. A common question is: "Couldn't AIG-CSAM serve as a harm reduction measure for people who would otherwise consume real abuse imagery?" There is no evidence to support this thesis; to the contrary, there is some evidence (albeit mixed) that CSAM consumers are at risk of committing hands-on abuse; viewing AIG-CSAM is still viewing CSAM. What's more, CSAM defendants are commonly found to possess both actual and virtual CSAM, which suggests it is unlikely that CSAM consumers would wholly forgo actual CSAM in favor of exclusively consuming AIG-CSAM. In any event, it is hard to imagine any way to ethically research this "harm reduction" thesis to verify or falsify it with new evidence.

4. In the March 2024 guidance from the PAI Synthetic Media Framework's first round of cases, PAI wrote of an emergent best practice: "Creative uses of synthetic media should be labeled, because they might unintentionally cause harm; however, labeling approaches for creative content should be different, and even more mindfully pursued, than those for purely information-rich content."

   Does your organization agree? If so, how do you think creative content should be labeled? What is your organization's understanding of "mindfully pursued"? If your organization does not agree, why not?

Stanford HAI does not have an organizational viewpoint on these questions; this is the viewpoint of the authors.

As noted, AIG-CSAM is inherently harmful regardless of how "creative" it is and whether and how it is labeled; there is no "mindful" approach to AIG-CSAM. That said, it is also possible to use AI image generation models to create innocuous imagery of children, which presents different, but related, considerations.

Since models may generate explicit content unprompted, any model that has been trained on both sexually explicit imagery and imagery of children could generate AIG-CSAM, even inadvertently (see Section 3.5). As a mitigation measure, one view is that **Builders** should pick one approach: If **Builders** are OK with their model being used to create sexually explicit content, no child imagery should be included in the training data, and vice versa.

Another view goes further: that imagery of children should never be included in training data at all. This is a challenging stance. After all, children are a commonality of life on Earth and presumptively fair game for objective representation in visual works. But in this view, there are distinct downsides to the ability to create AIG imagery of children that outweigh the upsides. AIG-CSAM is the biggest but not the only harm. Some harms can't be fully mitigated by disclosure, such as the privacy and consent issues with training a model on real children's images. Others might prove more susceptible to disclosure, like labeling AIG pictures of unrealistically thin or muscular bodies to avoid contributing to negative body image or disordered eating. (Some policymakers have proposed mandatory disclosure of digital alterations to bodily images, though, in the U.S., this would likely be unconstitutional.) By contrast, in this view, there are limited innocuous use cases for AIG imagery of children — even if they are creative or artistic — and those can readily be fulfilled by conventional means (e.g., making a children's book with traditional photography or illustration). That is, there is little marginal upside and major downside.

Direct disclosure of non-sexual AIG imagery of children could help inform the debate over children's proper place (if any) in AI imagery. Disclosure could aid with documentation of what innocuous use cases exist, whose marginal benefit and prevalence could then be weighed against the downsides and prevalence of the harmful use cases. Disclosure might also help track when unforeseen harms arise from seemingly innocuous uses.

5. Overall, what role(s) does your organization believe Builders, Creators, and Distributors play in directly disclosing AI-generated or AI-edited media to users?

The authors believe that, ultimately, the responsibility has to be with the **Builders** of AI image generation models — there are no legitimate **Creators** or **Distributors** of AIG-CSAM. Likewise, there are no legitimate "users" of AIG-CSAM; as categorized above, AIG-CSAM may be created by minors for minors (i.e., with "nudify" apps), by adults for adults, or by adults for (grooming) minors. None of these is a legitimate use case. Where **Builders**' guardrails against CSAM prove inadequate, it is imperative that they invest in improvements. For example, the **Builders** of the LAION-5B dataset re-launched a cleaned-up dataset after the original was found to contain confirmed CSAM and taken down. The re-launch took eight months.

Given this and other pitfalls, as said, **Builders** must consider whether to include children's imagery in their training sets at all, and **Distributors** can decide whether to impose prerequisites for hosting or distribution of models. **Creators** who wish to create (non-explicit) images of children should do a good-faith cost-benefit analysis of the potential harms, marginal benefit, and effectiveness of disclosure as a mitigation.

The one legitimate purpose of disclosure in the AIG-CSAM context is to support the identification, removal, and investigation of online AIG-CSAM, which occurs through the CyberTipline ecosystem for routing instances of CSAM on online platforms to the appropriate law enforcement agency via NCMEC. The standard CyberTipline report form now includes a "Generative AI" checkbox, but it is up to platforms to check that box accurately and consistently.

As noted above, there are different potential audiences for the disclosure that a piece of CSAM is AI-generated. When enabled by **Builders**, direct and/or indirect disclosure (especially the latter) could help the following audiences:

**Platforms (email, social media, other user-generated content sites)**

- Purpose: Streamlining and triaging the detection, review, and reporting to NCMEC of CSAM
    - Removal of AI-CSAM under terms of service; determining which images to surface for human moderator review; determining whether reporting is required
    - Open legal question: Federal law requires reporting "apparent" CSAM; can platforms rely on disclosure that an image is fully virtual to justify not reporting?

**National Center for Missing & Exploited Children (NCMEC)**

- Purpose: Triaging and annotating incoming CyberTips received from platforms before passing them to law enforcement; knowing that a real child isn't depicted would be immensely helpful in conserving resources at NCMEC & legal enforcement agencies that would otherwise be wasted (i.e., no victim identity efforts need be undertaken, there's no child who needs to be rescued)

**Law enforcement (investigators, prosecutors)**

- Purpose: Triaging incoming reports typically from NCMEC but sometimes from members of the public to determine the legality of material, know whether there's a real child involved, and decide which cases to prioritize for investigation
- In other countries, AIG-CSAM may be wholly illegal (whereas some AIG-CSAM is First Amendment-protected in the United States), but indirect/direct disclosure would still aid in triage as imagery of real children should be higher-priority than fully virtual CSAM

6. How important is it for those Building, Creating, and/or Distributing synthetic media to all align collectively, or within stakeholder categories, on a singular threshold for:

   1) the types of media that warrant direct disclosure, and/or

   2) more specifically, a shared visual language or mechanism for such disclosure?

   Elaborate on which values or principles should inform such alignment, if applicable.

Again, as an initial matter, AIG-CSAM is a category of synthetic media that cannot legitimately exist, which renders this question largely irrelevant to this particular use case. However, it is important to the broader governance of generative AI technologies, especially synthetic media.

CSAM is just about the only category of content that is illegal worldwide, meaning the incentives of **Builders**, **Creators**, and **Distributors** of AIG-CSAM are, at most, aligned around minimizing legal risk. They may align on disclosure as a norm only if they believe (rightly or wrongly) that it will insulate them from legal liability. Of course, the majority of people want nothing to do with CSAM and are similarly motivated to minimize their legal exposure for it, so disclosure could aid them as well.

Previous Stanford research revealed that the LAION-5B dataset included hundreds of links to confirmed CSAM. As the news spread, it functioned akin to a disclosure that all images created with LAION-5B had been trained on a "tainted" model, which its **Builders** took down. The news also put downstream parties (**Creators**, **Distributors**, etc.) on notice of potential legal risk for continued possession of the LAION-5B dataset. The incident revealed an implicit reliance on **Builders**, especially open source ones, to vet their training data — at least for material known to be universally illegal.

The episode suggests that both good-faith and AIG-CSAM-focused **Builders**, **Creators**, **Distributors**, etc., could align around a norm of **Builder** disclosure of their training data curation and vetting practices — if only for CSAM (not, e.g., copyrighted or Global Internet Forum to Counter Terrorism [GIFCT] material), given its singular legal status. A potential norm is disclosing that an AIG-CSAM image was generated with a model trained on a "clean" dataset that didn't contain actual CSAM. Such disclosure is also possible at the dataset level. When launching "RE-LAION 5B," LAION called their updated dataset "the first web-scale, text-link to images pair dataset to be thoroughly cleaned of known links to suspected CSAM." Whether someone who wished to download that image or use that dataset would trust and act in reliance on the "clean" representation is a different question.

Research into AIG-CSAM trading communities (for example, child safety organizations such as Thorn and the IWF conduct) could provide insight into what, if any, norms those communities develop regarding disclosure. However, possessing or creating an image trained on "clean" data might still incur liability, given legal nuances, country-by-country variations, and the current state of flux in the law regarding AIG-CSAM and NCII. There is a similar legal ambiguity around knowingly possessing a “tainted” dataset or model. In short, direct and indirect disclosures are incapable of reliably and definitively conveying the **legality** of an AI-generated image; they can only convey **facts** about a model or dataset or how an image was created or modified. Accurate factual disclosure can, at most, help downstream actors assess legal risk for themselves.

That said, unlike good-faith actors, not all **Builders**, **Creators**, and/or **Distributors** of AIG-CSAM will know or care about their legal liability. Therefore, we can distinguish Builders of general-purpose image generation models (or publishers of general-purpose content hosting platforms) from **Builders** of models tailor-made to generate CSAM. Good-faith **Builders** may take measures to prevent the misuse of their models, such as building in direct or indirect disclosure mechanisms. However, malicious Creators and Distributors may fine-tune the model to circumvent Builders’ disclosure mechanisms and/or strip out disclosures from images. In this way, the incentives of good-faith, general-purpose model **Builders** (or Publishers) may diverge from those of AIG-CSAM-specific model **Builders**, **Creators**, and **Distributors**.

Since motivated bad actors in the AIG-CSAM space will disregard the law, policy interventions might more constructively focus on incentivizing and assisting good-faith actors to make synthetic media and open-source models more robust against adversarial manipulation. Policies should also be realistic about the degree to which it is feasible for good-faith actors to stop bad-faith ones from circumventing protective guardrails.

*Since motivated bad actors in the AIG-CSAM space will disregard the law, policy interventions might more constructively focus on incentivizing and assisting good-faith actors to make synthetic media and open-source models more robust against adversarial manipulation.*

# 5 Approaches to Direct Disclosure, in Policy and Practice

1. What does your organization believe are the most significant socio-technical challenges to successfully achieving the purpose of directly disclosing content at scale? (Refer to question 2.3 for reference to PAI's description of direct disclosure)

Child exploitation and nonconsensual sexual content are listed among the potential harms from synthetic media that PAI's Framework seeks to mitigate. However, the harms from AIG-CSAM are less amenable to mitigation via direct disclosure. As a use case for generative AI, AIG-CSAM is abusive regardless of context; its existence is harmful, so prevention is the most responsible practice. Bad actors are generally not incentivized to disclose that an image is AI-generated in any event. But many misuses of generative AI (e.g., election disinformation, faked evidence in court) are hindered by "baking in" direct (or indirect) disclosure so that bad actors cannot avoid it. This is not so with AIG-CSAM — harm is done when the image is created, whether with disclosure or no disclosure. The Framework could do more to highlight that (as one author's past research has found) CSAM poses a unique challenge for harm mitigation and strategies that are effective against other types of harmful content may not be as effective for CSAM.

In other words, the most significant challenges in this domain aren't the technical challenge of ensuring direct disclosure at scale, the social challenges of confirmation bias, particular vulnerability to being fooled by fake imagery (e.g., young children), etc. Rather, it's the social problems of misogyny (image-based sexual abuse disproportionately affects girls) and of a small but seemingly intractable percentage of the population being sexually attracted to children.

That said, this case study repeatedly discusses the value of disclosing AIG-CSAM for reporting, triage, and investigation purposes. For those purposes, significant sociotechnical challenges include the collective action problem of having **Builders** of general-purpose image generation models, particularly open-source models, commit to the **prevention** of their models' misuse, in addition to addressing the technical challenges of "baking in" direct disclosure in a way that is difficult to strip out. Open-source models present the thorniest problem since disclosure and other guardrails can currently be fine-tuned away by motivated bad actors.

2. What goals should organizations be trying to accomplish when implementing direct disclosure? Does your organization believe directly disclosing ALL AI-generated or modified, as several policies are recommending, is useful in helping accomplish those goals?

Stanford HAI does not have an institutional position on this question, which is a non sequitur in the AIG-CSAM context. There is no legitimate use case for CSAM, so there are few legitimate goals to be served by direct disclosure that CSAM is AI-generated or -modified. The authors believe the only valid goals that direct disclosure helps to accomplish are those associated with reporting, triage, investigation, and legal proceedings. Being able to identify the face of the image that a piece of CSAM does not involve a real child helps stakeholders (e.g., platform Trust & Safety teams, NCMEC, law enforcement, schools, parents) respond more efficiently. Given its unique harms, AIG-CSAM shows that the goals of synthetic media direct disclosure are use case-dependent rather than universal. Direct disclosure alone cannot serve as a panacea for all the harms of synthetic media. Even if it were possible to build tamper-proof direct disclosure into all AI-generated or -modified content, this could not mitigate the harms inherent in the creation of AIG-CSAM.

3. Please share your organization's insight into how direct disclosure can impact:

   1) Accuracy
   2) Trustworthiness
   3) Authenticity
   4) Harm mitigation
   5) Informed decision-making
   6) Anything else we're missing that is relevant here

   NOTE: You can also discuss your understanding of the relationship between these concepts (for example, authenticity could impact trustworthiness, harm mitigation, etc.)

Stanford HAI does not have an organizational viewpoint on these questions; this is the viewpoint of the authors.

It is hard to see how direct disclosure that CSAM is AI-generated has much relevance to most of these factors. Of this list, the ones where direct disclosure has the greatest impact are:

- **Harm mitigation:** Especially in the "deepnude" context. While someone victimized by a deepnude suffers harm (emotional, privacy, etc.) merely from being depicted in it, direct disclosure may potentially mitigate the harms that occur (or that a victim fears will occur) if others believe the image is real, for example, reputational harms among the victim's peers, or punishments meted out — perhaps violently — by authority figures (such as parents or law enforcement). In a society still rife with sexual shaming (particularly of female and LGBTQ+ people), the stigma attached to real nudes results in victim-blaming for NCII. If it is clear that an image is fake, that may mitigate some, though not all, victim-blaming.
    - In this respect, trustworthiness and authenticity interact with harm mitigation — the harm mitigation flows from the trust that (per the direct disclosure) the image is not authentic. Deepnude victims are harmed differently and potentially even more when it is plausible that the nude image is real.
- **Informed decision-making:** Stakeholders such as mandatory reporters, online platforms, NCMEC, law enforcement, and the courts can deal more efficiently with CSAM if it has been disclosed as AI-generated or -modified, as that information assists with decision-making about whether the image must be reported under applicable law, how to triage and process CSAM reports, whether to initiate victim identification and location efforts, and the legality of the image.

Direct disclosure might have a harm mitigation impact to the degree it shapes informed decision-making about creating, distributing, or consuming AIG-CSAM. However, CSAM **Creators**, **Distributors**, and consumers are not uniform in their incentives. For example, a **Creator** who makes a deepnude to harass and humiliate their victim might not care if it's disclosed as AI-generated, whereas one who creates it for sextortion purposes may be stymied by direct disclosure, as that vitiates the image's power as blackmail. Likewise, some AIG-CSAM **Creators**, **Distributors**, and consumers may (for purposes of sexual gratification) prefer imagery that is not labeled as AI and thus is plausibly authentic. By contrast, others may believe (perhaps misguidedly) that they are on safer legal ground when AIG-CSAM is labeled as such. In that case, the principle of direct disclosure actually provides a perverse incentive to create, distribute, and consume imagery that, as we have said repeatedly, cannot legitimately exist. This is a potent illustration of the shortcomings of the direct disclosure framework when applied to the unique context of AIG-CSAM.

4. Does your organization believe there will be a tipping point to the liar's dividend (that people doubt the authenticity of real content because of the plausibility that it's AI-generated or AI-modified)? Why or why not? If yes, have we already reached it? How might we know if we have reached it?

Stanford HAI does not have an organizational viewpoint on these questions; this is the viewpoint of an individual contributor from HAI.

The liar's dividend has been a concern in the CSAM context for decades. Earlier computer graphics tools such as Photoshop gave rise to the worry that criminal defendants would escape liability for CSAM charges by claiming that real CSAM was computer-generated, prompting federal legislation banning "virtual" CSAM, which was struck down as unconstitutional by the Supreme Court in 2002. Several justices speculated that the decision might warrant revisiting if the "liar's dividend" defense started to succeed as technology evolved further.

That worry was largely illusory prior to the advent of generative AI. The state of the art did not yet enable highly realistic imagery, and prosecutors (sometimes aided by expert testimony) could prove beyond a reasonable doubt that images were real, so this defense routinely failed. But, as one of this case study's authors explained in a Lawfare paper, even if generative AI has reached or does reach the point of enabling truly photorealistic AIG-CSAM, it doesn't follow that the liar's dividend will doom criminal prosecutions. When prosecutors introduce sufficient proof that an abuse image is real (e.g., lay witness testimony, expert testimony, image metadata), raising reasonable doubt in a jury's mind requires actual evidence to the contrary, not speculation and conjecture.

Regardless, the few data points available so far (discussed below in Section 6) suggest prosecutors are sidestepping the entire "liar's dividend" issue by charging defendants with crimes to which "it's AI-generated or -modified" is not a defense, particularly obscenity laws.

5. As AI-generated media becomes more ubiquitous, what are some of the other important questions audiences should be asking in addition to "is this content AI-generated or AI-modified," especially as more and more content today has some AI-modification?

- Who benefits and is harmed if I believe this content is AI-generated or -modified when it's not? Who benefits and is harmed if I believe it's real and authentic when it's not?
- What, if any, harms does this content cause merely by existing or being distributed, whether it's AI-generated or -modified, or not? Are there different or additional harms if it's real vs. being AI-generated or -modified?
- Does viewing this content, individually or cumulatively, leave me better or worse off?
- How would I go about determining whether this content is AI-generated or -modified? What sources would I consult?
- How would I respond differently to this AI-generated or -modified content if I knew it was real and authentic?

6. How can research help inform development of direct disclosure that supports user/audience needs? Please list out key open areas of research related to direct disclosure that, the answers to which, would support your organization's understanding of direct disclosure.

As described above, evidence about an image's provenance and (in)authenticity can potentially incriminate or exculpate a criminal defendant. What's more, as the Lawfare paper describes, AIG-CSAM's legality may depend on whether or not there was real CSAM in the image's training data, but that will often be unknown once an image has been disseminated beyond its Creator. Direct and especially indirect disclosure practices could thus aid the justice system by helping to expeditiously resolve key evidentiary questions in court cases, as well as at earlier stages of investigation (as when law enforcement is attempting to identify a potential child victim). Open areas of research include:

- Which content provenance and authentication technologies are accepted (or rejected) as sufficiently reliable to hold up in actual court cases, and on what basis?
- What other evidence have judges and juries relied upon in court cases to decide whether the evidence is sufficient to convict a CSAM defendant?
- The development of tamper-resistant and/or tamper-evident direct and indirect disclosure techniques for still images and video
- Methods for determining (to a high confidence level) the training data underlying a particular image — either a specific image (a holy grail not just for CSAM but other domains such as copyright) or, even if not the specific image, a particular category, or set of images

# 6 Media Literacy and Education

Stanford HAI does not have an organizational viewpoint on these questions; this is the viewpoint of the authors.

Public education plays an important role in interventions against AIG-CSAM, though perhaps not as much in relation to direct disclosure. As we discussed above, the Creators and Distributors of AIG-CSAM are rarely incentivized to disclose that their content is AI-generated. The ability of those who see the content to understand whether it is synthetic also doesn't prevent harm from occurring even if it may alter it, as discussed above.

Nevertheless, broader public education is sorely needed to combat the harms of AIG-CSAM, especially in the teen "deepnude" context. Teaching teens and tweens that deepfake nudes cause serious harm and are illegal must be a part of sex education in schools and at home, in the same way that we teach children about consent and respect for bodily autonomy. Without such education, minors may think of deepnudes as funny and harmless pranks or as less harmful alternatives to watching porn. They need to learn about the consequences of such synthetic content, including the very real possibility that once they share the content, they will lose control over its circulation, the serious harm to victims, and the legal ramifications.

Education can also play a role for adults. A common line of thought is that types of AIG-CSAM that are not generated by uploading images of real minors is acceptable because the content does not depict real children and no real children are harmed. It is, therefore, crucial that there are education efforts by NCMEC or similar organizations that help clarify that the creation and dissemination of AIG-CSAM is illegal in many cases and just as harmful as traditional CSAM.

Ultimately, most adults who engage in AIG-CSAM creation, dissemination, or consumption know that it is legally and ethically wrong, so education will likely have a limited impact on that audience (though fear of legal liability could have some deterrent effect). However, educational efforts could be more fruitful with other groups of adults. For example, parents may not know about the deepnudes phenomenon or how to help their victimized child or keep their child from victimizing others. School personnel may fail to take deepnudes of students seriously because they are "not real nudes," and local law enforcement may not be trained on how relevant laws apply to CSAM made with generative AI, particularly at a time when many states are introducing new laws on this topic. Education on the legal issues and harms associated with AIG-CSAM would help ensure that these adults in positions of trust and authority are not failing children.

1. In the March 2024 guidance from the Synthetic Media Framework's first round of cases, PAI wrote of an emergent best practice: "Broader public education on synthetic media is required for any of the artifact-level interventions, like labels, to be effective."

Does your organization agree? If so, why? Has your organization been working on "broader public education on synthetic media"? How? (please provide examples.) If your organization does not agree, why not? What responsibility do Builders, Creators, or Developers (as defined in the Framework) have in educating users? What about civil society organizations?

2. What would you like to see from other institutions as it relates to improving public understanding of synthetic media? Which stakeholder groups have the largest role to play in educating the public (e.g., civic institutions; technology platforms; schools)? Why?

Please see our thoughts about this in Section 6.1.

3. What support does your organization need in order to advance synthetic media literacy and public education on evaluating trustworthiness?

Stanford HAI currently primarily engages in public education on synthetic media literacy through translating technical research on related issues for policy and general audiences.

The authors believe that beyond the much-needed public education outlined in Section 6.1., it is also crucial to educate policymakers on the nuances of direct disclosure. As various state and federal legislators move closer to mandating the watermarking of AI-generated or -modified content, it's important they consider what impact direct (and indirect) disclosure can and can't have in different contexts. While direct disclosure mechanisms may be an effective harm reduction measure in the mis- or disinformation context, our case study shows a more complex picture for the specific context of AIG-CSAM. As we discussed above, direct disclosure can still have an important role to play, as it may, for example, enable better streamlining of law enforcement resources. Yet, to substantially mitigate harm, policy solutions must look beyond direct disclosure and consider interventions during the training data curation, model development, and model hosting stages of AI development.