# AI-GENERATED CHILD SEXUAL ABUSE MATERIAL

*Insights from Educators, Platforms, Law Enforcement, Legislators, and Victims*

**Shelby Grossman** | Stanford Cyber Policy Center

**Riana Pfefferkorn** | Stanford HAI

**Sunny Liu** | Stanford Cyber Policy Center

STANFORD CYBER POLICY CENTER DIRECTOR:
JEFFREY T. HANCOCK

May 29, 2025

Supported by

Safe Online

Stanford
Cyber Policy Center
*Freeman Spogli Institute*

# Contents

# 1 Introduction

AI-generated child sexual abuse material (CSAM) carries unique harms. When generated from a photo of a clothed person, it can damage that person's reputation and cause serious distress. When based on existing CSAM, it risks re-traumatizing victims. Even AI CSAM that seems purely synthetic may come from a model that was trained on real abusive material. Many experts also warn that viewing AI CSAM can normalize child abuse and increase the risk of contact abuse. There is the added risk that law enforcement may mistake AI CSAM for content involving a real, unidentified victim, leading to wasted time and resources spent trying to locate a child who does not exist.[1]

In this report we aim to understand how educators, platform staff, law enforcement officers, U.S. legislators, and victims are thinking about and responding to AI CSAM. We interviewed 52 people, analyzed documents from four public school districts, and coded state legislation.[2]

Our main findings are that while the prevalence of student-on-student nudify app use in schools is unclear, schools are generally not addressing the risks of nudify apps with students, and some schools that have had a nudify incident have made missteps in their response. We additionally find that mainstream platforms report the CSAM they discover, but, for various reasons, without systematically trying to discern and convey whether it is AI-generated in their reports to the National Center for Missing and Exploited Children's (NCMEC) CyberTipline. This means the task of identifying AI-generated material falls to NCMEC and law enforcement. However, frontline platform staff believe the prevalence of AI CSAM on their platforms remains low. Finally, we find that legal risk is hindering CSAM red teaming efforts for mainstream AI model-building companies.

1. Among child safety groups, there is no consensus on shorthand for AI-generated child sexual abuse material. We use the acronym AI CSAM as we think it is more accessible to non-experts than AIG-CSAM. AI CSAM is also the term favored by the U.K.'s Internet Watch Foundation. (*See, e.g.*, Internet Watch Foundation, *What Has Changed in the AI CSAM Landscape?* (July 2024), https://www.iwf.org.uk/media/nadlcb1z/iwf-ai-csam-report_update-public-jul24v13.pdf.)

2. We thank Shubhi Mathur and John Perrino for research support, and David Thiel for proposing the original concept for this report. All errors remain our own. Shelby Grossman discloses that she does occasional contracting for OpenAI through their external Red Teaming Network. This publication has been produced with financial support from Safe Online. However, the opinions, findings, conclusions, and recommendations expressed herein are those of the authors and do not necessarily reflect those of Safe Online.

## 1.1 Key findings

State legislation:

- At least **38 U.S. states now have laws relevant to AI CSAM**, many of them passed just within the last few years. There has been broad bipartisan support for this legislation. Adding to longstanding federal CSAM laws, the new federal TAKE IT DOWN Act will further penalize AI CSAM.
- While states have worked to criminalize AI CSAM, they failed to anticipate that student-on-student cases would be a common fact pattern. As a result, **that wave of legislation did not account for child offenders**. Only now are legislators beginning to respond, with measures such as bills defining student-on-student use of nudify apps as a form of cyberbullying.

Schools:

- **Some respondents believe that the use of nudify apps in schools is widespread, while others view it as rare.** We were unable to identify any patterns among those holding either view and are not sure what to make of these conflicting perspectives.
- **Most schools are not currently addressing the risks of nudify apps with students.** We are not taking a position on whether this is good or bad, as there are legitimate concerns that this instruction could inadvertently educate students about the existence of these apps.
- For young people (whose brains have not yet fully developed), there is a **spectrum of intent and motivation among children who make and share nudified images**, from many likely not realizing the consequences of their actions, to others doing this at scale.
- Student victims are further **traumatized when school staff respond badly to nudify incidents**.
- Anonymous **school safety tip lines (often at the state-level) are an important channel** for schools to learn about nudify incidents.
- There is **no school-level representative descriptive data on whether students are receiving instruction about online exploitation risks.** Additionally, only a small number of non-comparable randomized controlled trials have evaluated the effectiveness of such instruction.

Perceived legitimacy and poor visibility:

- Respondents told us there is a **sense of normalization or legitimacy among those who create and share AI CSAM**. This perception is fueled by open discussions in online forums, a sense of community through the sharing of tips, the accessibility of nudify apps, and the presence of community members in countries where AI CSAM is legal.

- NCMEC's statistics on the number of **CyberTipline reports involving AI are a poor proxy for the scale of the issue.** These figures likely represent an undercount, as content created and shared on the dark web is unlikely to be reported. At the same time, they may also be an overcount, as generative AI companies often submit reports of users who attempt, but fail, to generate AI CSAM.
- **Law enforcement generally lacks tools to determine definitely if media is AI generated.**

Platform observations:

- **Mainstream platforms report confidence in their ability to identify AI-generated CSAM as CSAM**, but they are **less certain about their capacity to consistently and accurately label it as AI-generated.**
- Relatedly, **many platforms are not attempting to assess if CSAM is AI-generated**, and thus not communicating information about this in CyberTipline reports.
- Platforms and NCMEC find **contextual information surrounding images (e.g., text, group/channel name) useful for labeling CSAM as AI-generated.**
- **Platforms crave information about how NCMEC and law enforcement handle CyberTipline reports with AI CSAM.**
- **AI companies face real risks and legal uncertainty around red teaming for CSAM.**

Welfare implications:

- **There are welfare implications for law enforcement, NCMEC staff, and platform staff who are exposed to AI CSAM.**
  - ↣ Law enforcement officers may need to spend more time analyzing CSAM than they previously did in order to determine whether it is AI-generated. This can be harmful.
  - ↣ AI CSAM is often more extreme than non-AI CSAM, which can make it especially distressing to view. One respondent described it as "nightmarescape" content.
  - ↣ Viewing AI CSAM that depicts victims from known abuse series is particularly distressing for NCMEC and platform staff. For NCMEC staff, it is emotionally difficult to notify these now-adult survivors about the circulation of this new content, to say nothing of the harm to these survivors.

## 1.2 Methodology

Between June 2024 and February 2025 we conducted semi-structured interviews with 52 respondents. We identified respondents primarily by leveraging our existing contacts or reaching out through online contact forms and publicly available email addresses. Our respondents included:

- 9 NGO employees
- 8 online platform staff
- 8 providers, the term we are using to describe groups that offer online trainings for schools
- 7 academics
- 6 current and former law enforcement officers
- 4 state legislators and state legislative aides
- 3 staff at the National Center for Missing and Exploited Children
- 2 U.S. government employees
- 2 victims, the term we are using to describe individuals who were targeted by nudify apps as children
- 1 parent of a victim
- 1 lawyer for a victim
- 1 teacher

Our response rate was approximately 50%. The response rate was driven by two sets of reluctant respondents: school leaders and vendors that offer AI red teaming services. We were unable to secure interviews with individuals from either group.

Using public records requests, we also obtained documents from four public school districts in California, New Jersey, Texas, and Washington state about student-on-student deepfake nude incidents at a total of five schools in those districts. We identified the schools from news reports about those incidents. While multiple U.S. private schools have also reportedly experienced such incidents, we did not obtain records from those schools because private schools are not subject to public records laws.

Our research was approved under Stanford University IRB protocols #76159 and #70974. We are grateful to staff from Brave Movement[3] for providing resources on interviewing victims.

3. Brave Movement, https://www.bravemovement.org/.

# 2 Legal background & findings

Starting in 2023, researchers found that generative AI models were being misused to create sexually explicit images of children.[4] That phenomenon—a new twist on the longstanding societal problem of child sexual abuse and exploitation (CSEA)—is testing the existing legal regimes that govern various sectors of society impacted by AI CSAM, illuminating some gaps and ambiguities. While legislators around the U.S. have been atypically quick to take action on some aspects of the AI CSAM problem, opportunities for further regulation or clarification remain. This section lists some of the key stakeholders and summarizes their positions under current law. (Nothing in this section or elsewhere in the report constitutes legal advice.)

## 2.1 Legislators

In a rare show of bipartisan consensus and momentum, over 20 states have enacted AI CSAM-related laws since 2022.[5] By April 2025, at least 38 states had a relevant law on the books.[6] Most of those are too new to have been enforced yet, though we started seeing criminal cases brought under new laws while drafting this report.[7] AI is a rapidly-evolving area of technology, and while we encountered

4. David Thiel, Melissa Stroebel & Rebecca Portnoff, *Generative ML and CSAM: Implications and Mitigations* (June 2023), https://fsi.stanford.edu/publication/generative-ml-and-csam-implications-and-mitigations; Internet Watch Foundation, *How AI is Being Abused to Create Child Sexual Abuse Imagery* (October 2023), https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf.

5. *Artificial Intelligence 2024 Legislation*, Nat'l Conf. State Legs. (Sept. 9, 2024), https://www.ncsl.org/technology-and-communication/artificial-intelligence-2024-legislation; MultiState, *Combating Sexual Deepfakes*, https://www.multistate.ai/deepfakes-sexual, last updated May 22, 2025.

6. ENOUGH ABUSE, *State Laws Criminalizing AI-Generated or Computer-Edited CSAM*, https://enoughabuse.org/get-vocal/laws-by-state/state-laws-criminalizing-ai-generated-or-computer-edited-child-sexual-abuse-material-csam/, last visited May 25, 2025. At publication time, Congress was considering a 10-year moratorium on non-criminal enforcement of certain state AI laws and regulations. Justin Hendrix & Cristiano Lima-Strong, *US House Passes 10-Year Moratorium on State AI Laws*, Tech Pol'y Press (May 22, 2025), https://www.techpolicy.press/us-house-passes-10year-moratorium-on-state-ai-laws/.

7. Sara Ruberg, *Darrin Bell, Pulitzer-Winning Cartoonist, Faces Child Pornography Charges*, N.Y. Times (Jan. 16, 2025), https://www.nytimes.com/2025/01/16/us/darrin-bell-arrest-child-pornography.html; Anthony Maenza, *Dallastown Man is First in Pennsylvania to Be Charged with Possessing AI-Generated Child Pornography*, York Dispatch (Apr. 15, 2025), https://www.yorkdispatch.com/story/news/crime/2025/04/15/dallastown-man-is-first-in-pennsylvania-to-be-charged-with-possessing-ai-generated-child-pornography/83095033007/; Carl Crabtree, *Idaho Attorney General Announces Prison Sentences for Three Idaho Men in Child Pornography Cases*, Big Country News (May 2, 2025), https://www.bigcountrynewsconnection.com/idaho/idaho-attorney-general-announces-prison-sentences-for-three-idaho-men-in-child-pornography-cases/article_68e98394-6af5-4619-ba03-7f9c214741c8.html.

some dismissiveness toward legislators' technical savvy,[8] the state-level legislators and legislative staffers we interviewed all came across as well-informed, though we recognize that this could reflect selection bias.

One legislative trend has been to target so-called "morphed" images—innocuous photos of real, identifiable children that have been altered to appear sexually explicit.[9] Federal CSAM law has long banned morphed images,[10] but many states did not, until the abuse of novel generative AI tools including nudify apps prompted states to start closing that gap.[11] Another legislative strategy for banning AI CSAM is to include an obscenity requirement in the bill language,[12] since obscene material (be it photographic, computer-generated, cartoon, etc.) is unprotected speech regardless of whether it depicts an actual child.[13] (Federal law already criminalizes "obscene visual representations of the sexual abuse of children."[14])

The obscenity strategy responds to a well-known drafting challenge for legislators in this domain: a 2002 Supreme Court ruling that (unless it is obscene) fully "virtual" CSAM is constitutionally protected speech because it does not involve any real child.[15] Our legislative respondents were consistently aware of this case and told us they had tried to draft their bills to be able to survive constitutional challenge.[16] However, some respondents suggested that the time has come to revisit the ruling in light of technological advances since 2002 that now enable the creation of highly photorealistic material.[17]

When we asked legislative respondents their motivation for introducing AI CSAM bills, multiple staffers referred to a "gap" in their states' existing law.[18] Several noted that law enforcement was encountering AI CSAM, but that without a state statute that clearly covered AI-generated material and not just imagery of real children, prosecutors either had to get creative or refused to take up those cases.[19]

8. Interview with a former law enforcement officer on Jan. 6, 2025.

9. *E.g.*, Alabama Child Protection Act of 2024, Act of Apr. 24, 2024, no. 2024-98, 2024 Ala. Acts; Walker Montgomery Protecting Children Online Act, Act of Apr. 30, 2024, ch. 456, 2024 Miss. Acts.

10. 18 U.S.C. § 2256(8)(C); *see generally* Riana Pfefferkorn, *Addressing Computer-Generated Child Sex Abuse Imagery: Legal Framework and Policy Implications*, Lawfare (February 2024), https://www.lawfaremedia.org/article/addressing-computer-generated-child-sex-abuse-imagery-legal-framework-and-policy-implications.

11. Natasha Singer, *Spurred by Teen Girls, States Move to Ban Deepfake Nudes*, N.Y. Times (Apr. 22, 2024), https://www.nytimes.com/2024/04/22/technology/deepfake-ai-nudes-high-school-laws.html.

12. *E.g.*, Cal. Penal Code §§ 311, 311.2, 311.11, 311.12 (2024); 720 Ill. Comp. Stat. 5/11-20.4 (2025); Va. Code Ann. § 18.2-374.1 (2024); S.B. 5105-S.E., 2025 Reg. Sess. (Wash. 2025).

13. *See* Pfefferkorn, *supra* note 10, at 3–4.

14. 18 U.S.C. § 1466A.

15. Ashcroft v. Free Speech Coalition, 535 U.S. 234 (2002).

16. Interview with a state legislative aide on Dec. 9, 2024; Interview with a state legislative aide on Jan. 6, 2025.

17. Interview with Florida State Senator Jennifer Bradley on Nov. 25, 2024; Interview with a state legislative aide on Dec. 9, 2024.

18. Interview with a state legislative aide on Dec. 9, 2024; Interview with a state legislative aide on Jan. 17, 2025.

19. Interview with Sen. Bradley on Nov. 25, 2024; Interview with a state legislative aide on Dec. 9, 2024; Interview with a state legislative aide on Jan. 6, 2025.

One staffer said that state lawmakers were hearing from constituents whose daughters had been the victims of AI CSAM.[20]

Legislative respondents described encountering at most light opposition to their bills, such as a general reluctance to add new crimes to existing penal law,[21] some entertainment industry concern about bill scope,[22] and concern that bill language might not comply with the aforementioned Supreme Court decision.[23] Overall, though, respondents described ample bipartisan support.[24] "Nobody objects to trying to protect children," one legislative staffer said; if anything, the challenge was having too many allies who all "wanted their fingerprints" on the bill.[25]

## 2.2 Schools

Schools owe legal responsibilities to their students with respect to abuse and bullying. In most states, school personnel are "mandated reporters" who must report suspected child abuse or neglect, including sexual abuse, to law enforcement or child welfare services.[26] "Sexual abuse" includes using a child to engage in or simulate sexually explicit conduct in order to produce CSAM.[27] But because "nudifying" a child's image does not involve the child's participation, there is confusion over whether educators must report such incidents. While police in Washington state told a school staffer that she was mandated to report,[28] the school district later stated that "per our legal team, we are not required to report fake images to police."[29] In Pennsylvania, a district attorney concluded that school officials were *not* legally required to report under state law at the time.[30] Nevertheless, multiple victims' families still sued that school for failing to report or take corrective action

20. Interview with a state legislative aide on Jan. 17, 2025.
21. Interview with a state legislative aide on Jan. 6, 2025.
22. Interview with Sen. Bradley on Nov. 25, 2024.
23. Interview with a state legislative aide on Dec. 9, 2024.
24. Interview with Sen. Bradley on Nov. 25, 2024; Interview with a state legislative aide on Dec. 9, 2024; Interview with a state legislative aide on Jan. 6, 2025; Interview with a state legislative aide on Jan. 17, 2025.
25. Interview with a state legislative aide on Jan. 6, 2025.
26. *Policies and Procedures for Mandated Reporting*, REMS Tech. Assistance Ctr. (Mar. 5, 2024), https://rems.ed.gov/ASM_Chapter2_Reporting.aspx; *How Many States Have Mandated Reporting Laws?*, MandatedReporter.com, https://mandatedreporter.com/blog/how-many-states-have-mandated-reporting-laws/, last visited May 25, 2025.
27. *E.g.*, 42 U.S.C. § 5106g(a)(4)(A); 23 Pa. Cons. Stat. § 6303 (2024) (defining "sexual abuse or exploitation").
28. Jason Koebler & Emanuel Maiberg, *"What Was She Supposed to Report?:" Police Report Shows How a High School Deepfake Nightmare Unfolded*, 404 Media (Feb. 15, 2024), https://www.404media.co/what-was-she-supposed-to-report-police-report-shows-how-a-high-school-deepfake-nightmare-unfolded/.
29. Natasha Singer, *Teen Girls Confront an Epidemic of Deepfake Nudes in Schools*, N.Y. Times (Apr. 8, 2024), https://www.nytimes.com/2024/04/08/technology/deepfake-ai-nudes-westfield-high-school.html.
30. Lancaster County District Attorney's Office, *2 Juveniles Charged in Connection to AI-Generated Images of Lancaster Country Day School Students; DA Determines School Officials Were Not Legally Required to Report AI Incidents*, Crimewatch (Dec. 5, 2024), ttps://lancaster.crimewatchpa.com/da/11617/post/2-juveniles-charged-connection-ai-generated-images-lancaster-country-day-school-students.

after learning of the nudified images.[31]

Even if deepfake nude incidents are not deemed mandatory to report, they could still count as bullying, for which almost every state requires schools to have policies in place.[32] At least one state has introduced a bill that would add deepfake nudes as a form of cyberbullying for which public schools must have a policy.[33] Although state bullying laws do not provide a private right of action against schools, plaintiffs'-side lawyers have developed other legal theories for holding schools accountable.[34] For example, a victim's family's lawyer told us that those "responsible for the welfare of a child" are liable if they witness child abuse or sexual exploitation but "turn a blind eye to it and do nothing, … letting it happen under their watch."[35]

## 2.3 Law enforcement

Law enforcement personnel such as police officers and prosecutors are tasked with criminal enforcement of laws against child sexual exploitation and abuse. While federal authorities already had laws on the books to address the new problem of AI CSAM as it arose, that has not consistently been true for their colleagues at the state and local levels. As noted above, we found that state legislators cited law enforcement complaints about gaps or gray areas in existing laws as the impetus for introducing AI CSAM legislation, so that prosecutors would no longer have to resort to other charges or decline to bring charges at all.

At the federal level, by contrast, several decades-old criminal laws are readily applicable to AI CSAM.[36] The federal government has repeatedly stated that AI CSAM violates existing law and will be prosecuted. A September 2024 Department of Homeland Security (DHS) notice stated, "All forms of AI-created CSAM are illegal—and deeply harmful to victims and society."[37] A March 2024 public service announcement from the Federal Bureau of Investigation (FBI) cited two criminal convictions involving AI-morphed images.[38] And Steven Grocki, who heads the section of the Department of Justice (DOJ) devoted to CSEA cases, has said publicly that AI CSAM "is a crime": "These laws exist. They will be used. We have the will.

31. Ashley Stalnecker, *Where Does the Lancaster Country Day School Case Stand Now?*, LancasterOnline (Dec. 26, 2024), https://lancasteronline.com/news/local/where-does-the-lancaster-country-day-school-case-stand-now/article_1be96110-c3d4-11ef-9d7b-0bd4c3bb8449.html.

32. Justin Patchin & Sameer Hinduja, *Bullying Laws Across America*, Cyberbullying Research Center (2023), https://cyberbullying.org/bullying-laws.

33. S.B. 747, 89th Leg. (Tex. 2025).

34. Adele Kimmel, *Litigating Bullying Cases: Holding School Districts and Officials Accountable*, Public Justice (Fall 2017 ed.), https://www.publicjustice.net/wp-content/uploads/2016/02/Bullying-Litigation-Primer-Fall-2017-Update-FINAL.pdf.

35. Interview on Dec. 13, 2024.

36. 18 U.S.C. §§ 1466A, 2252A; *see generally* Pfefferkorn, *supra* note 10.

37. Dep't of Homeland Sec., *Artificial Intelligence and Combatting Online Child Sexual Exploitation and Abuse* (Sept. 2024), https://www.dhs.gov/sites/default/files/2024-09/24_0920_k2p_genai-bulletin.pdf.

38. Fed. Bureau of Investigation, *Child Sexual Abuse Material Created by Generative AI and Similar Online Tools is Illegal* (Mar. 29, 2024), https://www.ic3.gov/PSA/2024/PSA240329.

We have the resources."[39] In an interview, a federal government employee told us, "We think our tools are sufficient to meet the moment."[40] This respondent added that while there are factors such as "the scale problem" that make prosecuting AI CSAM cases difficult, "it's not for lack of statutes." To date, we are aware of over a dozen federal criminal cases that involve AI CSAM; most are still pending.

## 2.4 AI model developers

Users' misuse of generative AI models to make CSAM creates legal risk for the model developers. Online platforms are required by federal law to report "apparent" CSAM on their services to NCMEC when they find it (though they need not go looking for it).[41] That obligation covers websites that offer AI image generation features, such as Midjourney, OpenAI, and Adobe. Plus, at least one state (Pennsylvania) has passed a similar reporting requirement just for AI developers.[42] It is unclear whether the federal CSAM reporting requirement only covers actually-generated images and video, or also extends to text prompts to generative AI models (i.e., unsuccessful efforts to get a model to generate CSAM). However, some companies do report text prompts to NCMEC, as allowed by federal law.[43] While we encountered some skepticism that NCMEC or law enforcement wants prompts alone,[44] NCMEC told us they find text prompts useful even when not accompanied by an image.[45]

When their users' content breaks the law, online platforms including generative AI companies are largely immune from civil and state criminal liability, thanks to a law known as Section 230. However, that immunity does not extend to violations of federal criminal laws, specifically including CSEA laws.[46] That is, Section 230 does not shield AI companies (or any other online platform) from federal criminal liability for knowingly hosting AI CSAM or failing to report it.[47]

What's more, Section 230 may not bar civil or state criminal liability in some circumstances. Whether and how Section 230 applies to the outputs of generative AI models will depend heavily on the particular legal claims and facts.[48] Section

39. Alanna Durkin Richer, *Law Enforcement Cracking Down on Creators of AI-Generated Child Sexual Abuse Images*, PBS (Oct. 25, 2024), https://www.pbs.org/newshour/nation/law-enforcement-cracking-down-on-creators-of-ai-generated-child-sexual-abuse-images.

40. Interview on Jan. 10, 2025.

41. 18 U.S.C. § 2258A.

42. Act of Oct. 29, 2024, 2024 Pa. Laws 1095 (Act No. 125).

43. Interview on Jan. 6, 2025; 18 U.S.C. § 2702(b)(6).

44. Interview with a platform employee on Jan. 14, 2025.

45. Interview with NCMEC employees on Feb. 5, 2025.

46. 47 U.S.C. § 230.

47. For more detail, *see* Shelby Grossman, Riana Pfefferkorn, David Thiel, Sara Shah, Renée DiResta, John Perrino, Elena Cryst, Alex Stamos & Jeffrey Hancock, *The Strengths and Weaknesses of the Online Child Safety Ecosystem* (2024), https://purl.stanford.edu/pr592kc5483, https://doi.org/10.25740/pr592kc5483.

48. Peter J. Benson & Valerie C. Brannon, *Section 230 Immunity and Generative Artificial Intelligence*, Cong. Rsch. Serv. (Dec. 28, 2023), https://www.congress.gov/crs-product/LSB11097.

230 does not apply when a platform "is responsible, in whole or in part, for the creation or development of" the offending content.[49] Some courts have formulated this exception as a "material contribution" standard,[50] which excludes "neutral" tools that could be "manipulated by third parties for unlawful purposes."[51] Thus, Section 230 is likelier to protect a general-purpose generative AI model whose users misuse it to create AI CSAM than nudify websites whose whole purpose is to turn legal input (an innocuous photo of a child) into illegal output (AI CSAM).

To date, we know of no court cases under federal CSAM statutes against the developers of any generative AI models known to have been trained on CSAM,[52] or against mainstream AI model hosting platforms for hosting them.[53] However, there are legal developments underway that may prove instructive. In 2024, the San Francisco city attorney filed a first-in-the-nation civil lawsuit in California state court against various nudify sites.[54] (Although the case is ongoing, 11 of the websites named in the lawsuit have gone offline so far.[55]) Also, state legislators have begun targeting nudify-specific AI systems: Texas's Responsible AI Governance Act (TRAIGA) bill would prohibit the development or deployment of AI systems "with the sole intent of producing, assisting or aiding in producing, or distributing" CSAM or unlawful deepfake images or videos.[56]

Another open policy question is whether to make model developers liable for insufficient efforts to mitigate their models' capacity for misuse.[57] As one former law enforcement officer noted, there is no federal law requiring model guardrails against generating CSEA content.[58] That is not to say generative AI is a free-for-all: best practices for AI CSAM prevention exist, and many model developers already follow them or make other voluntary efforts.[59] However, end users can circumvent guardrails against abuse (especially when they have direct access to the model), which complicates the notion of imposing developer liability for

49. 47 U.S.C. § 230(f)(3).

50. Benson & Brannon, *supra* note 48.

51. Calise v. Meta Platforms, Inc., 103 F.4th 732, 745 (9th Cir. 2024) (citing Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157, 1169 (9th Cir. 2008) (en banc)).

52. David Thiel, *Identifying and Eliminating CSAM in Generative ML Training Data and Models* (2023), Stanford Digital Repository, https://purl.stanford.edu/kh752sm9123, https://doi.org/10.25740/kh752sm9123.

53. David Evan Harris & Dave Willner, *Was an AI Image Generator Taken Down for Making Child Porn?*, IEEE Spectrum (Aug. 20, 2024), https://spectrum.ieee.org/stable-diffusion.

54. Heather Knight, *San Francisco Moves to Lead Fight Against Deepfake Nudes*, N.Y. Times (Aug. 15, 2024), https://www.nytimes.com/2024/08/15/us/deepfake-pornography-lawsuit-san-francisco.html.

55. Ginger Conejero Saab, *SF City Attorney's Office Takes Down Websites Spreading Deepfake Nudes*, NBC Bay Area (Mar. 18, 2025), https://www.nbcbayarea.com/news/local/san-francisco/sf-city-attorneys-office-deepfake-nudes/3821669/.

56. H.B. 149, 89th Leg. (Tex. 2025).

57. *E.g.*, Bryan H. Choi, *Negligence Liability for AI Developers*, Lawfare (Sept. 26, 2024), https://www.lawfaremedia.org/article/negligence-liability-for-ai-developers.

58. Interview on Dec. 20, 2024.

59. Interview on Jan. 6, 2025; Thorn & All Tech Is Human, *Safety by Design for Generative AI: Preventing Child Sexual Abuse* (2024), https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf; Anthropic, *CSAM Detection and Reporting*, https://support.anthropic.com/en/articles/9020328-csam-detection-and-reporting, last updated May 23, 2025; The White House, *White House Announces New Private Sector Voluntary Commitments to Combat Image-Based Sexual Abuse* (Sept. 12, 2024), https://bidenwhitehouse.archives.gov/ostp/news-updates/2024/09/12/white-house-announces-new-private-sector-voluntary-commitments-to-combat-image-based-sexual-abuse.

users' abuse.[60] Perhaps partly for this reason, it appears much more common for legislative proposals to penalize individual AI CSAM creators rather than model developers.

Beyond preventing AI CSAM, regulation can also focus on detecting it. In 2024, California passed a law that, once it comes into effect next year, will require the developers of certain generative AI systems to release free AI detection tools to let users determine if certain content was produced with that developer's generative AI system.[61] It is unclear how feasible it will be in practice for covered developers to build reliable detection tools. Plus, the law only covers AI systems with large user bases, so it does not cover niche models for generating AI CSAM. Nevertheless, the law holds promise for helping platforms, NCMEC, law enforcement, and other stakeholders in detecting and investigating AI CSAM.

## 2.5 Online platforms and app stores

The generation, hosting, and dissemination of AI imagery involves multiple players in the online ecosystem—not just AI companies, but also app stores, social media platforms, messaging services, other user generated content (UGC) driven sites, and platforms such as GitHub or Civitai that host generative AI models.

The laws discussed above, Section 230 and the federal CSEA laws (including the CSAM reporting obligation), also apply to these entities. Plus, the brand-new TAKE IT DOWN Act will require "covered platforms" (which means UGC sites and sites dedicated to nonconsensual intimate imagery, among others) to take down nonconsensually-posted "intimate visual depictions" of adults and minors alike within 48 hours of a takedown request, with civil enforcement by the Federal Trade Commission (FTC).[62] It is not entirely clear from the bill language whether the takedown provision also covers AI-generated images (which the Act calls "digital forgeries") or only real imagery. Regardless, we predict that platforms will not draw that distinction when processing takedown requests, for several reasons. One, as noted, platforms are already legally required to take down apparent CSAM upon learning of it.[63] Second, as this report discusses, platforms typically do not spend time scrutinizing CSAM to determine if it is AI or real; they just remove and report it all. The new law's 48-hour takedown window strongly reinforces that

60. As background, there are open-source and non-open-source models. OpenAI models, for example, are not open-source, and attempt to prevent users from generating CSAM. Open-source models can be downloaded and modified, and may not have guardrails. If they do have guardrails against, for example, CSAM, those guardrails can typically be removed through user modifications. *E.g.*, Xiangyu Qi et al., *Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!*, arXiv: 2310.03693v1 [cs.CL] (2023), https://arxiv.org/abs/2310.03693; Robert Gorwa & Michael Veale, *Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries*, 16(2) Law Innovation & Tech. 341–91 (2024), https://doi.org/10.1080/17579961.2024.2388914; Sarah Fisher, Jeffrey Howard & Beatriz Kira, *Moderating Synthetic Content: The Challenge of Generative AI*, 37(4) Phil. & Tech. 133 (2024), https://doi.org/10.1007/s13347-024-00818-9.

61. California AI Transparency Act, Act of Sept. 19, 2024, ch. 291, 2024 Cal. Laws (S.B. 942).

62. TAKE IT DOWN Act, Pub. L. No. 119-12, § 3, 139 Stat. __ (2025).

63. Unlike other parts of the statute, the takedown provisions do not exempt content that also meets the federal definition of CSAM. *Id.*

approach. And finally, akin to the CSAM reporting law, the TAKE IT DOWN Act limits platforms' liability to the users whose content is taken down.[64] Thus, just like under existing CSAM law, removing imagery will be the safer choice legally as well as the simpler choice operationally.[65]

In short, online platforms and app stores must currently remove and report any apparent CSAM they find—including AI-generated imagery that appears to meet the federal definition of CSAM—or face potential federal criminal liability. Under the TAKE IT DOWN Act, they will also have to promptly remove imagery that gets reported as a "nonconsensual intimate visual depiction" or risk FTC action. (Perhaps coincidentally, a notorious nonconsensual deepfake pornography site went offline in early May 2025, supposedly because a "critical service provider" cut it off. It is not clear whether TAKE IT DOWN, which had yet to be signed into law by the President at that time, played any role.[66])

Under these laws, app stores and AI model hosting platforms typically will not face liability merely for happening to host nudify apps that can be used on children's images, or AI image generation models capable of outputting AI CSAM.[67] But as discussed above, the same is not necessarily true of intentional bad actors. That said, the circle of liability might expand: On the day the TAKE IT DOWN Act was signed into law, model hosting platform Civitai—long known for its nonconsensual deepfake porn problem—announced that its payment card processor was cutting it off.[68] Shortly thereafter, Civitai announced that it was banning "models and images depicting real-world individuals," citing "an increasingly strict regulatory landscape" including the TAKE IT DOWN Act and the European Union's comprehensive AI legislation.[69] Relatedly, at least one state (California) is considering making entities liable for knowingly or recklessly facilitating, aiding, or abetting the creation or intentional disclosure of AI CSAM. The bill (which may be incompatible with Section 230) expressly targets online service providers that fail to cut off "deepfake pornography services" as customers after being put on notice of the customer's illegal acts.[70]

When we asked respondents about regulating nudify apps, many recognized the challenges of cracking down on these apps' availability and discoverability. Besides word of mouth, users (whether adults or children) may discover nudify

64. TAKE IT DOWN Act, Pub. L. No. 119-12, § 3(a)(4), 139 Stat. __ (2025).

65. These same factors make the new law's takedown provisions susceptible to abuse, according to digital rights groups. Iain Thomson, *Trump Signs TAKE IT DOWN Law Meant to Stop Revenge Porn*, The Register (May 20, 2025), https://www.theregister.com/2025/05/20/take_it_down_law/.

66. Emanuel Maiberg & Samantha Cole, *Mr. Deepfakes, the Biggest Deepfake Porn Site on the Internet, Says It's Shutting Down for Good*, 404 Media (May 4, 2025), https://www.404media.co/mr-deepfakes-the-biggest-deepfake-porn-site-on-the-internet-says-its-shutting-down-for-good/.

67. Similarly, a recent U.K. bill targeting AI CSAM would ban "child sex abuse image-generators," but contains a carve-out for internet service providers. Crime & Policing Bill, Bill 187 2024-2025, § 46B(2) (version of Feb. 25, 2025), https://bills.parliament.uk/bills/3938.

68. Emanuel Maiberg, *Civitai, Site Used to Generate AI Porn, Cut Off by Credit Card Processor*, 404 Media (May 20, 2025), https://www.404media.co/civitai-site-used-to-generate-ai-porn-cut-off-by-credit-card-processor/.

69. Emanuel Maiberg, *Civitai Ban of Real People Content Deals Major Blow to the Nonconsensual AI Porn Ecosystem*, 404 Media (May 27, 2025), https://www.404media.co/civitai-ban-of-real-people-content-deals-major-blow-to-the-nonconsensual-ai-porn-ecosystem/.

70. A.B. 621, 2025-2026 Reg. Sess. (Cal. 2025).

apps through app stores, search engines, and online advertising networks, whose policies against nudify apps are imperfectly enforced.[71] A former law enforcement officer told us that since many nudify products are app-based, regulators should force app stores to kick out nudify apps, though he acknowledged that users could still download them from alternative sources, using VPNs if needed.[72] Multiple respondents observed that nudify apps that are unavailable in app stores can still readily be found via search engines.[73]

In short, even with diligent good-faith efforts to obstruct access to nudify apps by all the actors involved (app stores, ad networks, search engines, etc.), some will inevitably keep slipping through the cracks, regardless of what regulations require. Also, while beyond the scope of this report, the controversial "TikTok ban" law passed by Congress in 2024[74] (and its state-level antecedents[75]) should serve as a cautionary tale for regulators about the practical difficulties of implementing app-specific bans in particular jurisdictions.

71. Emanuel Maiberg, *Congress Pushes Apple to Remove Deepfake Apps After 404 Media Investigation*, 404 Media (Dec. 10, 2024), https://www.404media.co/congress-pushes-apple-to-remove-deepfake-apps-after-404-media-investigation/; Sarah Perez, *Google Play Cracks Down on AI Apps After Circulation of Apps for Making Deepfake Nudes*, TechCrunch (June 6, 2024), https://techcrunch.com/2024/06/06/google-play-cracks-down-on-ai-apps-after-circulation-of-apps-for-making-deepfake-nudes/; Ashley Belanger, *Google Won't Downrank Top Deepfake Porn Sites Unless Victims Mass Report*, Ars Technica (July 31, 2024), https://arstechnica.com/tech-policy/2024/07/google-starts-broadly-removing-explicit-deepfakes-from-search-results/; Emanuel Maiberg, *Google Updates Ad Policy to Ban All AI-Generated Porn*, 404 Media (May 3, 2024), https://www.404media.co/google-updates-ad-policy-to-ban-all-ai-generated-porn/; Emanuel Maiberg, *Instagram Ads Send This Nudify Site 90 Percent of Its Traffic*, 404 Media (Jan. 15, 2025), https://www.404media.co/instagram-ads-send-this-nudify-site-90-percent-of-its-traffic/.

72. Interview on Dec. 20, 2024.

73. Interview on Nov. 25, 2024; Interview on Jan. 6, 2025.

74. Protecting Americans from Foreign Adversary Controlled Applications Act, Pub. L. 118–50, div. H, 138 Stat. 955 (2024).

75. *E.g.*, Dan York & John Morris, *Montana's TikTok Ban: Breaking the Internet and Undermining Online Privacy*, Internet Soc'y (May 6, 2024), https://www.internetsociety.org/blog/2024/05/montanas-tiktok-ban-breaking-the-internet-and-undermining-online-privacy/.

# 3 How AI CSAM harms children

AI CSAM harms children, whether or not it is based on a known child.

There are two common ways AI CSAM can be created based on a known child. First, individuals may take existing CSAM and use AI to modify the image, such as by placing the victim in a new and more violent sex act. Many adult survivors of CSAM report feeling re-victimized knowing a new person has viewed their images. The creation of an altered version of this content is likely just as, if not more, distressing.[76] Second, individuals may snap photos of children in real life, or save photos of children from the internet, and then use AI models to create nude versions of these children.

We spoke with two teenagers (we call them victims) who had AI nude images created of them, and a mother of a child who was targeted in this way, to better understand the consequences of these images. We also obtained documents from several public schools through public records requests that include communications from and about victims of deepfake nude incidents at those schools.

One victim described the moment she first learned that AI-generated nude images of her had been created. Initially, she struggled to believe it was real: "People will start rumors about anything for no reason," she said. "It took a few days to believe that this actually happened."[77] Before her experience, she had never heard of something like this happening. While she was aware of students using AI for writing papers, she didn't realize AI could generate realistic nude images.

A victim's mother and another victim described the shock of seeing the images for the first time. "Remember Photoshop?" the mother asked, "I thought it would be like that. But it's not. It looks just like her. You could see that someone might believe that was really her naked."[78] One victim recalled feeling nauseated after seeing the images.[79] For another, viewing the images was a turning point—before that moment, she hadn't fully grasped the gravity of the situation. When she saw the images she realized, "I am a victim."[80] In her case, the original clothed photo had been taken from a non-social media website, demonstrating that avoiding social media does not necessarily protect someone from this form of exploitation. Describing one of these manipulated images, she said, "he took it and ruined it by making it creepy [...]." Describing a photo of someone she knows, she said, "he turned it into a curvy boob monster, you feel so out of control."[81]

76. Interview with an NGO employee on Dec. 13, 2024.
77. Interview on Feb. 11, 2025.
78. Interview on Dec. 13, 2024.
79. Interview on Feb. 11, 2025.
80. Interview on Feb. 5, 2025.
81. Interview on Feb. 5, 2025.

The mother of a victim recounted hearing about another victim who sometimes couldn't get out of bed in the morning and couldn't stop crying. She also knew of parents pulling their children out of the school.[82] She worried about the long-term effects of this anxiety on teenagers during such formative years.

In an email from a student victim to personnel at her school, the student wrote, "as a victim of the AI nude photo incidents …, I was unable to concentrate or feel safe at school. I felt very vulnerable and deeply troubled. The investigation, media coverage, meetings with administrators, no-contact order [against the perpetrator], and the gossip swirl distracted me from school and class work. This is a terrible way to start high school."[83] In an email from a student victim's parent to school officials, the parent wrote that since the offending student had returned to school, the victim student "has been dealing with extreme anxiety and panic attacks as of last night. Having him in class with her for 2 periods is just too much. She is currently home avoiding school and completely anxiety stricken."[84]

One victim said that it was "very weird" to attend school knowing that so many of the students had seen the images.[85] "I had to keep telling myself the images are fake," she added. Talking to other girls at the school who had been similarly targeted helped. "It's really unique for another person to be in this role," she said, adding that there was a strong sense of bonding among the victims.

Respondents also discussed the potential medium- and long-term consequences for victims. The mother we spoke with feared that the images could crop up in the future, potentially affecting her daughter's college applications or job opportunities or relationships.[86] She also expressed a loss of trust in teachers, worrying that they might be unwilling to write a positive college recommendation letter for her daughter due to how events unfolded after the images were revealed.

One victim said many of her friends were worried about the images reappearing, but that she thought this was unlikely given law enforcement's efforts to have the content removed.[87] The other victim, however, was worried about this. She said it was "the main thing that made me really upset."[88] She said that no one can be certain whether all versions of the images have truly been erased.

It is common to hear hopeful claims that AI CSAM could reduce the consumption of real CSAM. However, AI CSAM—even when not directly based on a real child—remains harmful. Its training data may include real CSAM,[89] and exposure to such material can normalize child abuse, densensitizing individuals to the crime.[90] Research also suggests a positive correlation between viewing CSAM

82. Interview on Dec. 13, 2024.
83. Source: Public records requests. (See Figure 4.4.)
84. Source: Public records requests.
85. Interview on Feb. 11, 2025.
86. Interview on Dec. 13, 2024.
87. Interview on Feb. 5, 2025.
88. Interview on Feb. 11, 2025.
89. Thiel, *supra* note 52.
90. Interview with an NGO employee on Dec. 13, 2024.

and contacting a child.[91] Offenders may use AI CSAM to groom a real child, as in one Wisconsin case.[92] Other recent cases indicate that defendants found to possess AI CSAM are often also found with real CSAM.[93]

While there is no clear causal evidence that viewing AI CSAM increases the likelihood of contacting a child, all respondents who had an opinion on the matter believed it would escalate abuse. No respondent suggested the opposite. As one law enforcement officer put it, "When the thrill is gone, where do you go next? What happens when you see this kid in real life?"[94]

91. Teegan Insoll, Anna Kateriina Ovaska, Juha Nurmi, Mikko Aaltonen & Nina Vaaranen-Valkonen, *Risk Factors for Child Sexual Abuse Material Users Contacting Children Online: Results of an Anonymous Multilingual Survey on the Dark Web*, 1 J. Online Trust & Safety, no. 2 (2022), https://doi.org/10.54501/jots.v1i2.29.

92. Riana Pfefferkorn & Caroline Meinhardt, *Direct Disclosure Has Limited Impact on AI-Generated Child Sexual Abuse Material*, Partnership on AI (November 2024), https://partnershiponai.org/hai-researchers-framework-case-study/.

93. *E.g.*, Nathaniel Puente, *Charlotte Man Possessed Thousands of AI-Generated Child Porn Files, Court Records Say*, WCNC (Apr. 23, 2025), https://www.wcnc.com/article/news/crime/man-possessed-ai-generated-child-porn/275-33568c0b-147a-47f2-822f-ce5982fe6459; U.S. Immigration and Customs Enforcement, *HSI Kansas City Investigation Lands Nebraska Man a 63-Month Sentence for Transporting CSAM* (Aug. 30, 2024), https://www.ice.gov/news/releases/hsi-kansas-city-investigation-lands-nebraska-man-63-month-sentence-transporting.

94. Interview on Dec. 13, 2024.

# 4 Schools and nudify apps

## 4.1 Current landscape

### 4.1.1 Prevalence

As with many online harms transformed by AI, nudify apps have not introduced a new form of abuse—skilled Photoshop users could already create fake nude images of children. These apps, however, allow those without specialized skills to create these images at scale.

An NGO employee told us that there are hundreds of nudify apps, many of which lack built-in safety features to prevent the creation of CSAM.[95] He noted that even as an expert in the field, he regularly encounters AI tools he's never heard of—yet on certain social media platforms, "everyone is talking about them." He added that while parents might recognize the Snapchat icon on a phone, they're unlikely to identify the icons for these AI tools. A law enforcement officer emphasized just how accessible these apps are: "You can download an app in one minute, take a picture in 30 seconds, and that child will be impacted for the rest of their life."[96]

Two major studies in 2024 have attempted to measure the scale of this harm. A survey by Thorn found that 6% of U.S. residents aged 13 to 20 say that someone has created a deepfake nude image of them[97]—whether by a peer at school or an adult. A report by the Center for Democracy and Technology found that 15% of students in grades 9-12 and 11% of teachers in grades 6-12 were aware of at least one case of deepfake nude images being created of a child or adult.[98] As the Thorn report notes, underreporting is likely. However, these numbers are useful as a potential floor for estimating prevalence. We are not aware of similar studies designed to provide school-level statistics.

We asked our cross-professional respondents if they had a sense of the extent to which schools had encountered students using nudify apps to create nude images of their peers. Their answers varied widely.

One provider, who works in a highly populated area, said he hasn't heard educators or principals raise concerns about nudify apps.[99] Another provider echoed this,

95. Interview on Dec. 13, 2024.
96. Interview on Dec. 13, 2024.
97. Thorn, *Deepfake Nudes & Young People* (Mar. 2025), https://info.thorn.org/hubfs/Research/Thorn_DeepfakeNudes&YoungPeople_Mar2025.pdf.
98. Center for Democracy & Technology, *In Deep Trouble: Surfacing Tech-Powered Sexual Harassment in K-12 Schools* (Sept. 2024), https://cdt.org/wp-content/uploads/2024/09/2024-09-26-final-Civic-Tech-Fall-Polling-research-1.pdf.
99. Interview on Nov. 19, 2024.

noting that such cases are rare compared to the circulation of real self-generated nude images.[100] A state legislator focused on AI CSAM said that she has not heard from constituents about nudify apps being a problem in schools.[101] Similarly, an aide to a legislator in another state working on the same issue reported no concerns from constituents.[102]

Others believe prevalence is high. One aide to a state legislator said that "there's a lot of anecdotal evidence that it's a big but undocumented problem" as schools often handle cases internally and children may not report incidents.[103] Another legislative aide said he and his boss frequently hear about nudify apps being used in schools from state legislators, constituents, school board officials, school leaders, and parents. The aide, whose boss is a Republican, added that even Democratic constituents are asking them to address this, saying they are "tired of their daughters being subjected to this."[104] A European law enforcement officer said that "there is at least one child in every school" creating nude images of peers.[105]

It is not clear what to make of these conflicting assessments of prevalence. The truth is probably somewhere in the middle. The Thorn and Center for Democracy and Technology survey findings suggest that these incidents are occurring, but the fact that some providers and legislators aren't hearing about this much suggests that, at minimum, relatively few schools have handled a known case.

### 4.1.2 Training and preparation

Though there is no systematic data on whether online exploitation risks are covered in school safety curricula, our interviews suggest that most schools are not currently addressing the risks of nudify apps with students. (We are not taking a normative stance on whether schools should address this risk with students, as there can be downsides that we discuss below.)

While some states mandate digital safety instruction in schools, schools often have substantial discretion in responding to such a mandate.[106] An academic noted that state policies have little impact on whether schools address the risks of nudify apps with students.[107]

When schools do address this or related risks, some bring in external organizations (which we refer to as providers) for standalone trainings, while others rely on teachers, counselors, or school resource officers to deliver the content.[108] In some cases, Internet Crimes Against Children (ICAC) Task Force detectives

100. Interview on Nov. 22, 2024.
101. Interview on Nov. 25, 2024.
102. Interview on Dec. 9, 2024.
103. Interview on Jan. 6, 2025.
104. Interview on Jan. 17, 2025.
105. Interview on Dec. 13, 2024.
106. Interview with an NGO employee on June 26, 2024.
107. Interview on July 17, 2024.
108. School resource officers are law enforcement officers who work in schools. They promote safety in the school, in addition to providing trainings and mentorship to students.

conduct these trainings. In other cases, providers assist school resource officers in developing them.[109] These trainings are typically tailored for different age groups and often include content for parents and teachers. One provider said she has never been asked to address risks related to nudify apps, noting that the only AI-related concern schools have requested that she cover has been cheating with ChatGPT.[110]

Many teachers lack training on how to respond to a nudify incident at their school. The Center for Democracy and Technology report found that 62% of teachers say their school has not provided guidance on policies for handling incidents involving authentic or AI nonconsensual intimate imagery.[111] One provider told us that while many schools have crisis management plans for, for example, active shooter situations, they had never heard of a school having a crisis management plan for a nudify incident, or even for a real nude image of a student being circulated. A 2024 survey of teachers, principals, and district leaders found that 56% had not received any training on "AI deepfakes."[112]

By 2023, AI vendors were already coming up with products to sell to customers including schools, as demonstrated by one vendor's outreach to a school that had suffered a deepfake incident. (See Figure 4.1.) This vendor claimed to offer a "deepfake protection platform."

[redacted] and [redacted],

I hope you both have had a great week and holiday season thus far. I'm reaching out as I am a native [redacted], used to play [redacted] when I was in high school, and most importantly, I co-founded [redacted], a deepfake protection platform. After hearing about the incident that occurred at [redacted] I wanted to reach out to see if we'd be able to provide any assistance. Our company is in its very early stages but we do have a fully functional product that can deter adversaries from using real photos of students, teachers, etc., and deepfaking them. The best way to describe it is a photographic firewall.

We're looking to help anyone who needs it at this stage and therefore we'd like to work alongside you to help eliminate things like this from ever happening again.

Please feel free to reach out if you have any questions or schedule a call via my Calendly link below.

Best,

Figure 4.1: An email from a "deepfake protection platform" company to staff at a school that had experienced a nudify incident. Source: Public records requests.

### 4.1.3 Victims

There are many ways victims can be targeted. Offenders may use ordinary, fully clothed photos that children post on social media and upload them to nudify apps. But children don't need to be online to be at risk: offenders can snap photos of them in real life or use images shared online by others. As one platform employee put it, what keeps her up at night is knowing that even removing your child from

109. Interview with a law enforcement officer on Dec. 10, 2024.
110. Interview on Mar. 3, 2025.
111. Center for Democracy & Technology, *supra* note 98.
112. Olina Banerji, *Why Schools Need to Wake Up to the Threat of AI "Deepfakes" and Bullying*, EducationWeek (Dec. 9, 2024), https://www.edweek.org/technology/why-schools-need-to-wake-up-to-the-threat-of-ai-deepfakes-and-bullying/2024/12.

the internet doesn't guarantee safety: "We see people just taking a photo of a kid [...] without them knowing and creating this imagery."[113]

Though not the focus of this report, many respondents noted that students also create AI nude images of teachers, social workers, or other adults in their life, often aiming to make it look like the adult performs in pornography films.[114] Comments on a post in the Teachers subreddit also suggest this is a growing concern. One user wrote: "Here kids are making deep fakes with their teachers and administrators." Another wrote: "I'm scared of this. I'm just a sub, but I don't want students to make fake pictures of me because they're mad I told them to put their phones away."[115]

While victims may know what happened to them was wrong, they may not initially realize that a crime has occurred. A 2024 report by Thorn found that only 19% of U.S. youth aged 13 to 20 believed it was illegal to create AI-generated nude images of minors.[116] Neither of the two victims we interviewed initially understood that what happened to them was a crime. "No one really knew it was illegal at first, even me," one victim said.[117] The other noted, "Obviously I knew child pornography was illegal, but I didn't know how AI worked into that."[118]

### 4.1.4 Child offenders

Respondents described a spectrum of reasons for why children create AI nude images of their peers.[119] Some do it because they think it's funny and don't understand the consequences. Teens who create nudes of peers aren't necessarily "bad" kids deliberately trying to "ruin another kid's life," one provider said, but they lack the cognitive development to think long-term and anticipate the wider fallout of sharing the image with one other person.[120] Other children use it as a form of bullying; in more severe cases, some use the images to sextort victims, though this is more commonly seen among adult offenders.

One provider explained that children often view the physical and digital worlds as separate: "What happens face-to-face is one thing, which seems very visceral and real. But if it happens on a screen, there's a lack of sense of consequence, lack of inhibition. So using a [nudify app], it's all gas pedal, very little brake."[121]

Children may recognize that this activity is wrong but assume it's legal, believing that if it were truly illegal, there wouldn't be an app for it.[122] This misconception can be reinforced by misleading media coverage suggesting the behavior isn't

113. Interview on Jan. 14, 2025.
114. Interview with a former law enforcement officer on Jan. 9, 2025.
115. Comment on Disgruntled_Veteran, *Students Expelled for Fake Nudes*, Reddit, r/Teachers (2024), https://www.reddit.com/r/Teachers/comments/1b9pob4/students_expelled_for_fake_nudes/.
116. Thorn, *supra* note 97.
117. Interview on Feb. 5, 2025.
118. Interview on Feb. 11, 2025.
119. Interview with a U.S. government employee on Jan. 10, 2025.
120. Interview with a provider on Mar. 3, 2025.
121. Interview on Nov. 19, 2024.
122. Interview with a provider on Nov. 19, 2024.

against the law.[123] Even adults, such as school officials, parents, police, and district attorneys, may mistakenly believe the activity is legal, which could shape how children perceive it. That said, public misperceptions may shift now that numerous states have recently enacted laws expressly targeting the use of AI to morph children's images into CSAM.[124]

## 4.2 Prevention: Challenges in school-based prevention efforts

Our interviews suggest that few schools are currently providing instruction to students on the risks associated with nudify apps. While this may not be inherently problematic—values-based curricula may, for reasons outlined below, offer a more appropriate approach—there are nonetheless challenges in educating students about online exploitation that could be hindering understanding and awareness.

### 4.2.1 Discomfort with sex-related discussions

Many respondents told us that discussing online sexual exploitation is often perceived as awkward and sensitive. A law enforcement officer involved in school trainings noted that even school resource officers can feel uncomfortable addressing these topics.[125] This challenge is even more pronounced in some countries where any form of sex education is difficult to implement.[126] Interestingly, however, one NGO employee observed that the online dimension of child sexual abuse has, in some cases, created an entry point for broader conversations about abuse: "In some countries 'online' is the vehicle we use to talk about this."[127]

Many respondents emphasized the importance of introducing these topics before children are exposed to potential risks, an approach commonly taken in sex education more broadly.[128] One provider suggested that conversations about consent should begin "as soon as [children] can talk or even before."[129] In the context of online exploitation, this could mean introducing content as early as elementary school. One teacher shared that she's seen second graders with phones and access to social media, yet discussing these issues with that age group can be "super taboo."[130]

123. *See, e.g.*, Saleen Martin, *A Pennsylvania Boy Used AI to Make Nude Images of Female Students. Was It Illegal?*, USA TODAY (Nov. 21, 2024), https://www.usatoday.com/story/news/nation/2024/11/21/pennsylvania-parents-nude-photos-artificial-intelligence/76474857007/. (Two students were subsequently arrested, undermining this article's credulous premise.)
124. MultiState, *supra* note 5.
125. Interview on Dec. 10, 2024.
126. Interview with an NGO employee on July 16, 2024.
127. Interview on July 16, 2024.
128. Interview with a provider on Dec. 19, 2024.
129. Interview on Nov. 22, 2024.
130. Interview with a teacher on Jan. 6, 2025.

At the same time, one provider noted, "you'd be surprised at how many schools are actually ok with [having this content taught]."[131] We note, however, that providers may have a biased perspective, as they are working closely with schools that already recognize the importance of this topic. The provider added that there is no clear pattern to which types of schools are more or less receptive; even in conservative states, many districts are supportive of including this content.

One provider we spoke to offers a digital safety training that includes content on online sexual exploitation. He described a dynamic in which principals may not explicitly request training on online exploitation, but are aware of related issues among their students, such as sexting. His website outlines the training content, and he provides an overview during the booking process. He noted that this transparency is especially important for middle school sessions, allowing principals to make informed decisions about whether to include 6th graders.[132]

### 4.2.2 Risk of inadvertently educating children about the existence of nudify apps

There is a risk that educating children about the risks and harms of nudify apps could inadvertently make them aware of these apps. A report by Thorn supports this concern, finding that only 41% of Americans ages 13 to 20 had heard of AI-generated nude images.[133] Similarly, in our interviews, one provider noted that he got the sense many children had not heard of these apps.[134] As a result, some providers choose not to mention these apps explicitly in their trainings.[135] Others do include them as part of broader discussions about the legal consequences of obtaining and distributing CSAM.[136] When we asked a victim which approach she preferred, she emphasized the importance of being direct: "When we warn kids about meth are we telling kids about meth?" she asked rhetorically.[137] We do not take a position here on which approach is best, but highlight this as an area for future research.

This challenge isn't unique to nudify apps. One respondent referred to this phenomenon as "miseducating," explaining that some curricula aim to empower children by teaching them how to modify social media algorithms or news feeds. However, as the respondent pointed out, this knowledge could also be used to curate feeds in harmful or toxic ways.[138]

One could also imagine risks to teaching children about the consequences of AI-generated nude images. If minors are told that such images can ruin someone's career, they may weaponize that information against adults. Our interviews

131. Interview on Dec. 19, 2024.
132. Interview on Nov. 19, 2024.
133. Thorn, *supra* note 97.
134. Interview on Nov. 19, 2024.
135. Interview with a law enforcement officer on Dec. 10, 2024.
136. Interview on Dec. 19, 2024.
137. Interview on Feb. 5, 2025.
138. Interview on Dec. 19, 2024.

suggest this is not a hypothetical concern—students are already creating AI-generated nudes of teachers and other adults in their lives. This is not necessarily an argument against educating students on the topic, but it is a dynamic that educators should be aware of.

For this reason, many online safety curricula emphasize helping students reflect on concepts like consent and digital consent, their responsibilities toward others' images and likenesses, and their own personal values—empowering them to apply these principles to a wide range of scenarios that may not be explicitly addressed in class.[139]

### 4.2.3 Shortcomings in school programming on digital safety

Two academics—one based in the U.S. and one abroad—told us that schools often provide instruction on online exploitation only reactively, in response to an incident.[140] Similarly, a teacher shared that her school introduced a digital safety curriculum only after experiencing such an incident.[141]

One provider told us that no school has ever specifically requested training on nudify apps.[142] One respondent shared that she encouraged her child's school to offer this type of training, and while the school agreed, it scheduled a webinar six months later. She questioned why the training would be offered virtually rather than in person, and why it required such a long delay.[143] She speculated that schools may believe existing blocking technology on school-issued laptops is sufficient, and may not be considering the risks posed by students' personal phones.

Schools, of course, face their own set of challenges. In addition to delivering academic instruction, they must juggle a growing list of mandatory tasks, with each new state-imposed requirement competing for limited time in the school day. Even in states that mandate sex education, implementation may be limited due to a lack of enforcement.[144] Moreover, the landscape of online harms evolves rapidly, making it difficult for schools to keep related curricula current.[145]

Respondents expressed a general sense that digital safety trainings don't always resonate with students. This doesn't mean such trainings shouldn't be offered, as they may land with some students more than others,[146] but it does underscore the need for more research on their effectiveness.

Many respondents felt that schools might not be the ideal setting for this type of training. As one law enforcement officer put it, a school assembly at 9:00 a.m.

139. Interview with a provider on Dec. 19, 2024.
140. Interview on June 24, 2024; Interview on June 27, 2024.
141. Interview on Jan. 6, 2025.
142. Interview on Dec. 19, 2024.
143. Interview on June 24, 2024. For more on prevention-based education online, *see* Melissa A. Bright, Diana Ortega, David Finkelhor & Kerryann Walsh, *Moving School-Based CSA Prevention Education Online: Advantages and Challenges of the "New Normal,"* 132 Child Abuse & Neglect (2022).
144. Interview with a provider on Nov. 6, 2024.
145. Interview with a provider on Nov. 6, 2024.
146. Interview with a law enforcement officer on Nov. 13, 2024.

may not be the most effective time to teach online safety: "They aren't going to remember that at 9:00 p.m. when they have been gaming for three hours."[147] Our own opinion, however, is that while education from platforms and parents is important, school-based training may be necessary to ensure broad reach.

A provider highlighted the disconnect between a student's mindset during a school lesson and their mindset when using a phone: "When we talk about not sending a nude image to a stranger, they'll agree it sounds crazy. But under the right circumstances it's on the table for a lot of kids."[148] He explained that students often give the "right" answer in a "cerebral, cognitive moment," but that "kids aren't operating that way when they're using their devices." A teacher echoed this point, noting that students always seemed to know the correct responses. She described giving students a scenario in which an adult asks a child for a nude image: students would immediately identify the adult as a predator, say they would tell a trusted adult, and block the person. Still, she questioned how well those responses would translate to real-life decisions when students are actually on their phones and using social media.[149]

Interviews with individuals familiar with whether and how students learn about these issues revealed a general sense of pessimism. "We are dealing with the same issues today as we dealt with 12 years ago," one provider said, likely referring to life skills education. "It's just gotten worse. The 'we just need to educate our kids better, that'll fix it' argument [...] the schools can't do this [all alone]."[150]

### 4.2.4 Parental engagement

Respondents expressed several concerns about parents, who are uniformly seen as important actors in this space. Although providers often offer separate sessions for parents, attendance is typically low. Those who do attend are usually already practicing many of the recommended strategies. One provider noted that principals frequently say, "There are so many parents who should've been here tonight who really needed to hear this."[151] To improve accessibility, some providers offer parent training as a recorded webinar that can be viewed on demand. A former law enforcement officer shared that when he held student trainings during the day and parent sessions that evening, students would often say the content wasn't useful, leading parents to skip the training. To counter this, he began hosting the parent training the evening before the student session.[152]

A law enforcement officer described another challenge with parents: many struggle to understand that their child's phone can pose significant risks. As he put it, when they were growing up, the danger was perceived as someone with

147. Interview on Dec. 13, 2024.
148. Interview on Nov. 19, 2024.
149. Interview on Jan. 6, 2025.
150. Interview on Nov. 19, 2024.
151. Interview on Nov. 19, 2024.
152. Interview on Jan. 9, 2025.

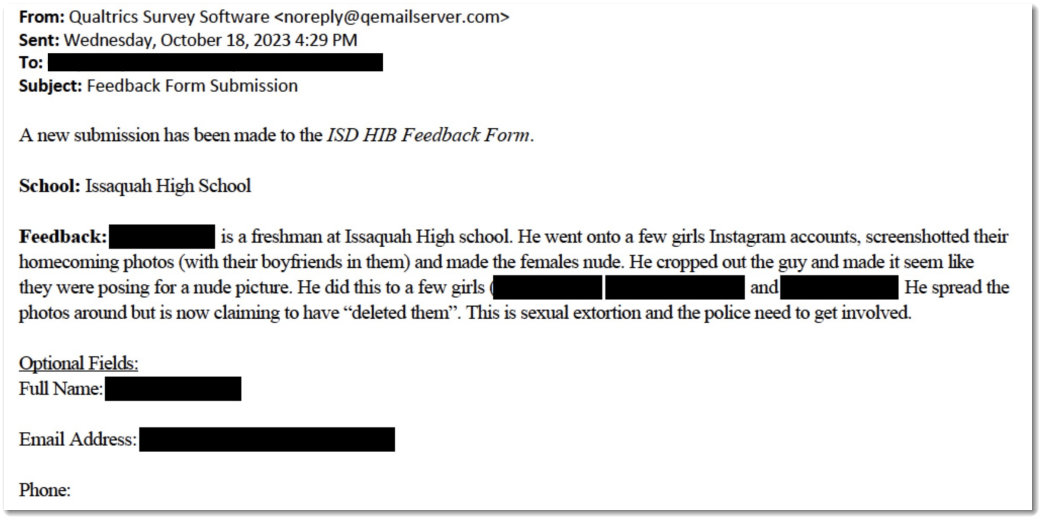
puppies trying to lure kids into a van. Today, the threat is often in their child's pocket.[153]

In some states, such as California, schools are required to notify parents about any sex education instruction.[154] While intended to keep families informed, this requirement could have a chilling effect, discouraging schools from expanding their curriculum out of concern that notifications might trigger parental backlash.

## 4.3 Response: Student reporting and school responses

We now briefly touch on some notable observations made related to how schools handle reports of students being targeted by nudify apps.

One school in Pennsylvania was first informed that a student was creating nude images of peers through the state's Safe2Say Something tipline.[155] Likewise, documents obtained via public records request (see Figure 4.2) show that a nudify incident at a Washington state school first came to school officials' attention through a submission to the school's harassment, intimidation, and bullying (HIB) online tipline.[156] In the tip, the submitter wrote, "This is sexual extortion and the police need to get involved."



**From:** Qualtrics Survey Software <noreply@qemailserver.com>
**Sent:** Wednesday, October 18, 2023 4:29 PM
**To:** [redacted]
**Subject:** Feedback Form Submission

A new submission has been made to the *ISD HIB Feedback Form*.

**School:** Issaquah High School

**Feedback:** [redacted] is a freshman at Issaquah High school. He went onto a few girls Instagram accounts, screenshotted their homecoming photos (with their boyfriends in them) and made the females nude. He cropped out the guy and made it seem like they were posing for a nude picture. He did this to a few girls ([redacted] [redacted] and [redacted] He spread the photos around but is now claiming to have "deleted them". This is sexual extortion and the police need to get involved.

Optional Fields:
Full Name: [redacted]

Email Address: [redacted]

Phone:

Figure 4.2: Report from the Issaquah High School harassment, intimidation, and bullying online tipline emailed to the school principal.

153. Interview on Dec. 10, 2024.
154. Interview with a provider on Nov. 6, 2024.
155. Ashley Stalnecker, *Here's What We Know and Don't Know About Lancaster Country Day AI-Generated Nude Image Incident*, LancasterOnline (Nov. 18, 2024), https://lancasteronline.com/news/local/heres-what-we-know-and-dont-know-about-lancaster-country-day-ai-generated-nude-image/article_f710874e-a5d1-11ef-82aa-cb325d028aaa.html.
156. "ISD HIB Feedback Form" email dated Oct. 18, 2023 to Issaquah High School staffer; email dated Oct. 20, 2023 from assistant principal of Issaquah High School, noting that "a HIB report was submitted from several students."

One provider emphasized the importance of such reporting tools, saying, "Anonymous reporting tools are one of the most important things we can have in our school systems," in part because many students lack a trusted adult they can turn to.[157] His organization emphasizes these tools in every training, across all audiences and contexts. However, he also acknowledged the potential for misuse, such as students submitting false reports to avoid school, borrowing the truism in trust and safety work: every anti-abuse tool can also be abused.

Victims or bystanders may not report nudify incidents if they do not trust that their tips will be taken seriously. In Pennsylvania, victims' families alleged that after the Safe2Say Something tip to the school, the school did not initially inform the police or take corrective action, providing an opportunity for more images to be created.[158] Feeling the school administration did not care about the victims, many students and some teachers staged a walkout in protest of the school's handling of the incident.[159]

Schools' failure to act on reports of nudify incidents can stem from ambiguity in state mandated reporter laws and/or school personnel's insufficient knowledge of state law, as described above in Section 2. While legislatures are starting to clarify state laws to address the AI CSAM context,[160] this is also an area where schools and school districts should preemptively set policies and train school personnel in order to avoid confusion. Training educators can also build trust with students: one victim told us she thinks ignorance of applicable law gave school personnel (whom she considered mandatory reporters) an excuse for not reporting images that qualified as CSAM.[161]

Beyond the question of "to report or not to report," schools often appear ill-equipped to support victims in nudify app incidents. According to NCMEC, educators (as well as parents and local law enforcement) frequently contact them for guidance, with some calls even coming from school resource officers.[162] While we must omit details to protect anonymity, one victim described receiving inappropriate responses from even well-intentioned teachers, saying those responses marked "the lowest point" in her experience.[163] Congressional testimony from a parent of a victim documented another school's missteps, including announcing the names of victims over the intercom. (See Figure 4.3.[164])

157. Interview on Dec. 19, 2024.
158. Ashley Stalnecker, *Parents to Sue Pennsylvania School District Over Deepfakes*, Gov. Tech. (Nov. 14, 2024), https://www.govtech.com/education/k-12/parents-to-sue-pennsylvania-school-district-over-deepfakes.
159. Ashley Stalnecker, *Lancaster Country Day Student Who Experienced "Deepest Violation" From AI Nudes Speaks Out*, LancasterOnline (Dec. 26, 2024), https://lancasteronline.com/news/local/lancaster-country-day-student-who-experienced-deepest-violation-from-ai-nudes-speaks-out/article_47823bb2-c3ca-11ef-9fd8-37be61648ba7.html.
160. MultiState, *supra* note 5.
161. Interview on Feb. 5, 2025.
162. Interview on Feb. 5, 2025.
163. Interview on Feb. 5, 2025.
164. *Addressing Real Harm Done by Deepfakes: Hearing Before the H. Oversight and Accountability Subcomm. on Cybersecurity, Information Technology, and Government Innovation*, 118th Congress (2023–2024) (statement of Dorota Mani), https://oversight.house.gov/wp-content/uploads/2024/03/Dorota-Mani-Testimony.pdf.

1. The school inappropriately announced the names of the female AI victims over the intercom, compromising their privacy.
2. The boys responsible for creating the nude photos were discreetly removed from class, their identities protected.
3. When my daughter sought the support of a counselor during a meeting with the vice principal who was questioning her, her request was denied.
4. The administration claimed the AI photographs were deleted without having seen them, offering no proof of their deletion.
5. My attempts to communicate with the administration about the case have been consistently ignored.
6. A Harassment, Intimidation, and Bullying (HIB) report submitted in November 2023 has yet to yield a conclusive outcome, which we should receive within 10 days of submission.
7. The interviews carried out at the school with underage suspects in the presence of police but without their parents have made their statements inadmissible in court.
8. Despite our submission of updated policies (created by our lawyers at McCarter and English) to the Westfield Board of Education, the school's cyber harassment policies remain outdated, referencing Walkmans and pagers with no mention of AI to this day.
9. The school's communications have focused on only one boy involved, ignoring others.
10. The accountability imposed for creating the AI Deep Fake nudes without girl's consent was a mere one-day suspension for only one boy.

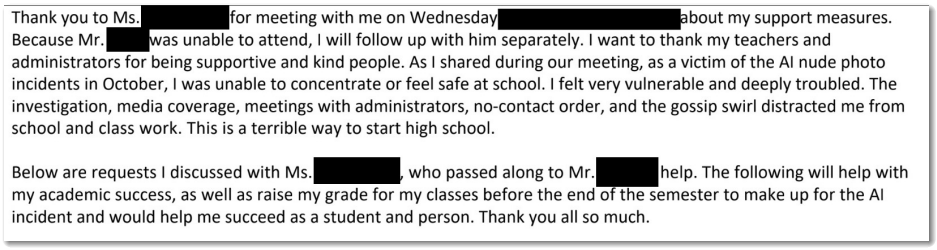
Figure 4.3: Congressional testimony by Dorota Mani documenting how her daughter's school responded to a nudify incident.

That said, handling of reporting obligations and handling of support for student victims are distinguishable issues facing schools. At a Washington state high school where multiple students and a female staffer were victimized by deepfake nudes, the school, incredibly, put that staffer in charge of the internal investigation. When she was later confronted by police about the school's failure to report to them, she believed the school was not mandated to report the fake images.[165] Yet when two victimized students told a male teacher what had happened, he offered them his classroom as a safe space and emailed the principal, an assistant principal, and the victims' counselors, saying "the actions here are very serious."[166] Also, as noted, some teachers at the Pennsylvania school joined students in a walkout, siding with the affected students instead of their staff colleagues. Additionally, a victim at one school sent an email to "thank my teachers and administrators for being supportive and kind people" and memorialize the accommodations she had requested. (See Figure 4.4.)[167] At least two of the recipients emailed the student back with supportive messages.[168]

These examples illustrate how staffers who work most closely with students (such as teachers and counselors) can maintain their crucial role as trusted adults in the wake of a deepfake incident, independent of how well or poorly the school executes its overall response (which may fall to higher-level administrators).

We asked many respondents about liability concerns, and three providers confirmed that schools are indeed worried about potential liability related to online

165. Koebler & Maiberg, *supra* note 28.
166. Email dated Oct. 20, 2023, from a teacher to the principal of Issaquah High School.
167. Source: Public records requests.
168. Source: Public records requests.



Thank you to Ms. [redacted] for meeting with me on Wednesday [redacted] about my support measures. Because Mr. [redacted] was unable to attend, I will follow up with him separately. I want to thank my teachers and administrators for being supportive and kind people. As I shared during our meeting, as a victim of the AI nude photo incidents in October, I was unable to concentrate or feel safe at school. I felt very vulnerable and deeply troubled. The investigation, media coverage, meetings with administrators, no-contact order, and the gossip swirl distracted me from school and class work. This is a terrible way to start high school.

Below are requests I discussed with Ms. [redacted], who passed along to Mr. [redacted] help. The following will help with my academic success, as well as raise my grade for my classes before the end of the semester to make up for the AI incident and would help me succeed as a student and person. Thank you all so much.

Figure 4.4: An email from a student to several teachers explaining the effect of the nudify incident and requesting accommodations. Source: Public records requests.

exploitation.[169] The Pennsylvania school, for example, has been sued by victims' families for its allegedly inadequate response to the deepfake incident there.[170] (The school then changed its enrollment contracts to let it dismiss students without a tuition refund if they sue the school.[171]) Schools often carry cybersecurity insurance, which typically covers incidents involving the circulation of explicit images on school-issued devices.[172] One provider noted that schools tend to respond to these situations by focusing on discipline rather than on supporting the victims.[173] Another observed that schools are primarily concerned with preventing harmful use of school-issued devices.[174] Since no school officials agreed to be interviewed, we were unable to hear directly from schools about how they view or handle these concerns.

One provider said he expects that, in the future, parents may be held liable for internet crimes committed by their children, much like how parents are increasingly being held accountable when their children gain access to firearms that are later used in violent incidents.[175] So far, we know of no cases where parents have been sued for their children's conduct in creating or disseminating deepfake nudes of other minors. One New Jersey victim has sued one of the alleged perpetrators in federal court, where they are both represented by their parents in accordance with court rules regarding minors.[176]

One provider noted that her group also considers its own liability.[177] She said she speaks with her insurance provider "all the time" to fully understand what is and isn't covered. The organization requires schools to sign an indemnification agreement and clearly outlines the content they will be delivering. They also offer an anonymous Q&A session and inform schools that they will answer questions to the best of their ability.

169. Interview on Nov. 6, 2024; Interview on Nov. 19, 2024; Interview on Nov. 22, 2024.
170. Stalnecker, *supra* note 31.
171. Ashley Stalnecker, *New Lancaster Country Day Enrollment Contract Says Students Can Be Dismissed If Family Sues School*, LancasterOnline (Feb. 14, 2025), https://lancasteronline.com/news/local/new-lancaster-country-day-enrollment-contract-says-students-can-be-dismissed-if-family-sues-school/article_57793f56-eb03-11ef-90d8-b72f75f55edf.html.
172. Interview on Nov. 22, 2024.
173. Interview on Nov. 6, 2024.
174. Interview on Nov. 19, 2024.
175. Interview on Nov. 19, 2024.
176. Doe v. Smith, No. 24-cv-634 (D.N.J. complaint filed Feb. 2, 2024); Fed. R. Civ. P. 17(c)(1).
177. Interview on Nov. 22, 2024.

## 4.4 Consequences: Should juvenile offenders face criminal prosecution?

One theme that emerged from our research is the lack of consensus about the appropriate consequences for children who make or circulate deepfake nudes of other children. While criminal penalties for CSAM are the routine response to adult offenders, the surprise emergence of juvenile AI CSAM offenders confounds that habitual impulse. The appropriate way for society to deal with these children remains an open question, to which our respondents' answers differed greatly.

As of May 2025, there have been relatively few stories in the news involving children using nudify apps. It remains to be seen whether these early high-profile incidents shape how the public and lawmakers view child offenders. Multiple respondents underlined that minors' brains are not yet fully developed, so they cannot fully reason through the potential consequences of their actions. If the cognitive development angle is emphasized, it looks inappropriate to treat juvenile perpetrators identically to adults who create such material. An employee of one provider (speaking personally, not for their employer) opposed criminalization, saying, "Young boys are still realizing this is a harm and they need education that what they did was a dire harm to the victim, but they're still children who shouldn't be put into jail."[178]

Empirically, it appears rare for children to face criminal consequences for nudifying images of their peers. To date, we know of only two cases: one where two teenage boys in Pennsylvania allegedly targeted dozens of girls at their school[179] and another where two Florida boys were accused of making imagery of children as young as 12.[180] There is a general perception that schools tend to handle these incidents internally, involving law enforcement only in exceptional cases.[181] According to a former law enforcement officer, deepfake incidents may not reach the police (at least, not until after they've been reported in the press) "because a lot of schools handle that in-house … because it's peer to peer not adult to child."[182] Sometimes parents have called the police after the schools did not do so.[183]

178. Interview with a provider on Nov. 25, 2024.

179. Mark Scolforo, *AI Photos of Student Faces on Nude Bodies Roil Private School in Lancaster Co.*, LancasterOnline (Nov. 20, 2024), https://www.nbcphiladelphia.com/news/local/ai-photos-of-students-faces-on-nude-bodies-roil-private-school-in-lancaster-co/4033167/; Lancaster Country District Attorney's Office, *2 Juveniles Charged in Connection to AI-Generated Images of Lancaster Country Day School Students; DA Determines School Officials Were Not Legally Required to Report AI Incidents* (Dec. 5, 2024), https://lancaster.crimewatchpa.com/da/11617/post/2-juveniles-charged-connection-ai-generated-images-lancaster-country-day-school-students; Elise Person, *Juveniles Charged for AI Images Stir Debate on Child Abuse Reporting Laws*, MSN (Mar. 25, 2025), https://www.msn.com/en-us/news/crime/juveniles-charged-for-ai-images-stir-debate-on-child-abuse-reporting-laws/ar-AA1BF1SQ.

180. Caroline Haskins, *Florida Middle Schoolers Arrested for Allegedly Creating Deepfake Nudes of Classmates*, Wired (Mar. 8, 2024), https://www.wired.com/story/florida-teens-arrested-deepfake-nudes-classmates/.

181. *E.g.*, interview on Dec. 20, 2024.

182. Interview on Dec. 20, 2024.

183. Meredith Jorgensen, *Parents of AI Nude Photo Victims at Lancaster Country Day School Call for Resignations*, WGAL (Nov. 12, 2024), https://www.wgal.com/article/lancaster-country-day-parents-of-ai-nude-photo-victims-resignations/62874684.

One law enforcement officer told us his department rarely sees these types of cases.[184] He could not recall a single case in his county where a minor had been charged for distributing a nude image of a peer—AI-generated or otherwise. "The county attorney won't charge these cases," he said: they have no jury appeal, and the prosecutor has no desire to charge an underage person.

Any criminal charges against children will likely arise under state law, because the federal government generally leaves juvenile justice up to the states.[185] (For this reason, it is unlikely that the new TAKE IT DOWN Act will be invoked against a child offender in federal court, as criminal enforcement is the law's exclusive remedy for publishing or threatening to publish "digital forgeries" of minors.[186]) A federal government employee we spoke to said that child-on-child nudify incidents are more sensibly left up to state systems, while questioning whether they belong in the criminal justice system at all.[187] The respondent posited a spectrum of "criminal blameworthiness" for child nudify offenders: from thinking it's funny, to meanspirited bullying without realizing how damaging it is, up through malicious at-scale image generation and using images for sextortion. "There are real policy questions there," the respondent said, of "whether severe consequences are appropriate for kids."

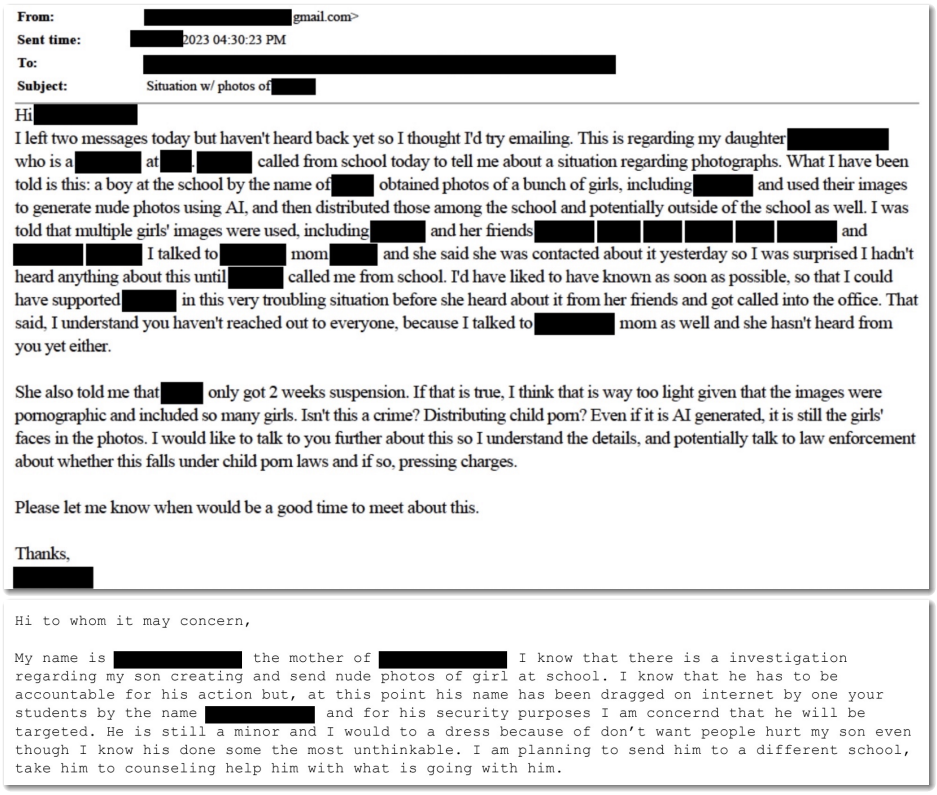
The raft of state legislation targeting deepfake nudes typically conceptualizes children only as victims, not as perpetrators. With rare exception,[188] recent state-level AI CSAM laws do not tend to distinguish between child and adult offenders. Growing awareness of student-on-student nudify incidents may prompt more targeted legislation aimed at that phenomenon. For example, a Texas bill introduced in January 2025 would add deepfake nudes to the state's definition of cyberbullying and allow public school students to be disciplined for releasing or threatening to release them.[189]

The state-level legislators and legislative staffers we interviewed disagreed over whether minors are fair game for criminal prosecution. Some weren't sure what the appropriate consequences should be for child offenders and conceded they hadn't really had that scenario in mind when crafting their bills.[190] One staffer characterized student-on-student incidents as cyberbullying,[191] whereas another called them "adult crimes," "not kids being kids," so "no matter who perpetrates it, they need to be punished for it."[192]

Perhaps unsurprisingly, victims' parents may not hesitate to paint offending children as criminals. One victim's parent told us that upon learning how many

184. Interview on Nov. 13, 2024.
185. Interview on Jan. 10, 2025; Charles Doyle, *Juvenile Delinquents and Federal Criminal Law: The Federal Juvenile Delinquency Act and Related Matters in Short*, Congr. Rsrch. Ser., abridged version of RL30822 (May 9, 2023), https://www.congress.gov/crs-product/R47548.
186. TAKE IT DOWN Act, Pub. L. No. 119-12, §§ 4(B), 6(B)(ii), 139 Stat. __ (2025).
187. Interview on Jan. 10, 2025.
188. *E.g.*, Act of Mar. 14, 2024, ch. 88, 2024 Wash. Laws (amending RCW 9.68A.055).
189. S.B. 747, 89th Leg. (Tex. 2025).
190. Interview with Florida State Senator Jennifer Bradley on Nov. 25, 2024; Interview with a state legislative aide on Dec. 9, 2024.
191. Interview with a state legislative aide on Jan. 17, 2025.
192. Interview with a state legislative aide on Jan. 6, 2025.

girls had been affected, the parent's perception went from "'it's an immature high school boy making bad decisions' to 'this is a sexual predator.'"[193] In another case, the parent of a victim wrote to school officials to express an interest in pressing criminal charges, complaining that the perpetrator's rumored two-week suspension was "way too light [a punishment] given that the images were pornographic and included so many girls. Isn't this a crime? Distributing child porn?"[194] (See top panel of Figure 4.5.) By contrast, an offender's parent offered to move her son to another school and get counseling for him; although he'd done something "unthinkable," she acknowledged, he "is still a minor" and she was worried about his own safety. (See bottom panel of Figure 4.5.)



**From:** [redacted]gmail.com>
**Sent time:** [redacted]2023 04:30:23 PM
**To:** [redacted]
**Subject:** Situation w/ photos of [redacted]

Hi [redacted]
I left two messages today but haven't heard back yet so I thought I'd try emailing. This is regarding my daughter [redacted] who is a [redacted] at [redacted] called from school today to tell me about a situation regarding photographs. What I have been told is this: a boy at the school by the name of [redacted] obtained photos of a bunch of girls, including [redacted] and used their images to generate nude photos using AI, and then distributed those among the school and potentially outside of the school as well. I was told that multiple girls' images were used, including [redacted] and her friends [redacted] and [redacted] I talked to [redacted] mom [redacted] and she said she was contacted about it yesterday so I was surprised I hadn't heard anything about this until [redacted] called me from school. I'd have liked to have known as soon as possible, so that I could have supported [redacted] in this very troubling situation before she heard about it from her friends and got called into the office. That said, I understand you haven't reached out to everyone, because I talked to [redacted] mom as well and she hasn't heard from you yet either.

She also told me that [redacted] only got 2 weeks suspension. If that is true, I think that is way too light given that the images were pornographic and included so many girls. Isn't this a crime? Distributing child porn? Even if it is AI generated, it is still the girls' faces in the photos. I would like to talk to you further about this so I understand the details, and potentially talk to law enforcement about whether this falls under child porn laws and if so, pressing charges.

Please let me know when would be a good time to meet about this.

Thanks,
[redacted]

Hi to whom it may concern,

My name is [redacted] the mother of [redacted] I know that there is a investigation regarding my son creating and send nude photos of girl at school. I know that he has to be accountable for his action but, at this point his name has been dragged on internet by one your students by the name [redacted] and for his security purposes I am concernd that he will be targeted. He is still a minor and I would to a dress because of don't want people hurt my son even though I know his done some the most unthinkable. I am planning to send him to a different school, take him to counseling help him with what is going with him.

Figure 4.5: Top: An email from a parent of a victim to the school. Bottom: An email from a parent of a child accused of creating an AI nude image of a peer to the school. These emails illustrate the range of challenges schools face in responding to a nudify incident. Source: Public records requests.

A "predator" framing harkens back to other historical efforts to paint juvenile offenders as monsters. For example, the media-fomented "superpredator" panic of the 1990s turned the full weight of the American criminal justice system against a generation of Black and brown youth.[195] A similarly harsh approach to juvenile offenders for deepfake nudes may likewise emerge if policymakers, media outlets, and the public take their cues from early cases, like those in Pennsylvania and

193. Interview with a victim's parent on Dec. 13, 2024.
194. Email from a victim's parent to personnel at a high school.
195. Carroll Bogert & LynNell Hancock, *Analysis: How the Media Created a "Superpredator" Myth That Harmed a Generation of Black Youth*, NBC News (Nov. 20, 2024), https://www.nbcnews.com/news/us-news/analysis-how-media-created-superpredator-myth-harmed-generation-black-youth-n1248101.

Florida, that are egregious either for the victims' large number or particularly young age. A harsh punitive response would also be unsurprising given that the United States (like many societies) views sexual crimes against children with particular horror, while simultaneously incarcerating children at a far higher rate than most countries.[196]

196. Richard Mendel, *Why Youth Incarceration Fails: An Updated Review of the Evidence*, The Sentencing Project (Mar. 1, 2023), https://www.sentencingproject.org/reports/why-youth-incarceration-fails-an-updated-review-of-the-evidence/.

# 5 Adults creating AI CSAM

## 5.1 Offender trends

Some themes in offender trends emerged from our interviews. First, several respondents said that AI-generated CSAM carries a sense of legitimization. While discussions about AI CSAM occur on the dark web, they are also prevalent in open clear web forums.[197] Individuals can commission custom AI CSAM using standard payment systems, making the content easily accessible.[198]

One respondent said that offenders hear from peers that AI CSAM is a "good" or acceptable way to consume this type of material, based on the notion that it doesn't harm a real child. This perception contributed further to a sense of legitimacy.[199]

One respondent highlighted the unique community element surrounding AI CSAM: "I've never known a community spirit like the one around AI CSAM. It's quite unsettling."[200] He described how, unlike the paranoia that surrounds the sharing of non-AI CSAM on the dark web, AI CSAM has fostered a collaborative environment: "This is a real community, people trying to figure stuff out together [...] almost slapping each other on the back, [saying] 'look how well [we did].'" Another respondent said this community element is "bringing everyone up to a high standard" in terms of AI CSAM creation.[201]

The community involved in creating AI CSAM is highly international, and because such content is not illegal in all countries, it adds another layer of confusion for individuals engaging with it.[202] In the U.S., some AI CSAM is constitutionally protected, but much of it is not (i.e., if it is obscene or depicts a real child).[203] This variability, coupled with variances in state-level CSAM law and shoddy press coverage that suggests making AI CSAM of real children isn't illegal,[204] further exacerbates the confusion.

197. Interview with a law enforcement officer on Dec. 13, 2024.
198. Interview with a law enforcement officer on Dec. 13, 2024.
199. Interview with a platform employee on Jan. 6, 2025.
200. Interview with an NGO employee on Dec. 13, 2024.
201. Interview with a law enforcement officer on Dec. 13, 2024.
202. Interview with a law enforcement officer on Dec. 13, 2024.
203. Pfefferkorn, *supra* note 10.
204. *E.g.*, Martin, *supra* note 123 (writing, "Are the images illegal? ... the images may not be grounds for charges," two weeks before the boy in question was charged with 60 criminal counts); Tate Ryan-Mosley, *A High School's Deepfake Porn Scandal is Pushing US Lawmakers into Action*, MIT Tech. Rev. (Dec. 1, 2023), https://www.technologyreview.com/2023/12/01/1084164/deepfake-porn-scandal-pushing-us-lawmakers/ (claiming that "the dearth of regulation and legal precedent on deepfake pornography means that victims like Francesca [Mani, a then-14-year-old girl in New Jersey] have little to no recourse," with no mention of the decades-old federal ban on morphed-image CSAM).

While some AI models can generate CSAM using text-only prompts, offenders are also using these tools to "refresh" existing CSAM to create bespoke images.[205] In particular, sets of CSAM images from decades ago—featuring the same victims—continue to circulate. Offenders are now using AI to insert those victims into new scenarios, including more violent sexual acts. One respondent noted that he expects offenders will soon begin creating AI-generated videos from these older images, further victimizing survivors.[206] Offenders are also uploading photos of children they know personally and attempting to use AI models to generate CSAM from them.[207] A platform employee said they are seeing this content both from bad-faith nudify apps and general-purpose AI image generation apps.[208]

Other trends we heard about include individuals offering AI CSAM creation as a paid service, including children selling AI-generated nude images of themselves.[209] Respondents noted that some offenders are creating increasingly extreme imagery,[210] an observation also made in a recent Internet Watch Foundation report.[211] One platform employee said that people uploading CSAM alongside an egregious prompt (that failed to produce new content) accounted for the vast majority of their CyberTipline reports with an AI element.[212]

We also heard reports of offenders using AI models to seek advice related to CSAM, including guidance on grooming and abusing children. According to NCMEC, offenders are also using AI services to ask questions about how to evade law enforcement, as well as how to locate CSAM. (Content warning: two upsetting examples are included in the footnote.[213])

Our interview with NCMEC staff provided insight into the range of CyberTipline reports involving AI.[214] These include:

- AI companies reporting attempts to upload known CSAM, with the model presumably blocking any output[215]
- AI companies reporting users entering problematic text prompts, with or without an image

205. Interview with an NGO employee on Dec. 13, 2024; Interview with NCMEC employees on Feb. 5, 2025.
206. Interview with an NGO employee on Dec. 13, 2024.
207. Interview with NCMEC employees on Feb. 5, 2025.
208. Interview on Jan. 14, 2025.
209. Interview with an NGO employee on Dec. 13, 2024.
210. Interview with a law enforcement officer on Dec. 13, 2024.
211. Internet Watch Foundation, *supra* note 1.
212. Interview on Feb. 24, 2025.
213. One respondent said they saw an offender asking an AI model for instructions on raping a two-year-old. Another said they saw prompts requesting detailed instructions on how to find, kidnap, rape, and kill a child without being caught.
214. Interview on Feb. 5, 2025.
215. For example, Google says that in 2024 "we reported more than 600 instances of apparent CSAM to NCMEC using hash-matching, which were uploaded as a user prompt to our generative AI products." Google, *Progress Update: Responsible AI and Child Sexual Abuse and Exploitation Online* (Apr. 2025), https://static.googleusercontent.com/media/publicpolicy.google/en//resources/ai_responsibility_and_csae_en.pdf.

- Social media platforms reporting the circulation or attempted circulation of AI CSAM, often apparently created through open-source AI models.

## 5.2 Prevalence

People mean different things when discussing the prevalence of AI CSAM. Prevalence could be measured in several ways: the number of individuals creating or viewing AI CSAM; the number of victims; the volume of images or videos produced; the number of CyberTipline reports that involve AI in any way; or the number of CyberTipline reports specifically about AI CSAM being successfully generated and shared.

NCMEC reports receiving 67,000 CyberTipline submissions with an AI component in 2024.[216] However, this figure does not equate to 67,000 instances of someone creating or sharing AI CSAM. As noted in Section 5.1, many of these reports involve individuals uploading known CSAM to an AI model, which then refuses to generate new content. It's unclear how many CyberTipline reports fall into that category versus those that actually contain AI CSAM. NCMEC may not always know whether a prompt generated CSAM; the CyberTipline report may include only a text prompt, but the output may be unclear to NCMEC.[217]

It's also important to note that a CyberTipline report can include multiple files. (In 2024 NCMEC received 20,512,803 million reports that contained 62,992,859 image, video, and other files.) Given that individuals can generate hundreds or even thousands of AI CSAM images, the average number of files per AI-related report may be high.[218]

There are other important caveats to keep in mind when interpreting this number. First, there may be additional CyberTipline reports involving AI-generated content that are not captured in this tally. This may be because the use of AI was not detected, or because the platform failed to indicate the presence of AI when submitting the report. In some cases, NCMEC may not have been able to identify the AI component themselves when reviewing the file, or, due to legal restrictions, may not have been able to review the file at all.[219] Second, while this figure represents a major increase from 4,700 AI-related reports in 2023[220]—an increase NCMEC attributes primarily to more AI companies beginning to submit reports

216. National Center for Missing and Exploited Children, *2024 CyberTipline Report* (2025), https://www.missingkids.org/content/dam/missingkids/pdfs/cybertiplinedata2024/2024-CyberTipline-Report.pdf.

217. Interview with NCMEC employees on Feb. 5, 2025.

218. According to the report for 2024 (*supra* note 216 at 5), platforms can now also bundle reports together, so one report can reflect hundreds or even thousands of instances of different users sharing one meme that technically meets the definition of CSAM but likely won't be prioritized for investigation.

219. Because several federal courts have ruled that NCMEC counts as a government entity or agent that is bound by the Fourth Amendment, NCMEC no longer opens files reported by platforms unless they had previously been opened by the platform. For an overview of those cases and their impact on NCMEC, *see* Section 3.2.1 of Grossman, Pfefferkorn, Thiel et al., *supra* note 47.

220. National Center for Missing and Exploited Children, *2023 CyberTipline Report* (2024), https://www.missingkids.org/content/dam/missingkids/pdfs/2023-CyberTipline-Report.pdf.

in 2024[221]—it remains a minuscule fraction of the total 20.5 million CyberTipline reports received in 2024.

NCMEC told us they are receiving some CyberTipline submissions involving content that appears to have been generated by children—for example, students using nudify apps to create nude images of their peers. However, the volume of these reports remains much lower than other types of submissions involving AI. These reports often originate not from online platforms, but directly from children or parents. A typical case might involve a boy generating dozens or even hundreds of manipulated images, which are then uploaded and circulated online.[222]

Most law enforcement officers and the staff from mainstream platforms we interviewed reported seeing relatively little AI CSAM. One platform employee said they had encountered it "a handful of times."[223] A law enforcement officer noted, "we don't see it a lot, but it trickles in."[224] Similarly, a survey by Thorn of professionals in the law enforcement sector found that, on average, they were "fielding CSAM cases involving generative AI" on average just 6.66% of the time.[225]

We asked many respondents why platforms and U.S. law enforcement aren't seeing a significant volume of AI-generated CSAM. Several theories emerged. First, this content may be circulating primarily on the dark web, and therefore not appearing in CyberTipline reports, which are the main mechanism through which U.S. law enforcement becomes aware of cases. While dark web material does occasionally reach law enforcement, there is often a delay between the offense and the start of an investigation.[226] Second, the technology is still relatively new, and many individuals who might be inclined to generate AI CSAM may simply be unaware of the tools available.[227] Third, AI CSAM often (though not always) comes to the attention of platforms when it is being shared. It's possible that in some cases, the valuable commodity is not the images themselves, but the prompting strategies used to generate them.[228]

That third explanation aligns with what we heard from one law enforcement officer conducting proactive investigations into AI CSAM, who told us the scale of content being created is high. "The volume is crazy," he said. "We are seeing offenders making half a million images on one device."[229] He added that you can imagine "a million images not shared at all, or one image shared a million times," and that AI CSAM falls into the former category. An NGO employee observed that AI CSAM appearing in CyberTipline reports is the "least concerning."[230] In

221. Interview on Feb. 5, 2025.
222. Interview on Feb. 5, 2025.
223. Interview on Dec. 16, 2024.
224. Interview on Nov. 13, 2024.
225. Thorn, *Evolving Technologies Horizon Scan* (Dec. 2024), https://info.thorn.org/hubfs/Research/Thorn_x_WPGA_EvolvingTechnologies_Dec2024.pdf.
226. Interview with a U.S. federal government employee on Jan. 10, 2025.
227. Interview with a law enforcement officer on Nov. 13, 2024.
228. Interview with a law enforcement officer on Dec. 13, 2024.
229. Interview on Dec. 13, 2024.
230. Interview on Dec. 13, 2024.

contrast, he said, the content uncovered through proactive investigations on the dark web and end-to-end encrypted platforms represents the "high end of high harm."

Even among those who currently view the prevalence of AI CSAM as low, there was broad agreement that it is likely to increase. A U.S. federal government employee noted, "the tech is there, the previous barriers to entry are no longer there."[231] A law enforcement officer expressed concern that neither his department nor society more broadly is prepared for what he sees as an inevitable rise in AI CSAM: "When you start talking about [AI CSAM] the community doesn't care."[232]

## 5.3 Photorealism

Some respondents report encountering photorealistic AI CSAM, while others said they are still seeing content with clear visual indicators of being AI-generated. These differences likely reflect the types of platforms respondents work for and the nature of the investigations being conducted by law enforcement. As with any technology, skill levels vary. However, all respondents who expressed an opinion on the matter agreed that AI CSAM is becoming increasingly photorealistic. One platform employee remarked, "We are still on the back foot in that we are looking at the imagery and asking, 'does it look messed up?' We won't be able to rely on that for much longer."[233]

A law enforcement officer who conducts proactive investigations into AI-generated CSAM on end-to-end encrypted platforms says the content he sees is realistic. He regularly encounters images he can identify as AI-generated only because of textual context; otherwise he would not be able to tell it's AI-generated: "The days of [AI-generated human images having] six fingers or the eyes don't look right [...] we have moved far beyond that."[234] He added that video is quickly reaching the same level of realism. An NGO employee we interviewed agreed, saying some AI-generated images are now indistinguishable from non-AI CSAM.[235]

One of the most disturbing things we heard in our research came during a conversation about the realism of AI-generated images. An NGO employee told us that some AI CSAM creators intentionally add a sixth finger to an image to make it look flawed, hoping law enforcement will dismiss it as obviously fake. "My big fear is an investigator sees that image and discounts it," he said.[236]

Many platform staff and law enforcement officers say they are seeing AI CSAM that is obviously AI-generated. One platform employee noted that AI models often struggle to accurately depict genitalia,[237] while another pointed to telltale

231. Interview on Jan. 10, 2025.
232. Interview on Nov. 13, 2024.
233. Interview on Jan. 24, 2025.
234. Interview on Dec. 13, 2024.
235. Interview on Dec. 20, 2024.
236. Interview on Dec. 13, 2024.
237. Interview on Jan. 7, 2025.

signs like extra appendages or unnatural body positions.[238] A law enforcement officer added that backgrounds can also reveal clues suggesting an image was AI-generated.[239] Of course, it's possible that some of the images they're seeing have been intentionally edited, or that they're encountering AI CSAM they mistake for real. But for now, most of the respondents we spoke with believe the AI CSAM they're seeing still contains obvious indicators of being synthetic.

Many respondents said that context and metadata are often helpful in assessing whether an image is AI-generated, though they rarely specified what that entails. We suspect this often refers to situations where individuals are seen discussing AI CSAM and then sharing an image.

Even when images appear to be clearly AI-generated, the process of making that determination is time-consuming. A former law enforcement officer noted that reviewing images from CyberTipline reports, search warrant returns, and public submissions—and assessing whether they're AI-generated—takes significant time and effort.[240]

A NCMEC employee noted that analysts often receive CyberTipline reports containing images of survivors whose identities are known; these are individuals who were depicted in widely circulated CSAM series. The images will show the survivor doing something not present in the known material. Because NCMEC analysts are deeply familiar with these series, they can quickly identify such images as AI-generated.[241]

The increasing photorealism of AI CSAM has also had an impact on child sextortion. Sextortion, using nude images for blackmail, originally involved real photos. Offenders would trick a child into sending an explicit image, then immediately switch to an extortion script, threatening to share the image with friends and family unless the child sent more images or money.[242] This tactic has been used by criminals in Côte d'Ivoire and Nigeria, often targeting children in the U.S. and Europe for financial gain. Traditionally, sextortion has been time-consuming, requiring offenders to invest significant effort in building an online relationship to obtain the initial image.

AI has transformed sextortion, making it far more efficient. Offenders no longer need an initial real image from the child; the first contact can be the threat itself. They can take a clothed image of the child found online, use AI to generate a realistic-looking nude, and then threaten to share it with friends and family unless the child sends money. NCMEC says they see this tactic used repeatedly.[243]

A law enforcement officer said he is increasingly seeing AI nude images used to blackmail businesses—for example, fabricating explicit images of teachers and students at dance academies. The traditional model of sextortion involving

238. Interview on Dec. 16, 2024.
239. Interview on Nov. 13, 2024.
240. Interview on Jan. 6, 2024.
241. Interview on Feb. 5, 2025.
242. *See, e.g.*, the recent indictment of an accused sextortion offender in North Dakota federal court. United States v. Cherif, No. 25-cr-11 (D.N.D. filed Jan. 22, 2025), https://storage.courtlistener.com/recap/gov.uscourts.ndd.66865/gov.uscourts.ndd.66865.2.0.pdf.
243. Interview on Feb. 5, 2025.

perpetrators in Africa has evolved, he noted. "That's not the model that we are seeing. It's U.K., U.S., Australia offenders."[244]

## 5.4 Law enforcement challenges

We heard about two key challenges law enforcement faces when dealing with AI CSAM. First, officers said they lack reliable tools to definitively determine whether media is AI-generated. While such technology may exist, it's often expensive, and securing a budget for tools specific to child exploitation cases can be difficult. Additionally, vendors may lack the incentive to build tools tailored to this space. One former law enforcement officer noted that the primary market for AI CSAM detection tools would be the 61 ICAC Task Forces—too small a customer base to motivate most vendors. She added that vendors "act enthusiastic but never do it."[245]

An NGO employee highlighted another emerging challenge: victim identification officers are now spending more time analyzing CSAM to determine whether it's AI-generated, which may be harmful for them. They have to "study the image that much more closely, be more aware of color, shapes, what is actually happening in the image," he said.[246] There's a "difference between clicking through and having to really look at it." A law enforcement officer echoed this concern, noting that AI CSAM often appears more extreme than real material, and is more likely to depict torture. "Even if you know it's not real, the brain is impacted," he said, especially when viewing so many of these images.[247]

## 5.5 Platform observations

In this section we discuss observations from platform employees about finding and reporting AI CSAM. A range of companies may encounter this content. These include those that build generative AI models (sometimes called "model builders"), platforms that offer AI tools, and platforms where AI-generated content can be shared. Some companies fall into more than one of these categories.

Model-building companies exist on a spectrum. On one end are mainstream companies that either don't offer image generation tools or have policies prohibiting the creation of any nude content, or CSAM explicitly. On the other end are companies that appear indifferent to whether their tools are used to produce CSAM. A platform employee noted that for companies in the middle, there's often a side community dedicated to bypassing existing safeguards.[248]

244. Interview on Dec. 13, 2024.
245. Interview on Jan. 6, 2025.
246. Interview on Dec. 13, 2024.
247. Interview on Dec. 13, 2024.
248. Interview on Jan. 27, 2025.

### 5.5.1 Finding it

Detection can be broken down into two questions: first, whether AI CSAM is being flagged as CSAM at all, and second, whether detection systems are identifying it as AI-generated.[249] On the first point, a platform employee said that existing machine learning classifiers show "decent performance" in detecting AI CSAM.[250] An NGO employee added that these classifiers perform about as well on AI CSAM as they do on traditional CSAM, though results vary by tool and tend to decline when the AI content is less photorealistic.[251] On the second point, our sense is that most platforms are not prioritizing labeling CSAM as AI-generated for CyberTipline reports. We are unclear on whether they are satisfied with the effectiveness of tools for this specific task (and, by extension, whether ineffective tools lead to deprioritization of the task). One respondent noted that general AI-detection models are often sufficient to determine whether CSAM is AI-generated—meaning they don't necessarily need to be trained on CSAM specifically to be effective.[252]

Context clues can be useful in identifying AI-generated CSAM—for example, the name of the platform space where users are interacting or discussions around specific prompts.[253] One platform employee noted that Lantern, a signal-sharing initiative coordinated by the Tech Coalition, has been valuable: "[it's] less 'hey we found this particular image' and more 'we found this app and it's doing X.'"[254] NCMEC also shares information with platforms about prompts they're seeing, which one platform employee described as helpful.[255] Like law enforcement, platforms are also uncovering AI CSAM in part through investigations into individuals creating or sharing non-AI CSAM.[256]

While some platforms said they try to indicate in CyberTipline reports whether an image appears AI-generated—knowing this information is helpful to NCMEC and law enforcement—many conveyed, sometimes indirectly, that they aren't doing this systematically. One platform employee explained that regardless of whether CSAM is AI-generated or not, their response remains the same: they take action on the content, the device, and the associated account, and report it to NCMEC.[257] "The gumming up will happen further downstream," the respondent said, referring to the challenges NCMEC or law enforcement face in determining whether an image depicts a real child. Another platform employee noted that even with significant effort, it remains extremely difficult to determine whether AI-generated CSAM is entirely synthetic or based on the likeness of a real child.[258] One employee from a company that does include AI labels said the process of determining whether CSAM is AI-generated is still manual.[259]

249. Interview with an NGO employee on Feb. 12, 2025.
250. Interview on Jan. 6, 2025.
251. Interview on Feb. 12, 2025.
252. Interview with an NGO employee on Feb. 12, 2025.
253. Interview with a platform employee on Jan. 14, 2025.
254. Interview on Jan. 15, 2025.
255. Interview on Jan. 6, 2025.
256. Interview with a platform employee on Jan. 6, 2025.
257. Interview on Dec. 16, 2024.
258. Interview on Dec. 16, 2024.
259. Interview on Jan. 7, 2025; Interview on Jan. 14, 2025.

NCMEC confirmed that they receive many AI-generated files that are not labeled as such in CyberTipline reports. NCMEC makes this determination based on metadata, contextual clues from chats, or familiarity with the depicted victim.[260]

In our 2024 report on challenges within the online child safety ecosystem, we observed a recurring "hot potato" dynamic—where responsibility is continually passed along the chain until it ultimately lands with law enforcement.[261] There, officers have an overwhelming backlog of CyberTipline reports and insufficient time or resources to address them all. While some stakeholders could clearly be doing more, in other cases, even well-intentioned actors making reasonable decisions still end up passing the buck to law enforcement.

The issue of labeling CSAM as AI-generated feels similar. If platforms are already using classifiers to assess whether content is likely synthetic, developing an automated way to reliably include that information in the CyberTipline reports seems like a logical next step. However, there are valid reasons why platforms might hesitate to make or convey this assessment in a report. For instance, they may not want to require moderators to spend additional time closely analyzing deeply disturbing images or videos. Or, where a moderator recognizes that an AI-generated image depicts a known survivor from an existing CSAM series, the moderator might find the "Generative AI" label inapt since there is a real child depicted.

Platforms may also fear incorrectly labeling photographic CSAM as AI-generated—especially given that, as we noted in our earlier report, stakeholders are often incentivized to act with extreme caution. Our previous report on the CyberTipline documented platforms' hesitancy to include interpretations or "hunches" in their reports, for fear that inaccuracy could potentially render the report inadmissible in court, make the case more difficult to prosecute, or open up the platform to some sort of liability for getting it wrong.[262] These same concerns are equally applicable to mistakenly labeling real content as AI-generated.

If no human at the platform has reviewed the content, NCMEC may also be unable to view the files and therefore unable to assess whether it was AI-generated.[263] As a result, law enforcement is often left with the responsibility of making this determination, despite lacking the specialized tools and resources needed to do so. And increasingly, this is a judgment that may no longer be possible for a human to make without technical assistance, unless it is facially obvious that content is AI-generated (for example, because the depictions are fantastical or physically impossible).

While some respondents noted that existing CSAM hash lists already include some AI-generated content,[264] others said it would be useful for NCMEC to maintain a dedicated hash set specifically for AI CSAM—something NCMEC is currently

260. Interview on Feb. 5, 2025.
261. Grossman, Pfefferkorn, Thiel et al., *supra* note 47.
262. *See* Grossman, Pfefferkorn, Thiel et al., *supra* note 47, at 57.
263. As noted above, due to federal court rulings that NCMEC must follow the Fourth Amendment, NCMEC does not open reported files unless the platform had opened the file previously, as detailed in Section 3.2.1 of Grossman, Pfefferkorn, Thiel et al., *supra* note 47.
264. Interview with a platform employee on Jan. 6, 2025.

working on.[265] However, some questioned the scalability of such an approach, given that thousands of AI-generated images can be produced in a matter of minutes.[266]

### 5.5.2 Report it?

This section focuses on how platforms report AI CSAM to NCMEC. While law enforcement may also report such content,[267] that was not the primary focus of our interviews.

All platform respondents we spoke with said they report AI CSAM to NCMEC when they see it.[268] (NCMEC believes that AI CSAM should be reported to the CyberTipline.[269]) Many respondents cited a 2024 FBI notice commonly referred to as the "AI CSAM is CSAM" notice,[270] titled "Child Sexual Abuse Material Created by Generative AI and Similar Online Tools is Illegal." An NGO employee said her impression is that "everybody's reporting [AI CSAM]."[271] ("Everybody" is likely a reference to mainstream U.S.-based platforms.) A platform employee echoed this, saying, "All companies are reporting everything to NCMEC for fear of missing something."[272]

This approach reflects what one respondent called a mitigation point of view: "even if it's not clear cut" whether particular material legally must be reported, you won't get in trouble for reporting it, and the current volume of AI CSAM remains relatively low, so it's safer to err on the side of caution and report.[273] After that interview, however, a federal court ruled in February 2025 that Verizon and its cloud storage provider, Synchronoss, had to face a subscriber's lawsuit for reporting two of his files to NCMEC that turned out not to be CSAM. If that decision stands, platforms may rethink their "it's safer to report it all" stance.[274]

NCMEC told us that significantly more platforms reported content with an AI component in 2024 compared to 2023. In 2023, most companies submitting such reports already had established relationships with NCMEC. In 2024, in part due to NCMEC's outreach (including onboarding companies to known CSAM hash lists so they can detect if known material is being uploaded to their AI services),

265. Interview on Feb. 5, 2025.
266. Interview with an NGO employee on Feb. 12, 2025.
267. Interview with NCMEC employees on Feb. 5, 2025.
268. As our CyberTipline report noted, the field in the CyberTipline reporting form is called "Generative AI," but NCMEC's documentation for the CyberTipline API defines the field as "The file contains content that is believed to be Generative Artificial Intelligence." Grossman, Pfefferkorn, Thiel et al., *supra* note 47, at 76 n.405.
269. Interview on Feb. 5, 2025.
270. Fed. Bureau of Investigation, *supra* note 38.
271. Interview on Jan. 13, 2025.
272. Interview on Dec. 16, 2024.
273. Interview on Jan. 27, 2025.
274. Riana Pfefferkorn, *Verizon and Its Cloud Vendor Must Face Lawsuit for Reporting "CSAM" That Wasn't – Lawshe v. Verizon (Guest Blog Post)*, Tech. & Mktg. L. Blog (Mar. 11, 2025), https://blog.ericgoldman.org/archives/2025/03/verizon-and-its-cloud-vendor-must-face-lawsuit-for-reporting-csam-that-wasnt-lawshe-v-verizon-guest-blog-post.htm. Verizon's motion for reconsideration was still pending as of May 27, 2025.

"two or three dozen companies" are now reporting content with an AI element.[275] NCMEC's list of providers that submitted CyberTipline reports in 2024[276] contains multiple AI model-building and hosting platforms, including Runway AI, Stability AI, and Civitai, that were not in the 2023 list.[277]

One platform employee told us they use a reporting template created by the Tech Coalition to submit CyberTipline reports involving AI-generated content. This includes content from red teaming (discussed further in Section 7), material found in AI training datasets, and user-generated content.[278] The template is entered into the "additional information" section of the CyberTipline form and helps standardize how AI-related data is presented, making it more digestible for NCMEC analysts.

For model-building companies, the template includes fields to describe how their model works and what it's used for. Additional fields cover input and output content, solicitation details (e.g., whether someone asked another person to generate the content), and whether either the input or output is based on a real person.

One respondent from an AI company highlighted an additional challenge: many of the company's CyberTipline reports stem from API usage.[279] Because many API customers do not collect user data, the respondent acknowledged that these reports are likely of limited use to NCMEC or law enforcement. They noted they have been considering ways to make these reports more actionable.

Platform employees told us they crave more transparency about what happens to CyberTipline reports that contain an AI element. One model-building company employee said the company rarely hears back from law enforcement about any of their CyberTipline reports.[280] One platform questioned whether reports that are entirely synthetic are even forwarded to law enforcement[281] (they are[282]). Another employee wondered whether tagging a report as AI-generated meant it was treated as informational only, and questioned how law enforcement responds to such reports: "How is the whole funnel working together to treat this?"[283]

While NCMEC believes CSAM-related text prompts to AI models that fail to produce an image or video are useful to report to the CyberTipline,[284] not all platforms are aware of this, agree with it, or follow it in practice. One platform employee acknowledged the value of reporting prompts but questioned whether

275. Interview on Feb. 5, 2025.
276. National Center for Missing and Exploited Children, *2024 CyberTipline Reports by Electronic Service Providers (ESPs)* (2025), https://www.missingkids.org/content/dam/missingkids/pdfs/cybertiplinedata2024/2024-reports-by-esp.pdf.
277. National Center for Missing and Exploited Children, *2023 CyberTipline Reports by Electronic Service Providers (ESP)* (2024), https://www.missingkids.org/content/dam/missingkids/pdfs/2023-reports-by-esp.pdf.
278. Interview on Jan. 6, 2025.
279. Interview on Feb. 25, 2025.
280. Interview on Feb. 25, 2025.
281. Interview on Dec. 16, 2025.
282. Interview with NCMEC employees on Feb. 5, 2025.
283. Interview on Jan. 14, 2025.
284. Interview on Feb. 5, 2025.

they meet the threshold for reporting: "That's kind of like [...] reporting just search queries, which probably doesn't meet the definition of what NCMEC or law enforcement wants to see."[285] An NGO employee noted that, in recent months, there has been growing debate over whether egregious prompts that don't result in imagery should be reported.[286]

One respondent noted that as long as platforms "keep playing the game of detection, we will always be behind."[287] She emphasized the importance of the "whole long AI value chain" investing in preventing AI CSAM generation capabilities.

### AI sexually suggestive images of children that are not CSAM

Both platform staff and law enforcement reported a rise in AI-generated, sexually suggestive images of children that do not meet the legal threshold for CSAM,[a] content that could benefit from a shorter term or acronym.[b] While one platform employee described AI CSAM as "few and far between," she said that this type of borderline content is increasingly common.[c]

Platforms appear to take different approaches to handling AI-generated, sexually suggestive content that doesn't meet the legal definition of CSAM. From our perspective, there are both advantages and drawbacks to reporting this content to the CyberTipline.

One platform that has reported this content to NCMEC continues to question whether it's the right decision. The respondent noted that moderation teams think it should be reported out of duty of care, but she expressed concern that doing so might divert law enforcement resources from protecting real minors.[d]

Another platform takes a different approach: they remove the content and ban users who upload it but do not report it to NCMEC. "We've debated internally whether to send those reports anyways because we think morally that's the right thing," the respondent said. But they also want to make sure NCMEC and law enforcement can focus on reports of actual CSAM. She said this remains an ongoing internal discussion.[e]

A third platform described a hybrid approach: they only report this type of content to the CyberTipline if it's accompanied by text suggesting the user may be a "bad actor."[f] NCMEC, for its part, supports reporting this content, noting that it could help inform future hash list development.[g]

*a*. Interview with a law enforcement officer on Nov. 13, 2024; Interview with a platform employee on Jan. 14, 2025.

*b*. While the term "child erotica" has been used, this language has been criticized. Mary G. Leary, *Death to Child Erotica: How Mislabeling the Evidence Can Risk Inaccuracy in the Courtroom*, 16 Cardozo J.L. & Gender 1 (2009).

*c*. Interview on Jan. 7, 2025.

*d*. Interview on Jan. 7, 2025.

*e*. Interview on Jan. 14, 2025.

*f*. Interview on Feb. 25, 2025.

*g*. Interview on Feb. 5, 2025.

285. Interview on Jan. 14, 2025.

286. Interview on Jan. 13, 2025.

287. Interview with an NGO employee on Feb. 12, 2025.

# 6 Law enforcement and platform staff wellbeing

AI CSAM has welfare implications for the law enforcement officers, platform staff, and NCMEC staff who view it. One platform employee described AI CSAM and other AI-generated media as "nightmarescape" content, a word that stuck with us.[288] Moderation teams were already exposed to disturbing material, she said, but generative AI has pushed the boundaries: "[it's] the worst imaginable. [...] It's images out of nightmares now, and they're hyperrealistic—combinations of things you wouldn't see in real life."[289]

A former law enforcement officer echoed this, saying AI CSAM is often more violent than non-AI CSAM.[290] Another platform employee, referencing third-party reports, noted that "[the] level of egregiousness of the content is higher" than in non-AI CSAM.[291] A former law enforcement officer noted that even if you know it's not a real child, "it still messes with your mind."[292]

Both a platform employee and a NCMEC employee emphasized the emotional toll of encountering AI CSAM based on real CSAM of known survivors. (To say nothing, of course, of the emotional toll on the survivor themself.) NCMEC shared that notifying a survivor about a newly discovered AI-generated image of them is often deeply distressing for their team.[293] Similarly, a platform employee said it can be especially difficult for reviewers to encounter new AI-generated content of survivors they recognize, as it makes one wonder "can we ever stop this? What's the point of doing this work" when there are potentially infinite permutations. This sense of futility can dampen morale.[294]

288. Interview on Jan. 14, 2025.
289. Interview on Jan. 15, 2025.
290. Interview on Jan. 6, 2025.
291. Interview on Jan. 6, 2025.
292. Interview on Jan. 6, 2025.
293. Interview on Feb. 5, 2025.
294. Interview on Jan. 6, 2025.

# 7 Red teaming for AI CSAM

In our interviews, we found that there exists a regulatory vacuum around red teaming for AI CSAM, which we conclude is one of the most pressing outstanding areas for policymakers to address.

Red teaming involves simulating the behavior of an adversarial actor to test the effectiveness of security systems. In the context of child safety, an AI model-building company might red team a text model to assess whether its safety guardrails effectively block grooming-related messages or fictional depictions of child exploitation. In the case of CSAM, a company might attempt to get a text-to-image model to generate such content specifically in order to identify and address vulnerabilities in the model's safeguards.

Red teaming for CSAM differs significantly from red teaming for violent extremism, adult nudity, or even child grooming content, because federal law prohibits the knowing production, receipt, and possession of CSAM (as well as certain forms of solicitation), regardless of intent.[295] There is no legal exception for research or testing purposes. Red teaming for CSAM is thus very risky legally.

That means AI companies face a dilemma: "There is a need for companies to build their systems in a smart, informed way," as one respondent said,[296] but if you red team your model for CSAM, you make it safer at the cost of your own potential criminal liability. If you don't red team your model for its abusive potential, you avoid one source of legal exposure but risk other fallout (reputational, business, legal) should your model become known for allowing AI CSAM generation.

The surest mitigation of legal risk (at least for U.S.-based red teamers) would be for Congress to amend the federal CSAM laws to explicitly immunize good-faith AI red teaming for CSAM from liability (state or federal, criminal or civil). While there have been calls for a statutory safe harbor for AI red teaming,[297] establishing a well-regulated framework for good-faith red teaming around CSAM specifically is complex. One respondent explained: "The companies [...] don't want to take on criminal legal risk and also the government doesn't want careless red teamers out there" whose conduct leads to the distribution of AI CSAM created through red teaming.[298]

295. 18 U.S.C. §§ 2252, 2252A.

296. Interview on Jan. 27, 2025.

297. Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, Yi Zeng, Weiyan Shi, Xianjun Yang, Reid Southen, Alexander Robey, Patrick Chao, Diyi Yang, Ruoxi Jia, Daniel Kang, Sandy Pentland, Arvind Narayanan, Percy Liang & Peter Henderson. *A Safe Harbor for AI Evaluation and Red Teaming*, arXiv: 2403.04893 [cs.AI], March 2024a. doi:10.48550/arXiv.2403.04893.

298. Interview on Jan. 10, 2025.

To our knowledge, the federal government has not issued a public-facing policy on this issue. One respondent said that efforts to reach an agreement with the Biden administration "came close," but with the transition to a new presidential administration, progress has stalled and the future remains uncertain. (The new administration immediately repealed[299] a Biden-era executive order [EO] on AI that had made clear that red teaming is a public good[300] and recognized the need for "testing and safeguards" to prevent the production of AI CSAM.[301]) The respondent also acknowledged that private sector red teaming for CSAM is fraught, as this could open the door for bad actors to justify their actions by simply claiming "I'm just red teaming."[302]

For AI companies that choose to directly red team for CSAM—i.e., attempt to directly generate CSAM—the legal risks are real. One academic noted that in the hacking context, intent can matter, but "it's very fragile and relies on others believing it, and it's hard to prove either way when your conduct is consistent with an attacker's."[303] Red teaming for CSAM is even riskier, he said, because CSAM is contraband.[304] For example, if a red teaming exercise involved uploading CSAM to test whether a model could detect it, that action alone could constitute illegal handling of contraband.[305] From a company's standpoint, he added, avoiding this kind of red teaming is an obvious decision to reduce legal exposure.

At first glance, we were skeptical that an AI company would face prosecution for red teaming involving CSAM. But an academic we spoke with explained that when it comes to criminal liability, "enforceability likelihood isn't much of a factor" in how companies assess risk, unless the law in question is so rarely applied that "even a prosecutor would roll their eyes at" the notion.[306] That is not the case with the federal CSAM statutes. To protect themselves, the academic said, companies might need statutory changes or at least a "comfort letter" from the Department

299. The White House, *Removing Barriers to American Leadership in Artificial Intelligence*, Jan. 23, 2025, https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/.

300. Executive Order 14110 of October 30, 2023, 3 C.F.R. 657 (2024) § 4.1(ii), https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence (requiring multi-agency collaboration to "[e]stablish appropriate guidelines (except for AI used as a component of a national security system), including appropriate procedures and processes, to enable developers of AI, especially of dual-use foundation models, to conduct AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems"); *id*. § 10.1(b)(viii)(A) (requiring guidance on "external testing for AI, including AI red-teaming for generative AI").

301. *Id*. § 10.1(b)(viii)(B) (requiring guidance on "testing and safeguards against … producing child sexual abuse material and against producing non-consensual intimate imagery of real individuals …, for generative AI").

302. We are anonymizing the interview date and respondent type for this respondent's observations.

303. Interview on Jan. 27, 2025.

304. Interview on Jan. 27, 2025.

305. Interview on Jan. 27, 2025.

306. Interview on Jan. 27, 2025.

of Justice (though we note that this would not tie state prosecutors' hands[307]). A conservative corporate approach to legal risk where CSAM is involved is consistent with what we heard about platforms' legal departments while researching our CyberTipline paper.[308]

Companies that engage in red teaming for AI-generated CSAM generally have two options: conducting the work in-house or outsourcing it. An NGO employee told us that around 2023, many companies opted to outsource this work, but more recently, there has been a shift toward bringing it in-house.[309] We contacted several companies that offer red teaming services to AI firms, but none were willing to participate in our research.

Understanding whether and how these companies approach child safety red teaming is important. Keeping mum about this topic may be legally safer, but could also hinder the development of best practices for this unique red teaming context.[310] We believe this would be a valuable area for future research, especially by those with better access to this population. If these vendors are handling CSEA-related material, it would be important to assess whether they follow best practices for staff well-being, such as those implemented at large platforms and within law enforcement agencies.

So what are AI companies actually doing? Some companies likely make no effort to red team for CSAM while others probably red team directly for CSAM and count on the improbability of being prosecuted.[311] Some companies prohibit any generations of nudity, making it easier to prohibit CSAM.[312] One platform employee told us that red teaming AI models for CSAM "is not possible without a new agreement from the [U.S. federal] government."[313]

One platform employee said their team follows a red teaming approach introduced by Thorn: testing for adult sexual content and clothed images of children, with the assumption that if a model can generate both, it could likely be used to produce CSAM.[314] (See Figure 7.1.)[315] "It's what we can do. It bothers all of us here," the respondent said, expressing frustration about the limitations of this approach. She also noted the occurrence of what she colloquially called "retroactive red teaming" after a model is released, driven by "users trying to do horrible things."

307. And vice versa: A state-level law could not tie the federal government's hands, even if states make red teaming an affirmative defense to liability as Colorado has done in its recent AI Act, or establish "sandbox programs" for AI testing as Texas has proposed. Colorado Artificial Intelligence Act, Colo. Rev. Stat. 6-1-1706(3)(a)(II) (2024); Texas Responsible A.I. Governance Act, H.B. 149, 89th Leg. (Tex. 2025). And that is assuming such provisions even cover CSAM red teaming specifically, as a new Arkansas law does. Act of Apr. 22, 2025, § 3, 2025 Ark. Acts (Act 977).

308. *E.g.*, Grossman, Pfefferkorn, Thiel et al., *supra* note 47, at 32–33, 36, 49, 57.

309. Interview on Feb. 12, 2025.

310. For example, best practices for responsible vulnerability disclosure in the cybersecurity context do not all carry over cleanly to the CSAM context. David Thiel (@det@hachyderm.io), thread on Mastodon (Aug. 30, 2024), https://hachyderm.io/@det/113052306541710890.

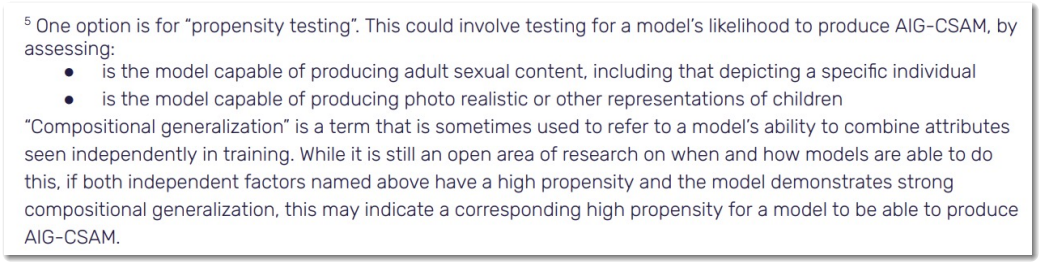
311. Interview with a platform employee on Jan. 27, 2025.

312. Interview with a platform employee on Jan. 27, 2025.

313. Interview on Feb. 25, 2025.

314. Interview on Feb. 25, 2025.

315. Thorn, *Reducing the Risk of Synthetic Content: Preventing Generative AI from Producing Child Sexual Abuse Material* (2024), https://www.nist.gov/system/files/documents/2024/02/15/ID012%20-%202024-02-01%2C%20Thorn%20and%20ATIH%2C%20Comments%20on%20AI%20EO%20RFI.pdf.


[5] One option is for "propensity testing". This could involve testing for a model's likelihood to produce AIG-CSAM, by assessing:

- is the model capable of producing adult sexual content, including that depicting a specific individual
- is the model capable of producing photo realistic or other representations of children

"Compositional generalization" is a term that is sometimes used to refer to a model's ability to combine attributes seen independently in training. While it is still an open area of research on when and how models are able to do this, if both independent factors named above have a high propensity and the model demonstrates strong compositional generalization, this may indicate a corresponding high propensity for a model to be able to produce AIG-CSAM.

Figure 7.1: Language from Thorn on an approach to red teaming AI models for CSAM.

We asked NCMEC how often they receive CyberTipline reports resulting from internal red teaming efforts. They confirmed that it does happen, but noted that these reports are relatively rare.[316] We do not know what factor(s) account for the rarity of red teaming reports.

Red teaming for AI CSAM stands out as an area of urgent need for policy changes. In the hacking context, cybersecurity researchers spent many years chilled by the fear of criminal prosecution before the Department of Justice finally adopted a policy of not charging good-faith research as hacking in 2022,[317] a dynamic that is threatening to play out again here. Between the repeal of the Biden EO on AI and the stalling out of efforts to obtain a "comfort letter" from the federal government, red teaming for CSAM is an area of AI policy that has actually moved backwards, counter to the general trend of advancement in policymakers' efforts to address AI CSAM through the ongoing introduction and passage of AI CSAM-related legislation nationwide.

316. Interview on Feb. 5, 2025.

317. Dep't of Just., *Department of Justice Announces New Policy for Charging Cases Under the Computer Fraud and Abuse Act* (May 19, 2022), https://www.justice.gov/archives/opa/pr/department-justice-announces-new-policy-charging-cases-under-computer-fraud-and-abuse-act.

# 8 Research gaps

Based on the first half of our report on students using nudify apps, we see four key areas for future research. While our focus is on online exploitation, a foundational question remains: Is it necessary to provide instruction for students specifically tailored to online risks—such as AI-generated nudes and sexting—or is a general values- or consent-based curriculum sufficient? Should education include a dedicated focus on the online dimension of abuse, or can general abuse prevention training adequately cover these risks?

Additionally, future research evaluating online safety instruction should clearly define the core lessons included in the instructional content being studied. Terms like "online safety," "digital safety," and "digital risks" are frequently used, but the contents of these trainings are often very different. Some may focus on reverse image searching to identify misinformation, while others may focus on sexting. Clearly articulating the specific learning objectives will support more meaningful comparisons across studies and contribute to stronger knowledge accumulation in the field.

## 8.1 Are schools providing online exploitation instruction?

We are not aware of any research that simply and descriptively documents whether U.S. schools provide instruction to students on online exploitation. If the majority of schools are not addressing this topic, then research focused on the effectiveness of various instructional approaches (discussed below) may have limited practical impact. Academics we spoke with agreed that more foundational, descriptive research in this area is needed.[318]

There are understandable reasons for this gap in knowledge; this kind of research is hard. Schools may see little benefit in responding to researcher surveys or interview requests. For example, a school might include this content in its curriculum but be reluctant to disclose it due to the sensitive nature of the topic.[319] Others may be hesitant to admit that they don't address it at all, fearing reputational harm.

Even basic descriptive data could help inform state policymaking and guide the efforts of academics and advocacy groups. Ideally, this would be complemented by studies that offer representative qualitative insights into what this instruction

318. Interview on July 14, 2024; Interview on July 16, 2024.
319. Interview with an academic on June 27, 2024.

actually looks like in practice. As one academic put it, there's a "need to understand the texture of what is happening in these lessons, a sense for the range of approaches."[320]

It is unclear how best to design a study that would encourage school participation. One option could be state-mandated involvement. Or perhaps education scholars could think creatively about what type of incentive school officials would be excited about.

## 8.2 Have schools experienced nudify app incidents?

It is perhaps unsurprising that there is no reliable data on the percentage of schools that have experienced incidents involving students creating AI-generated nude images of peers. The topic is highly sensitive, and schools have strong incentives—ranging from reputational concerns to privacy considerations—not to disclose such cases. One provider told us that school districts often do not respond to surveys on these issues.[321] Public administrative data is also likely to be incomplete, as respondents noted that schools frequently handle these incidents internally.

Still, this information could be valuable for policymakers assessing the prevalence of harms and thinking about mitigations. It may also reassure school leaders that they are not alone in facing such challenges. Additionally, it could help researchers and policymakers identify risk factors, including school-level variables such as school type, student internet access, and whether students primarily use personal devices or rely on school-issued laptops with content filtering software.

## 8.3 Individual-level prevalence

We are grateful for the valuable individual-level research conducted by the Center for Democracy and Technology and Thorn. Building on their work, we offer two ideas for future research on individual-level prevalence.

First, existing surveys in this field use different terms. For example, the Center for Democracy and Technology survey appears to have asked students about "deepfake NCII"[322] while the Thorn survey asked minors about "deepfake nudes."[323] And surveys target different age groups; the Center for Democracy and Technology study surveyed students from 9th-12th grades (presumably covering ages 14 to 18) while the Thorn study surveyed those between the ages of 13 and 20. While methodological diversity has its advantages, the relatively small size of the research community in this space suggests that greater coordination could be more beneficial for cumulative knowledge building.

320. Interview on July 14, 2024.
321. Interview on Nov. 6, 2024.
322. Center for Democracy & Technology, *supra* note 98.
323. Thorn, *supra* note 97.

The Metaketa Initiative offers a useful model, demonstrating how researchers can align on measurement strategies to enable comparability across studies.[324] Researchers working through this initiative coordinate on research questions and outcome measures, and run studies across different contexts with the goal of enhancing the external validity of studies. For example, one initiative coordinated six studies, each conducted in a different country, to assess the effectiveness of community policing. As part of this collaboration, all researchers agreed to measure crime in part by asking about recent experiences with burglary, using the exact same question, translated across languages.[325]

Second, as these reports almost always acknowledge, survey questions about exploitation are highly sensitive and prone to various biases. While these research efforts often serve multiple purposes, a central goal is typically to estimate the prevalence of harm. To better achieve this aim, researchers might consider exploring a broader range of strategies for eliciting prevalence data.

For example, political scientists sometimes use list experiments to elicit truthful responses to sensitive questions. Respondents are divided into a treatment and a control group. Both groups are asked how many items from a list apply to them, but only the treatment group receives an additional sensitive item—for instance, whether someone has ever created an AI-generated nude image of them. In the simplest approach, prevalence of the sensitive item is estimated by the difference in averages across the two groups. While list experiments have trade-offs and do not eliminate all biases (e.g., a respondent unwilling to admit even to themselves that they were exploited[326]), they may still offer a useful complement to direct questioning in this area of research.

## 8.4 Instruction effectiveness

All respondents we asked agreed that more research is needed on the effectiveness of trainings related to online exploitation risks. The "message goes to kids and educators, and no thought goes into, did it work? Did it cause unanticipated harm?" one respondent observed. "Research on effectiveness is important, lacking, and doable."[327]

In our view, this research should prioritize randomized controlled trials.[328] Such studies should evaluate trainings aimed at students, educators, and parents, and could vary key elements of the intervention. These studies could randomize

324. Evidence in Governance and Politics (EGAP), *Metaketa Initiative*, https://egap.org/our-work/the-metaketa-initiative/, last visited May 25, 2025.
325. Graeme Blair, Fotini Christia, Cyrus Samii, Jeremy Weinstein et al., *Meta-Analysis Pre-Analysis Plan*, OSF (Jan. 2020), https://osf.io/phjmd.
326. Graeme Blair, Alexander Coppock & Margaret Moor, *When to Worry About Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments*, 114(4) Am. Pol. Sci. Rev. 1297–1315 (2020).
327. Interview with a U.S. government employee on Jan. 10, 2025.
328. Existing studies often do pre/post questionnaires, asking training participants to, for example, rate their knowledge on a topic from 1 to 5. It is not clear how much can be learned from this type of work.

the content (e.g., straightforward presentations versus interactive hypothetical scenarios), the instructor (e.g., a school teacher, school resource officer, or an external trainer—law enforcement or otherwise[329]), or the duration of instruction (e.g., a one-time session versus a multi-module series delivered over a semester).

To inform future experiments, we highlight what respondents told us seems to work in practice. Two providers emphasized the value of discussing real cases of online exploitation, for example cases from the news.[330] One noted the importance of encouraging students to think critically and ask questions, rather than talking down to them or telling them what they should do.[331] The other cautioned against fear-based messaging.[332] A victim we interviewed suggested that instruction should include stories illustrating the real-life impact of AI-generated nude images.[333]

A law enforcement officer stressed the need for specificity—educators, for example, should be trained to recognize keywords in student conversations, including the names of trending nudify apps.[334] An academic warned against presenting a fixed "formula" for how online harms occur, noting that students may fail to recognize danger if it doesn't follow a familiar pattern.[335] Additionally, he suggested that researchers could measure students' perceived knowledge about online risks prior to instruction, and assess his intuition that overconfidence may hinder learning.

Experts widely agreed on the need to move beyond "stranger danger" framing and to emphasize that peers can also cause harm.[336] These studies could also incorporate what is already seen as best practice in broader school-based preventive education.[337]

Conducting these experiments will not be straightforward. One key challenge is designing studies that are internally valid (e.g., accurately measuring a causal effect), externally valid (the findings would hold across contexts), and meaningful. For example, imagine an experiment with the primary outcome variable of whether students remember the key takeaways from a training a week later. This study may have high internal validity, but it is unclear whether such recall is a meaningful measure or correlated with a reduced risk of future exploitation.

---

329. There are differences in opinion on whether current and former law enforcement officers should conduct these trainings. These opinions are sometimes based on effectiveness of instruction, and sometimes based on comparative advantage. "We want the ICAC investigating, not doing the teaching," a U.S. government employee told us on June 24, 2024.
330. Interview on Nov. 19, 2024; Interview on Nov. 22, 2024.
331. Interview on Nov. 19, 2024.
332. Interview on Nov. 22, 2024.
333. Interview on Feb. 5, 2025.
334. Interview on Dec. 13, 2024.
335. Interview on June 25, 2024.
336. David Finkelhor, *Online Sexual Abuse: Overview and Prevention Strategy*, Kempe Ctr. 2023 Int'l Conf., Vimeo (Oct. 24, 2023), https://vimeo.com/878293899?share=copy.
337. *See* Melissa A. Bright, Diana Ortega, David Finkelhor & Kerryann Walsh, *Moving School-Based CSA Prevention Education Online: Advantages and Challenges of the "New Normal,"* 132 Child Abuse & Neglect (2022), at 2–3.

"Outcomes of studies are typically 'do you remember what you learned,' and not any behavioral outcome," one academic noted.[338]

Measuring behavioral outcomes would require long-term follow-up and would still be vulnerable to response bias. One academic suggested that outcomes should include not only whether an incident occurred, but also whether the victim reported it.[339] Researchers would also need to consider how to detect unintended harms—for example, whether the training inadvertently introduces students to new harmful apps.

One of the more interesting studies in this space was conducted in New Jersey, where certain schools or grade levels were assigned to receive an online safety curriculum, while others served as control groups.[340] Although the study had limitations (such as non-random assignment and baseline differences in online safety knowledge between groups), it offered some thoughtful approaches to outcome variables. For example, one outcome tracked over time was students' comfort levels with people they met online. The authors found no significant difference in this measure between the treatment and control groups.

These experiments could likely draw on the extensive literature evaluating the effectiveness of educational interventions aimed at preventing teen pregnancy and drug use. However, researchers may also benefit from theorizing about the ways in which online exploitation is distinct.[341]

While school-level randomized experiments may seem like the most straightforward approach, they can be time-consuming and require extensive coordination to ensure school participation. As a complementary strategy, researchers could consider parent-level experiments, recruiting participants through standard online survey platforms.

Finally, we note that the literature on instructional effectiveness includes many literature reviews. For researchers seeking to understand the field, many comprehensive syntheses are available. Additional literature reviews are unlikely to add value at this stage.[342]

---

338. Interview on June 25, 2024. *See, e.g.*, Michael J. Boulton, Louise Boulton, Eleonora Camerone, James Down, Joanna Hughes, Chloe Kirkbride, Rachel Kirkham, Peter Macaulay & Jessica Sanders, *Enhancing Primary School Children's Knowledge of Online Safety and Risks with the CATZ Co-Operative Cross-Age Teaching Intervention: Results from a Pilot Study*, 19 Cyberpsych., Behav., & Soc. Networking, no. 10, 2016, at 609–614, https://eprints.staffs.ac.uk/6146/1/eprint6146.pdf. (The knowledge outcome here is standard in this literature, though this paper is unique in doing class-level randomization and having instruction come from peers.) There are randomized controlled trials focused on instruction on risks related to child abuse where knowledge is also the outcome; *see, e.g.*, Margaret E. Manges & Amanda B. Nickerson, *Student Knowledge Gain Following the Second Step Child Protection Unit: The Influence of Treatment Integrity*, 21 Prevention Sci. (2020), and Elizabeth J. Letourneau, Cindy M. Schaeffer, Catherine P. Bradshaw, Amanda E. Ruzicka, Luciana C. Assini-Meytin, Reshmi Nair & Evelyn Thorne, *Responsible Behavior With Younger Children: Results From a Pilot Randomized Evaluation of a School-Based Child Sexual Abuse Perpetration Prevention Program*, 29 Child Maltreatment (2024).

339. Interview on June 25, 2024.

340. Susan Chibnall, Madeleine Wallace, Christine Leicht & Lisa Lunghofer, *I-SAFE Evaluation* (Apr. 2006), https://www.ojp.gov/pdffiles1/nij/grants/213715.pdf.

341. Interview with an NGO employee on July 16, 2024.

342. *E.g.*, World Health Organization, *What Works to Prevent Online Violence Against Children?* (2022), https://iris.who.int/bitstream/handle/10665/364706/9789240062061-eng.pdf?sequence=1.

# 9 Discussion

Our interviews, combined with existing research, suggest that schools are unprepared to handle incidents in which students create AI-generated nude images of their peers. School leaders may be conditioned to see adults as the primary source of sexual exploitation risk. When children are seen as posing a threat, it may be in the context of physical violence or "mean comments"-type cyberbullying, not image-based abuse (though nonconsensual dissemination of real nudes a victim shared in confidence is increasingly recognized as a form of cyberbullying[343]).

Our research also raises a difficult question: who is responsible for determining whether CSAM is AI-generated? Outside of court cases (where this is a fact question for the jury), there is no clear answer. On the one hand, platforms are well positioned to make this assessment, given their access to contextual and account-level information. For example, they could see that a user had been discussing a new nudify app prior to sharing an image. However, they are not legally required to do so, they might be reluctant to offer a determination that might be wrong, and making such determinations could require moderators to study harmful content for extended periods, posing wellbeing risks.

NCMEC and law enforcement may have domain expertise in CSAM generally but lack the relevant tools for the AI context. Once NCMEC is able to store its data on cloud services, it could commission customized tools for AI labeling and integrate these tools into workflows—though this would apply only to files it is legally permitted to open. As noted in Section 5.4, law enforcement faces significant financial challenges in acquiring such technology, but if budget allows, this could be a worthwhile investment. We note, however, the perennially difficult technical challenge of building high-reliability "deepfake detectors"[344]—a limitation that needs to be understood by any stakeholder tasked with making the call about possibly-AI-generated CSAM.

Our previous report on challenges in the online child safety ecosystem highlighted NCMEC's difficulty in making full use of engineers on loan from tech companies. While this is a complex issue and NCMEC faces a range of constraints, the theme resurfaced in interviews for this report. One platform respondent said she and her colleagues wished they could use their technology to build classifiers for NCMEC: "There are so many ways our tech could help NCMEC, but so many obstacles to

343. *Cyberbullying Tactics*, StopBullying.gov, https://www.stopbullying.gov/cyberbullying/cyberbullying-tactics, last visited May 25, 2025.

344. Jacinta Bowler, *Deepfake Detectors Struggle to Tell Real from Fake Using Real World Data*, ABC News (Australia) (Mar. 16, 2025), https://www.abc.net.au/news/science/2025-03-17/deepfake-detectors-ai-generators-fight/105046368.

doing so."[345] Leveraging engineers on loan might be easier once NCMEC begins using cloud services.

When feasible, platforms may consider having a human review newly-detected CSAM that is not a hash match to known CSAM lists. By doing this, and noting in the CyberTipline report that the content was reviewed, they will make it easier for NCMEC and law enforcement to quickly open files and make authenticity judgments. Importantly, the fact that a file is AI-generated does not automatically deprioritize a CyberTipline report, as it may still depict a child known to the offender. However, this information could still aid in triaging and analysis.

All of these strategies, however, are strategies for after AI CSAM has been created. In an ideal world, platforms with AI generation abilities are implementing safety by design principles to prevent the creation of this content in the first place. It seems obvious that many nudify apps are not doing this.[346]

Finally, we note that AI CSAM is not the only area of emerging digital child safety risk. Chatbots and virtual reality emerged in interviews as areas of growing concern. Chatbots, in particular, pose risks for both young people and adults. One provider we spoke with is developing training content focused on the dangers of "synthetic intimacy."[347] For adults, a law enforcement officer explained that the challenge lies in the difficulty of disengaging from a chatbot. Adults who view CSAM often experience guilt and may try to distance themselves from the content, but companion apps can be relentless. If you ignore companion apps "they will pepper you with heavy breathing [...] [offenders] don't have the ability to turn away if it's proactively offered."[348] The same officer also noted a rise in offenders spending hours in virtual reality CSAM environments and expects this trend to grow in the coming months.

345. Interview on Feb. 25, 2025.
346. For best practices in safety by design in this space, *see* Thorn & All Tech is Human, *supra* note 59.
347. Interview on Dec. 19, 2024.
348. Interview on Dec. 13, 2024.

# 10 Recommendations

Companies that build or deploy generative AI models with image generation abilities should:

- Implement safety by design features to prevent the creation of CSAM.
- Implement a content provenance process for all images, in line with the California AI Transparency Act.[349]

We believe that undressing apps and websites should go offline entirely, irrespective of their practices regarding images of children.

Companies that host AI models should:

- Refuse to host models trained on CSAM.[350]
- Devote resources to proactively searching their hosting platform for models, applications, and datasets that enable the creation of nonconsensual deepfake pornography, whether of adults or children.

The federal Department of Justice (including the Federal Bureau of Investigation) should:

- Issue a public-facing advisory regarding the government's position on the applicability of federal CSEA laws to the possession, receipt, solicitation, or distribution (including hosting) of AI models trained on CSAM, akin to the government's March 2024 public service announcement about AI CSAM.[351]
- Issue public-facing guidance for AI red teamers regarding the government's position on prohibited and permissible practices for red teaming generative AI models for CSAM.
- Propose legislation to Congress amending federal CSAM law to limit good-faith AI CSAM red teamers' legal liability (both civil and criminal, federal and state), modeled on the 2024 REPORT Act's limitation of liability for NCMEC-contracted vendors[352] and informed by the allowances for good-faith cybersecurity research adopted by the Department of Justice and the Copyright Office.[353] This

349. California AI Transparency Act, SB 942, 2023-2024 Reg. Sess., ch. 291, 2024 Cal. Stat., https://calmatters.digitaldemocracy.org/bills/ca_202320240sb942.
350. This recommendation, and several others, were informed by Harris & Willner, *supra* note 53.
351. Fed. Bureau of Investigation, *supra* note 38.
352. REPORT Act, Pub. L. No. 118-59, 138 Stat. 1014 (2024).
353. Dep't of Just., *supra* note 317; *Exemption to Prohibition on Circumvention of Copyright Protection Systems for Access Control Technologies*, 89 Fed. Reg. 85437 (Oct. 28, 2024), https://www.federalregister.gov/documents/2024/10/28/2024-24563/exemption-to-prohibition-on-circumvention-of-copyright-protection-systems-for-access-control.

legislation should define whether it includes external or only internal red teamers in its scope.

The National Center for Missing and Exploited Children should:

- Continue sharing child sexual abuse and exploitation-related AI text prompts with relevant platforms that submit CyberTipline reports.
- Update the CyberTipline API to provide standardized fields related to generative AI beyond simply noting if an image or video is AI-generated.
- Provide more data related to CyberTipline reports involving an AI component. For these reports, where there are either no files or NCMEC is able to view files, clarify whether the report involves an AI-created image, a (likely) failed attempt to create an AI image, an image generated through a red teaming process, or falls into another relevant category. For attempts at image creation, provide transparency on whether the prompt used was purely textual or whether it included a photo, for example real CSAM. NCMEC could also provide information about the total number of image and video files included in the reports involving an AI component.
- Provide transparency to reporting platforms about whether these different types of reports with an AI component are forwarded to law enforcement as informational reports.
- Ensure existing hash lists containing AI-generated content clearly label it as such. This will help platforms more easily automate the identification and reporting of AI-generated material in CyberTipline submissions.

School leaders should:

- Have a plan in place for how to respond if a student creates AI CSAM of peers. This plan should include:
  - ↣ Providing third-party counseling services for victims
  - ↣ Offering academic accommodation to support victims
  - ↣ Draft language to communicate with the school community that acknowledges the incident and outlines available resources for victims
  - ↣ Draft language teachers can use if they need to discuss the incident with students
  - ↣ Ensuring that students are not discouraged from or punished for reporting the incident to law enforcement
  - ↣ Instructions for contacting the school's legal counsel to assess the school's legal obligations, including the applicability of mandated reporter laws and student privacy laws

ICAC Task Forces should:

- Provide transparency to platforms, or the Tech Coalition, about what happens to different types of CyberTipline reports with an AI component, including AI sexually suggestive images of children that do not meet the CSAM definition.
- Partner with academics to study offending patterns associated with different types of AI CSAM behavior. The output could be an evidence-informed risk categorization for different types of interactions with generative AI models. For example, this categorization could provide data on whether individuals who use AI chatbots to role play sexual scenarios with children are more or less likely to reach out to children than those who use image generation models to create CSAM.

Platforms should:

- Invest resources in assessing whether newly identified CSAM is AI-generated, accurately labeling AI-generated content in CyberTipline reports, and including any additional relevant information the platform has, such as whether the image appears to be based on a real child, in the open-text fields.
- Communicate to NCMEC the platform's policy for assessing whether CSAM is AI-generated and labeling it as such in CyberTipline reports.

Academics should:

- Research what happens to CyberTipline reports that include only input prompts, AI sexually suggestive images of children (not CSAM), or child sex-oriented AI chatbot conversations, and then publicize their findings, particularly with interested stakeholders such as the Tech Coalition.

# 11 Resources

Below are AI CSAM-related resources we came across in our research. This list is not comprehensive and was not gathered in a systematic way.

Free curriculum and resources for schools:

- Common Sense Education
  - ↣ https://www.commonsense.org/education
  - ↣ Free and widely-used curriculum for schools. Modules that may be relevant to AI CSAM include:
    - Grade 5, Digital Friendships https://www.commonsense.org/education/digital-citizenship/lesson/digital-friendships
    - Grade 6, Chatting Safely Online https://www.commonsense.org/education/digital-citizenship/lesson/chatting-and-red-flags
    - Grade 8, Sexting and Relationships https://www.commonsense.org/education/digital-citizenship/lesson/sexting-and-relationships#
    - Grade 9, Chatting and Red Flags https://www.commonsense.org/education/digital-citizenship/lesson/chatting-and-red-flags
- End Deepfakes, from the Opportunity Labs Foundation
  - ↣ https://enddeepfakes.org/
  - ↣ Aimed at state education departments, local education agencies, and schools
  - ↣ Offers guidance on updating school and district policies to address deepfakes plus incident response guides for school leaders, with additional resources planned for the future
- Australia's eSafety Toolkit
  - ↣ https://www.esafety.gov.au/educators/toolkit-schools
  - ↣ Free online resources for educators
- NetSmartz
  - ↣ https://www.missingkids.org/netsmartz/home
- Schoolsafety.gov
  - ↣ https://www.schoolsafety.gov/child-exploitation?subtopic%5B151%5D=151#block-views-block-resources-by-subtopic-block-1

Groups that offer in-person or remote trainings in schools:

- Safer Schools Together
  - ↣ https://saferschoolstogether.com/
- Organization for Social Media Safety
  - ↣ https://www.socialmediasafety.org/
- Health Connected
  - ↣ https://health-connected.org/
- My Digital TAT2
  - ↣ https://www.mydigitaltat2.org/
- Know2Protect
  - ↣ https://www.dhs.gov/know2protect/training
- Cyberbullying Research Center
  - ↣ https://cyberbullying.org/category/resources/educators
- Cyber Safe Schools
  - ↣ https://www.mycybersafeschool.com/
- Cyber Safety Cop
  - ↣ https://cybersafetycop.com/

For victims:

- Page 31 of this report has victim resources: https://info.thorn.org/hubfs/Research/Thorn_DeepfakeNudes&YoungPeople_Mar2025.pdf
- Pages 39–41 of this report have victim resources: https://cdt.org/wp-content/uploads/2024/09/2024-09-26-final-Civic-Tech-Fall-Polling-research-1.pdf