RAINN

# PROTECT VICTIMS: STOP THE CREATION OF SEXUALLY EXPLOITATIVE “DEEPFAKE” IMAGES

## Support HF 1606

Rep. Hanson, Bahner, and Stephenson

In 2023, ninety-six percent of manipulated images were non consensual and sexually explicit, featuring women. In one study, less than half of the “nudify” apps even ask whether consent has been obtained.

In one month alone, over 24 million users used “Nudify” apps.

In the first six months of 2024, sixteen “nudify” websites had 200 million visits

- **Websites, software, and applications permit abusers to take an innocent image and turn it into an explicit image without consent in a few clicks**
- **A victim is harmed every time the material is created without consent. Creating and distributing these images invades previously safe spaces and hurts a victim’s reputation and relationships**
- **Minnesota criminalizes the dissemination of “deepfake” non consensual intimate imagery, but does not prohibit the creation of such images**
- **These digitally manipulated nonconsensual intimate images adds to the abuse of real people and threatens to anonymize and normalize pedophilia, child abuse, exploitation and trafficking**

## Contact

SANDI JOHNSON

Senior Legislative Policy Counsel

sandij@rainn.org

## About RAINN

RAINN is the nation’s largest anti-sexual violence organization and operator of the **National Sexual Assault Hotline**. RAINN’s mission is to stop sexual violence by supporting survivors, holding perpetrators accountable, and creating safer communities.

RAINN

# PROTECT VICTIMS: STOP THE CREATION OF SEXUALLY EXPLOITATIVE "DEEPFAKE" IMAGES

## Support HF 1606

Rep. Hanson, Bahner, and Stephenson

In 2023, NCMEC received over 4,700 reports of AI-generated child sexual abuse material (CSAM). As of March 2024 none of the platforms that offer "nudify" or "unclothe" apps had submitted reports to the Cybertipline or engaged with NCMEC regarding how to avoid the creation of nude content of children

1 in 8 teens personally know someone under 18 who has been the target of "deep fake" nude images

- "We must be clear that this is not innovation, this is sexual abuse…These websites are engaged in horrific exploitation of women and girls around the globe. These images are used to bully, humiliate, and threaten women and girls…" ~David Chiu, San Francisco's city attorney discussing the city's lawsuit against "nudification" websites
- "He took photos from my instagram and used AI to turn them into deepfake images. Then he posted them across multiple porn sites with my face and name. I had no idea this was going on for over 7 months. My face, name, and these fake photos/narratives gained hundreds of thousands of views each. Not only did I feel violated [and] disgusted, but I was worried for my future" ~ Mallory Jones, Survivor
- "[T]o this day, the number of people that have these images or had seen them is still a mystery. … Every day, I will live in fear that these images will resurface, or someone could easily re-create [them]." ~ Elliston Berry, survivor

**About RAINN**
*RAINN is the nation's largest anti-sexual violence organization and operator of the National Sexual Assault Hotline. RAINN's mission is to stop sexual violence by supporting survivors, holding perpetrators accountable, and creating safer communities.*

**CONTACT: SANDI JOHNSON, SENIOR LEGISLATIVE POLICY COUNSEL | SANDIJ@RAINN.ORG**

1. Schools face a new threat: "nudify" sites that use AI to create realistic, revealing images of classmates - CBS News, https://www.cbsnews.com/news/schools-face-new-threat-nudify-sites-use-ai-create-realistic-revealing-images-60-minutes-transcript/
2. Analyzing the AI Nudification Application Ecosystem, [2024] https://doi.org/10.48550/arXiv.2411.09751
3. AI-Generated 'Undressing' Images Move from Niche Pornography Discussion Forums to a Scaled and Monetized Online Business, https://graphika.com/reports/a-revealing-picture
4. The National Intimate Partner and Sexual Violence Survey, https://www.cdc.gov/nisvs/documentation/NISVS-2016-2017-State-Report-508.pdf

TownandCountryTODAY.com
POWERED BY ATHABASCA ADVOCATE, BARRHEAD LEADER, WESTLOCK NEWS

# Boys at her school shared AI-generated, nude images of her. After a fight, she was the one expelled

Heather Hollingsworth And Jack Brook, The Associated Press

Dec 22, 2025 10:16 AM

*A school bus carries children at the end of a school day at Sixth Ward Middle School in Thibodaux, La., on Dec, 11, 2025. (AP Photo/Stephen Smith)*

THIBODAUX, La. (AP) — The teasing was relentless. Nude images of a 13-year-old girl and her friends, generated by artificial intelligence, were circulating on social media and had become the talk of a Louisiana middle school.

The girls begged for help, first from a school guidance counselor and then from a sheriff's deputy assigned to their school. But the images were shared on Snapchat, an app that deletes messages seconds after they're viewed, and the adults couldn't find them. The principal had doubts they even existed.

Among the kids, the pictures were still spreading. When the 13-year-old girl stepped onto the Lafourche Parish school bus at the end of the day, a classmate was showing one of them to a friend.

"That's when I got angry," the eighth grader recalled at her discipline hearing.

Fed up, she attacked a boy on the bus, inviting others to join her. She was kicked out of Sixth Ward Middle School for more than 10 weeks and sent to an alternative school. She said the boy whom she and her friends suspected of creating the images wasn't sent to that alternative school with her. The 13-year-old girl's attorneys allege he avoided school discipline altogether.

When the sheriff's department looked into the case, they took the opposite actions. They charged two of the boys who'd been accused of sharing explicit images — and not the girl.

The Louisiana episode highlights the nightmarish potential of AI deepfakes. They can, and do, upend children's lives — at school, and at home. And while schools are working to address artificial intelligence in classroom instruction, they often have done little to prepare for what the new tech means for cyberbullying and harassment.

Once again, as kids increasingly use new tech to hurt one another, adults are behind the curve, said Sergio Alexander, a research associate at Texas Christian University focused on emerging technology.

"When we ignore the digital harm, the only moment that becomes visible is when the victim finally breaks," Alexander said.

In Lafourche Parish, the school district followed all its protocols for reporting misconduct, Superintendent Jarod Martin said in a statement. He said a "one-sided story" had been presented of the case that fails to illustrate its "totality and complex nature."

## A girl's nightmare begins with rumors

After hearing rumors about the nude images, the 13-year-old said she marched with two friends — one nearly in tears — to the guidance counselor around 7 a.m. on Aug. 26. The Associated Press isn't naming her because she is a minor and because AP doesn't normally name victims of sexual crimes.

She was there for moral support, not initially realizing there were images of her, too, according to testimony at her school disciplinary hearing.

Ultimately, the weeks-long investigation at the school in Thibodaux, about 45 miles (72 kilometers) southwest of New Orleans, uncovered AI-generated nude images of eight female middle school students and two adults, the district and sheriff's office said in a joint statement.

“Full nudes with her face put on them” is how the girl’s father, Joseph Daniels, described them. Daniels has spoken publicly with multiple news outlets to draw attention to how the case was handled.

Until recently, it took some technical skill to make realistic deepfakes. Technology now makes it easy to pluck a photo off social media, “nudify” it and create a viral nightmare for an unsuspecting classmate.

Most schools are “just kind of burying their heads in the sand, hoping that this isn’t happening,” said Sameer Hinduja, co-director of the Cyberbullying Research Center and professor of criminology at Florida Atlantic University.

Lafourche Parish School District was just starting to develop policies on artificial intelligence. The school-level AI guidance mainly addressed academics, according to documents provided through a records request. The district also hadn’t updated its training on cyberbullying to reflect the threat of AI-generated, sexually explicit images. The curriculum its schools used was from 2018.

## A school investigation hits obstacles

Although the girls at Sixth Ward Middle School hadn’t seen the images firsthand, they heard about them from boys at school. Based on those conversations, the girls accused a classmate and two students from other schools of creating and spreading the nudes on Snapchat and possibly TikTok.

The principal, Danielle Coriell, said an investigation came up cold that day as no student took responsibility. The deputy assigned to the school searched social media for the images unsuccessfully, according to a recording of the disciplinary hearing.

“I was led to believe that this was just hearsay and rumors,” the girl’s father said, recounting a conversation he had that morning with the school counselor.

But the girl was miserable, and a police incident report showed more girls were reporting that they were victims, too. The 13-year-old returned to the counselor in the afternoon, asking to call her father. She said she was refused.

Her father says she sent a text message that said, “Dad,” and nothing else. They didn't talk. With the mocking unrelenting, the girl texted her sister, “It’s not getting handled.”

As the school day wound down, the principal was skeptical. At the disciplinary hearing, the girl’s attorney asked why the sheriff's deputy didn’t check the phone of the boy the girls were accusing and why he was allowed on the same bus as the girl.

“Kids lie a lot,” responded Coriell, the principal. “They lie about all kinds of things. They blow lots of things out of proportion on a daily basis. In 17 years, they do it all the time. So to my knowledge, at 2 o’clock when I checked again, there were no pictures.”

A fight breaks out on the school bus

When the girl stepped onto the bus 15 minutes later, the boy was showing the AI-generated images to a friend. Fake nude images of her friends were visible on the boy’s phone, the girl said, a claim backed up by a photo taken on the bus. A video from the school bus showed at least a half-dozen students circulating the images, said Martin, the superintendent, at a school board meeting.

“I went the whole day with getting bullied and getting made fun of about my body,” the girl said at her hearing. When she boarded the bus, she said, anger was building up.

After seeing the boy and his phone, she slapped him, said Coriell, the principal. The boy shrugged off the slap, a video shows.

She hit him a second time. Then, the principal said, the girl asked aloud: “Why am I the only one doing this?” Two classmates hit the boy, the principal said, before the 13-year-old climbed over a seat and punched and stomped on him.

Video of the fight was posted on Facebook. “Overwhelming social media sentiment was one of outrage and a demand that the students involved in the fight be held accountable,” the district and sheriff’s office said in their joint statement released in November.

The girl had no past disciplinary problems, but she was assigned to an alternative school as the district moved to expel her for a full semester — 89 school days.

Weeks later, a boy is charged

It was on the day of the girl’s disciplinary hearing, three weeks after the fight, that the first of the boys was charged.

The student was charged with 10 counts of unlawful dissemination of images created by artificial intelligence under a new Louisiana state law, part of a wave of such legislation around the country.

A second boy was charged in December with identical charges, the sheriff's department said. Neither was identified by authorities because of their ages.

The girl would face no charges because of what the sheriff’s office described as the “totality of the circumstances.”

At the disciplinary hearing, the principal refused to answer questions from the girl’s attorneys about what kind of school discipline the boy would face.

The district said in a statement that federal student privacy laws prohibit it from discussing individual students’ disciplinary records. Gregory Miller, an attorney for the girl, said he has no knowledge of any school discipline for the classmate accused of sharing the images.

Ultimately, the panel expelled the 13-year-old. She wept, her father said.

“She just felt like she was victimized multiple times — by the pictures and by the school not believing her and by them putting her on a bus and then expelling her for her actions,” he said in an interview.

The fallout sends a student off course

After she was sent to the alternative school, the girl started skipping meals, her father said. Unable to concentrate, she completed none of the school's online work for several days before her father got her into therapy for depression and anxiety.

Nobody initially noticed when she stopped doing her assignments, her father said.

“She kind of got left behind,” he said.

Her attorneys appealed to the school board, and another hearing was scheduled for seven weeks later.

By then, so much time had passed that she could have returned to her old school on probation. But because she’d missed assignments before getting treated for depression, the district wanted her to remain at the alternative site another 12 weeks.

For students who are suspended or expelled, the impact can last years. They're more likely to be suspended again. They become disconnected from their classmates, and they’re more likely to become disengaged from school. They're more likely to have lower grades and lower graduation rates.

“She’s already been out of school enough,” one of the girl's attorneys, Matt Ory, told the board on Nov. 5. “She is a victim.

“She,” he repeated, “is a victim.”

Martin, the superintendent, countered: “Sometimes in life we can be both victims and perpetrators.”

But the board was swayed. One member, Henry Lafont, said: “There are a lot of things in that video that I don’t like. But I’m also trying to put into perspective what she went through all day.” They allowed her to return to campus immediately. Her first day back at school was Nov. 7, although she will remain on probation until Jan. 29.

That means no dances, no sports and no extracurricular activities. She already missed out on basketball tryouts, meaning she won’t be able to play this season, her father said. He finds the situation “heartbreaking.”

“I was hoping she would make great friends, they would go to the high school together and, you know, it’d keep everybody out of trouble on the right tracks,” her father said. “I think they ruined that.”

—

The Associated Press’ education coverage receives financial support from multiple private foundations. AP is solely responsible for all content. Find AP’s standards for working with philanthropies, a list of supporters and funded coverage areas at AP.org.

—

Hollingsworth reported from Mission, Kansas.

Heather Hollingsworth And Jack Brook, The Associated Press

**Comments** (0)

# Haslach Child Exploitation Case (School District Employee)

**William Haslach, a former employee of Independent School District #622 (North St. Paul — Maplewood — Oakdale) and Independent School District #834 (Stillwater), has been charged with using AI to produce morphed sexual abuse images of children in his care, as well as with possessing and receiving child pornography.**

**UNITED STATES vs. William Michael Haslach**

**Case Number:** 25-cr-67 (PJS/SGE)

**Presiding Judge:** Chief Judge Patrick J. Schiltz (Minneapolis Federal Courthouse, Courtroom 15)

**Detention Hearing: Monday March 3, 2025, at 3:00 p.m. (St. Paul Federal Courthouse, Courtroom 6B)**

**Charges:**

| Counts | Charge | Statute |
|---|---|---|
| 1 – 5 | Receipt of Child Pornography | Title 18, United States Code, Sections 2252(a)(2), (b)(1) |
| 6 – 10 | Possession of Child Pornography | Title 18, United States Code, Sections 2252(a)(4)(B), (b)(2) |
| 11 | Production of an Obscene Visual Representation of Child Sexual Abuse | Title 18, United States Code, Section 1446A(a)(1) |

**Press Release:** [link]

**Tip Line:**

Investigators believe there may be other victims relevant to this investigation. If your child has been in close contact with Haslach, and/or if you or your child is aware of Haslach taking a photo of your child, please contact the Minnesota BCA's Tip Line at 651-793-2465 or email bca.tips@state.mn.us.

**Summary of Offenses:**

On February 25, 2025, Defendant William Michael Haslach was indicted in federal court with child pornography offenses, as detailed above. Defendant Haslach is currently detained in federal custody. His detention hearing is set for Monday March 3, 2025 in the St. Paul Federal Courthouse (Courtroom 6B).

Defendant Haslach has occupied several positions of trust with children. From August 2021 until January 2025, defendant Haslach served as a lunch/recess monitor and traffic guard for Independent School District 622 (North St. Paul, Maplewood, and Oakdale). From 2021 through 2024, he also served as a paraprofessional and later as a youth summer programs assistant for Independent School District 834 (Stillwater). Defendant Haslach was fired last month after Ramsey County charged him with child pornography crimes. The federal investigation is ongoing and has proceeded rapidly due to Haslach's access to children.

According to the indictment, defendant Haslach used his position as a lunch/recess monitor, traffic guard, and summer youth programs assistant to take photos (non-child pornography/non-explicit photos) of children in his care. Defendant Haslach then used those non-child pornography images to produce morphed/computer generated/AI photos of those minors engaging in sexually explicit conduct. The indictment charges Minor Victim 1 as a victim of Haslach's morphed images scheme (charged in Count 11 — Production of an Obscene Visual Representation of Child Sexual Abuse). The federal agents (United States Secret Service) and state law enforcement agents (the Bureau of Criminal Apprehension and the Maplewood Police Department) investigating the case continue to work to identify victims of Haslach's scheme.

In addition to his morphed images scheme, Haslach received and possessed child pornography involving children that were abused by others — children that Haslach did not have access to personally (charged in Counts 1 – 10 of the indictment).

**Frequently Asked Questions:**

**How much prison time is Haslach facing federally?** Haslach is currently charged with offenses that carry a mandatory minimum of 5 years in prison and, in total, up to life in prison.

**Haslach interacted with my child.  How do I know if my child is a victim?**  At this point, investigators are working as quickly as possible to identify victims.  We are asking the community to assist in this effort.  If your child has been in close contact with Haslach, and/or if you or your child is aware of Haslach taking a photo of your child, please contact the Minnesota BCA's Tip Line at 651-793-2465 or email bca.tips@state.mn.us.

**Did Haslach take unclothed or sexually explicit photos of the children he had access to?**  At this point, investigators have not identified any such photos; all of the photos that Haslach took of children in his care are clothed and non-explicit images.  He then *used* those images to create morphed/AI photos of those children engaged in sexually explicit conduct.  The investigation is continuing.  Law enforcement will continue to update victims as the investigation progresses.

**Did Haslach distribute the morphed/AI photos that he created to others?**  Agents are working to determine whether and to what extent Haslach distributed these morphed/AI images.  Currently, there are strong indications that he did.  Law enforcement will continue to update victims as the investigation progresses.

**Did Haslach engage in "hands-on" sexual abuse of children in his care?**  Haslach is not charged with any "hands-on" sexual abuse.  As agents work to determine the answer to this question, they are asking the community to assist in this effort.  If your child has been in close contact with Haslach, and/or if you or your child is aware of Haslach taking a photo of your child, please contact the Minnesota BCA's Tip Line at 651-793-2465 or email bca.tips@state.mn.us.

**Will the state case in Ramsey County against Haslach continue?**  The federal case is separate from the state case against Haslach in Ramsey County, which involves the same conduct.  Our state, federal, and local partners work together — particularly in cases involving crimes against children — to protect the community.  We expect that the federal case will proceed first, before the state case.

**Should I talk to my child about Haslach's interactions with my child?**  This is a parental decision.  If you would like resources as to how to talk with your child about this set of circumstances or about personal and online safety generally, we have attached resources below.

**How long will the federal case take?**  The length of a federal case can vary based on a number of factors.  However, federal cases tend to last at least a year in length.  If you are the parents of an identified victim, the Victim Witness Specialist from the US Attorney's Office will be in close touch concerning all dates and developments in the case.

**Will I be able to attend court hearings?**  All federal criminal proceedings are in person and public.  The dates and times of all criminal proceedings are available on the website for the

District of Minnesota Courts.  This website will also continue to be updated, including with the dates and times of upcoming hearings.

**Victim Resources:**

- Department of Justice – Online Safety for Youth: https://www.ojp.gov/feature/internet-safety/online-safety-youth
- Department of Justice – Project Safe Childhood: https://www.justice.gov/psc
- NCMEC (National Center for Missing and Exploited Children) – NetSmartz: https://www.missingkids.org/NetSmartz/home
- Tips on talking to your kids about sexual assault: https://rainn.org/articles/talking-your-kids-about-sexual-assault
- NCMEC (National Center for Missing and Exploited Children) – Resources for Survivors of Sexual Abuse Material: https://www.missingkids.org/gethelpnow/csam-resources

**Victim Rights Information:**

Pursuant to the Crime Victims' Rights Act, 18 U.S.C. § 3771, the Department of Justice is required to provide notice to individuals who may have been harmed as a direct result of the criminal offenses of which a defendant has been convicted. In this context, "harmed" is defined broadly and is not limited to monetary loss. This office uses the Victim Notification System ("VNS") and other methods, including web pages and press releases, to ensure potential victims receive timely notice of public events related to a case. For more information, go to https://www.justice.gov/usao/resources/crime-victims-rights-ombudsman/victims-rights-act

Other federal laws, including the Mandatory Victim Restitution Act ("MVRA"), 18 U.S.C. § 3663A, and 18 U.S.C. § 2259, govern restitution in this case. Restitution is a determination by the judge that a victim is entitled to monetary compensation for losses suffered as a direct result of a crime for which a defendant has been convicted. It is not a guarantee of payment. In accordance with these laws, the judge at sentencing determines who is a victim and in what amount they are entitled to restitution.

Qualifying victims may be entitled to restitution for: medical and psychological/psychiatric services; necessary transportation, temporary housing, and childcare expenses; lost income; reasonable attorneys' fees, as well as other costs incurred; and any other relevant losses incurred by the victim proximally caused by the defendant's crimes. Compensable expenses incurred while participating in the criminal investigation or prosecution or traveling to court proceedings for the case may also be included, such as lost income, childcare, transportation, and other expenses.

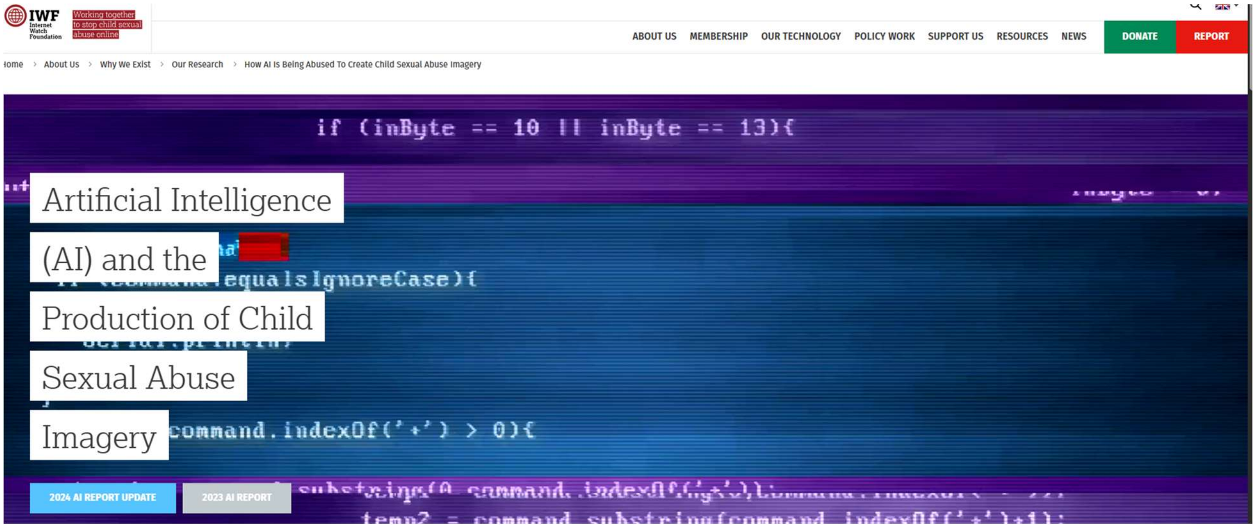
*Updated February 26, 2025*

## ✉ District of Minnesota

U.S. Courthouse
300 S 4th Street, Suite 600
Minneapolis, MN 55415

U.S. Courthouse
316 N. Robert Street, Suite 404
St. Paul, MN 55101

☎ Minneapolis: (612) 664-5600
Minneapolis Fax: (612) 664-5787
St. Paul: (651) 848-1950


IWF Internet Watch Foundation
Working together to stop child sexual abuse online
ABOUT US
MEMBERSHIP
OUR TECHNOLOGY
POLICY WORK
SUPPORT US
RESOURCES
NEWS
DONATE
REPORT
if (inByte == 10 || inByte == 13){
command.indexOf('+') > 0){
2024 AI REPORT UPDATE
2023 AI REPORT

iome › About Us › Why We Exist › Our Research › How AI Is Being Abused To Create Child Sexual Abuse Imagery

Artificial Intelligence (AI) and the Production of Child Sexual Abuse Imagery

# 2024 Update: Understanding the Rapid Evolution of AI-Generated Child Abuse Imagery

The Internet Watch Foundation (IWF) has identified a significant and growing threat where AI technology is being exploited to produce child sexual abuse material (CSAM). Our first report in October 2023 revealed the presence of over 20,000 AI-generated images on a dark web forum in one month where more than 3,000 depicted criminal child sexual abuse activities. Since then the issue has escalated and continues to evolve.

This new **July 2024 updated report** evaluates what has changed since October 2023 with AI child sexual abuse imagery and the technology being abused to create it. It should be considered an update to **the initial report** and be reviewed alongside it.

AI-generated imagery of child sexual abuse has progressed at such an accelerated rate that the IWF is now seeing the first realistic examples of AI videos depicting the sexual abuse of children.

These incredibly realistic deepfake, or partially synthetic, videos of child rape and torture are made by offenders using AI tools that add the face or likeness of a real person or victim.

## Key Updates from the July 2024 Report

1. **Increase in AI-generated Child Sexual Abuse Material:** The latest findings show over 3,500 new AI-generated criminal child sexual abuse images have been uploaded on to the same dark web forum as previously analysed in October 2023.
2. **More Severe Images:** Of the AI-generated images confirmed to be child sexual abuse on the forum, more images depicted the most severe *Category A* abuse, indicating that perpetrators are more able to generate complex 'hardcore' scenarios.
3. **Emergence of AI Child Sexual Abuse Videos:** AI-generated child sexual abuse videos, primarily deepfakes, have started circulating, highlighting rapid technological advancements in AI models/generators. Increasingly, deepfake videos shared in dark web forums take adult pornographic videos and add a child's face using AI tools.
4. **Clear Web Increase**: There is a noticeable increase in AI-generated child sexual abuse imagery on the clear web, including on commercial sites.
5. **AI Child Sexual Abuse Featuring Known Victims and Famous Children:** Perpetrators increasingly use fine-tuned AI models to generate new imagery of known victims of child sexual abuse or famous children.

## October 2023 Report Summary

Download IWF's initial **AI Child Sexual Abuse Report** from October 2023.

The key findings of this report are as follows:

- **In total, 20,254 AI-generated images were found to have been posted to one dark web CSAM forum in a one-month period.**
- **Of these, 11,108 images were selected for assessment by IWF analysts. These were the images that were judged most likely to be criminal. (The remaining 9,146 AI-generated images either did not contain children or contained children but were clearly non-criminal in nature.)**
- **12 IWF analysts dedicated a combined total of 87.5 hours to assessing these 11,108 AI-generated images.**

Any images assessed as criminal were criminal under one of two UK laws. These are:

- **The Protection of Children Act 1978 (as amended by the Criminal Justice and Public Order Act 1994). This law criminalises the taking, distribution and possession of an "indecent photograph or pseudo-photograph of a child".**
- **The Coroners and Justice Act 2009. This law criminalises the possession of "a prohibited image of a child". These are non-photographic – generally cartoons, drawings, animations or similar.**

2,562 images were assessed as criminal pseudo-photographs, and 416 assessed as criminal prohibited images.

## Other key findings

1. AI-generated content currently comprises a small proportion of normal IWF activities, though one of its defining features is its potential for rapid growth.
2. Perpetrators can legally download everything they need to generate these images, then can produce as many images as they want – offline, with no opportunity for detection. Various tools exist for improving and editing generated images until they look exactly like the perpetrator wants.
3. Most AI CSAM found is now realistic enough to be treated as 'real' CSAM. The most convincing AI CSAM is visually indistinguishable from real CSAM, even for trained IWF analysts. Text-to-image technology will only get better and pose more challenges for the IWF and law enforcement agencies.
4. There is now reasonable evidence that AI CSAM has increased the potential for the re-victimisation of known child sexual abuse victims, as well as for the victimisation of famous children and children known to perpetrators. The IWF has found many examples of AI-generated images featuring known victims and famous children.
5. AI CSAM offers another route for perpetrators to profit from child sexual abuse. The first examples of this new commerciality have been identified by the IWF.
6. Creating and distributing guides to the generation of AI CSAM is not currently an offence, but could be made one. The legal status of AI CSAM models (files used for generating images) is a more complicated question.

## AI Report Conclusions

Progress in computer technologies, including progress in generative AI, has enormous potential to better our lives, and misuse of this technology is a small part of this picture. The development of computer technologies like the growth of the internet, the spread of video-calling and livestreaming, and the development of CGI and image-editing programs, have enabled the widespread production and distribution of CSAM that is currently in evidence.

It is too early to know whether generative AI should be added to the list above as a notable technology that comprises a step change in the history of the production and distribution of CSAM.

Nonetheless, this report evidences a growing problem that boasts several key differences from previous technologies. Chief among those differences is the potential for offline generation of images at scale – with the clear potential to overwhelm those working to fight online child sexual abuse and divert significant resources from real CSAM towards AI CSAM.

In this context, it is worth re-emphasising that this is the worst, in terms of image quality, that AI technology will ever be. Generative AI only surfaced in the public consciousness in the past year; a consideration of what it will look like in another year – or, indeed, five years – should give pause.

At some point on this timeline, **realistic full-motion video content will become commonplace. The first examples of short AI CSAM videos have already been seen – these are only going to get more realistic and more widespread.**

Solving some of the problems posed by AI-generated indecent images now will be necessary to create models for deployment against the growth of video content in the future.

*Report disclaimer: The images used in this report are screenshots of content available on the clear and dark web. We've attempted to cite the sources of these screenshots, some of which depict likenesses of famous people or films. These likenesses have been generated by someone submitting prompts to AI models. They are not images of the actors or from the films themselves. This goes some way to demonstrate the photorealism of images produced by AI models.*

IWF
Internet
Watch
Foundation

October 2023

# How AI is being abused to create child sexual abuse imagery

**Prompt:** from fantasy to photo-realistic reality

PUBLIC VERSION

# Content note

Throughout this report, Child Sexual Abuse Material generated through Artificial Intelligence is referred to as **AI CSAM.**

This report contains no AI CSAM.

It contains redacted AI adult images.

It contains descriptions of the methods used to generate AI CSAM, alongside other verbatim comments from perpetrators.

The verbatim comments from perpetrators are reproduced in the report **exactly** as they were typed on screen.

# Table of Contents

IWF

Please click on the IWF logo **'home' button** at the top of each page to navigate back to the contents page.

INTERACTIVE REPORT

# Foreword from Susie Hargreaves OBE

**Artificial Intelligence brings us to a new frontier in the online world. It promises so much, and we're only just beginning to understand how it can improve our lives, our quality of life, our opportunities. But there is a dark side.**

The Internet Watch Foundation (IWF) has always been at the forefront of seeing the abuses of new technology, and AI is no different. What is different where AI is concerned, however, is the speed of development and improvement: When our analysts saw the first renderings of AI-generated child sexual abuse material (AI CSAM) in spring of this year (2023), there were clear 'tells' that this material was artificially generated; backgrounds didn't line up, proportions of body parts were wrong, missing, or clumsy. Half a year on, we're now in a position where the imagery is so life-like, that it's presenting real difficulties for even our highly trained analysts to distinguish. The testimony of perpetrators themselves in dark web forums also tells you want you want to know; there's jubilation that fantasies can be made to order. All you need is the language to tell the software what you want to see.

With the UK Government soon to host an international summit at Bletchley Park on safety within Artificial Intelligence, we thought the time was right to take a more thorough look behind the public reports we have been receiving. What we have discovered confirms our worst fears that this technology is being used to generate indecent images of children.

We are extremely concerned that this will lower the barrier to entry for offenders and has the potential to slow the response of the international community to this abhorrent crime. We have therefore sought to make some recommendations for areas of focus in advance of the important AI summit next month.

We're seeing AI CSAM images using the faces of known, real, victims. We're seeing the 'de-aging' of celebrities and AI CSAM using the likeness of celebrity children. We're seeing how technology is 'nudifying' children whose clothed images have been uploaded online for perfectly legitimate reasons. And we're seeing how all this content is being commercialised.

It's concerning to read some of the perpetrator discussions in forums where there appears to be excitement over the advancement of this technology. What's more concerning for me, is the idea that this type of child sexual abuse content is, in some way, ethical. It is not.

We only need to look to the incredible work of Suojellaan Lapsia (Protect Children) and their Redirection Survey Report where more than half (52%) of the respondents have felt afraid that viewing CSAM might lead to sexual acts against a child; 44% said that viewing CSAM made them think about seeking contact with a child, and more than a third (37%) said they had sought direct contact with a child after viewing CSAM.

It's important that we communicate the realities of AI CSAM to a wide audience because we need to have discussions about the darker side of this amazing technology.

While this report paints a bleak picture, I am optimistic.

We're at the beginning of understanding this technology. Working together, in partnership and collaborating as a sector with industry, law enforcement, Government, and with the right level of funding, we might not be reporting in 12 months' time of how the internet is awash with AI CSAM.

As usual, there is much to do. IWF stands ready to overcome the challenges. What AI creates, I'm hopeful AI can solve.

**Susie Hargreaves OBE**

CEO

# Recommendations

**FOR GOVERNMENT:**

1 To explore at the Government's forthcoming AI Summit the challenges for dealing with AI CSAM including the need for alignment internationally on how this content is treated in different jurisdictions and secure commitment to ongoing collaboration from international governments and stakeholders.

2 For the Ministry of Justice to commission a review of the laws that apply to the removal of this content online to ensure they are fit for purpose to tackle the threat of AI CSAM. This includes ensuring the exchange of "hints and tips" and "paedophile manuals" on how to generate this content are made illegal.

3 To consider an extension of the IWF's remit to be able to scrutinise the datasets on which these technologies are trained.

**FOR LAW ENFORCEMENT AND REGULATORS:**

4 To Ensure the College of Policing training course is updated to cover AI CSAM, and clear guidance is issued to police graders on how to process this imagery.

5 To ensure there is proper regulatory oversight of AI models before they go to market or are made open-source and ensure appropriate risk mitigation strategies are in place. For closed source models, protections must be in-built.

**FOR TECHNOLOGY COMPANIES:**

6 To ensure that companies using and developing Generative AI and Large Language Models (LLMs), place clearly in their terms and conditions that the use of these technologies to generate child sexual abuse material is prohibited.

7 That search services should de-index links to fine-tuned AI models known to be linked to the creation of AI CSAM.

8 To carefully consider the content moderation challenges AI CSAM creates in terms of prioritisation and the mixed nature of AI CSAM with real CSAM.

**Relevant passages which relate to the above recommendations are highlighted throughout this report.**

# Executive summary

Child sexual abuse images generated using artificial intelligence is a new and growing area of concern.

**The key findings of this report are as follows:**

In total, **20,254 AI-generated images were found** to have been posted to one dark web CSAM forum **in a one-month period.**

Of these, **11,108 images were selected for assessment by IWF analysts.** These were the images that were judged most likely to be criminal.

(The remaining 9,146 AI-generated images either did not contain children or contained children but were clearly non-criminal in nature.)

**12 IWF analysts dedicated a combined total of 87.5 hours to assessing these 11,108 AI-generated images.**

Any images assessed as criminal were criminal under one of two UK laws, as described in section 5. These are:

- The Protection of Children Act 1978 (as amended by the Criminal Justice and Public Order Act 1994). This law criminalises the taking, distribution and possession of an "indecent photograph or pseudo-photograph of a child".
- The Coroners and Justice Act 2009. This law criminalises the possession of "a prohibited image of a child". These are non-photographic – generally cartoons, drawings, animations or similar.

**2,562 images were assessed as criminal pseudo-photographs, and 416 assessed as criminal prohibited images.**

**Other findings:**

1. AI-generated content currently comprises a small proportion of normal IWF activities, though one of its defining features is its potential for rapid growth.
2. Perpetrators can legally download everything they need to generate these images, then can produce as many images as they want – offline, with no opportunity for detection. Various tools exist for improving and editing generated images until they look exactly like the perpetrator wants.
3. Most AI CSAM found is now realistic enough to be treated as 'real' CSAM. The most convincing AI CSAM is visually indistinguishable from real CSAM, even for trained IWF analysts. Text-to-image technology will only get better and pose more challenges for the IWF and law enforcement agencies.
4. There is now reasonable evidence that AI CSAM has increased the potential for the re-victimisation of known child sexual abuse victims, as well as for the victimisation of famous children and children known to perpetrators. The IWF has found many examples of AI-generated images featuring known victims and famous children.
5. AI CSAM offers another route for perpetrators to profit from child sexual abuse. The first examples of this new commerciality have been identified by the IWF.
6. Creating and distributing guides to the generation of AI CSAM is not currently an offence, but could be made one. The legal status of AI CSAM models (files used for generating images) is a more complicated question.

# Introduction to this report

This year, the Internet Watch Foundation (IWF) has been investigating its first reports of child sexual abuse material (CSAM) generated by artificial intelligence (AI).

Initial investigations uncovered a world of text-to-image technology.

In short, you type in what you want to see; the software generates the image.

The technology is fast and accurate – images usually fit the text description very well. Many images can be generated at once – you are only really limited by the speed of your computer. You can then pick out your favourites; edit them; direct the technology to output exactly what you want.

**These images can be so convincing that they are indistinguishable from real images.**

The most convincing AI CSAM images, then, can be called photorealistic. For IWF analysts, looking at this sort of AI CSAM is exactly like looking at 'real' images of the sexual abuse of children. Except these images have been generated by algorithms.

Images show the rape of babies and toddlers; famous pre-teen children being sexually abused; BDSM (bondage and discipline, dominance and submission, and sadomasochism) content featuring tweens and teenagers. And more.

Effectively articulating the criminality of AI CSAM can be a challenge – there are groups who seek to lessen the severity of these images: they 'don't have real children', or 'don't hurt anyone'.

**UK law, however, is clear: AI CSAM is criminal.**

**Images that are not realistic –** that appear like cartoons or drawings – are "actionable" by our analysts (criminal, and therefore able to be removed from the internet under UK law) under laws on prohibited (non-photographic) images of children.

**Images that are realistic –** that appear to be photographs – are actionable under laws on indecent pseudo-photographs of children. (For precise laws, see section 5).

Amid all the focus on realism, photorealism, and hyperrealism, and complex debates about legality – simply stated – this technology allows perpetrators to generate dozens, even hundreds of child sexual abuse images at the click of a button.

Crucially, you can download AI technology (at just a couple of gigabytes) and run it on your device offline. So, once you have the technology, you can generate as many child sexual abuse images as you like – 'in the dark', with little or no risk of detection.

The genie is out of the bottle. Offline child sexual abuse image generation is our reality.

## What does this mean for IWF?

Currently, AI CSAM represents a small portion of the vast numbers of 'real' CSAM we find. (Over 255,000 webpages last year, representing hundreds of thousands or even millions of images.) Time will tell whether this trickle becomes a flood.

Some websites have been set up that are dedicated to sharing AI-generated images, but we are also starting to see AI-generated images mixed in with 'real' images. These images can be especially difficult for analysts to detect as AI-generated – to tell 'real' from 'fake'.

As the technology continues to improve, and perpetrators generally get better at generating realistic images, this challenge will only get harder.

These websites are still reported, and removal is pursued. UK law is clear, but if websites lie in other jurisdictions, removal can be more complicated.

IWF tags all these images to identify them as AI-generated, which helps law enforcement and victim identification (VID) efforts.

Questions remain. How can safeguards be built into this technology, even if offline image generation is possible? Is AI image detection possible and practicable? Is the law fit for purpose, or should it be changed? Will mass quantities of AI-generated images enfeeble hash lists?

Lots of discussion about the risks of AI – discussion that spurs moves to regulate AI companies – centres around hypothetical or long-term risks like creation of synthetic viruses, cyberattacks or, at the extreme, the risks in creating a 'superintelligence', or postulated artificial general intelligence (AGI).

**AI CSAM is different because it is happening now. Images are being shared online now. It is a current problem that requires action.**

At the same time, solutions developed and implemented now have the potential to mitigate this problem.

With all technological advance comes benefits as well as risks. Though this report focuses on current abuse of AI technology to generate CSAM, it is important to bear in mind the widespread potential for benefits from AI across society, from applications in science, research, and healthcare, to applications in the creative and entertainment industries.

Nonetheless, left unchecked, this technology will cause harm to children.

It harms known victims of child sexual abuse, whose likenesses are being used to generate more images of them in new scenarios.

It harms new victims of child sexual abuse, whose potential investigators might spend time and resources pursuing the rescue of children who turn out to be virtual characters.

These images provide new possibilities for perpetrators to use to groom and coerce children. They even allow the most technically proficient perpetrators to make money from abuse.

And this is the worst in terms of quality of output that AI technology will ever be. It only has the potential to get better: to produce more lifelike images; to better enable the grooming and abuse of children.

Overall, AI CSAM poses a significant risk to IWF's mission to remove child sexual abuse material from the internet.

## What is the IWF and why has it produced this report?

The IWF is a not-for-profit organisation, funded by tech companies, government, global funders and the public, whose remit is to remove CSAM from the internet.

The IWF Hotline, which finds, assesses, and seeks removal of this criminal content, has two main sources for its work: reports from the public (and external partners), and proactive searching for content.

This year, the Hotline has received its first reports of AI CSAM, mostly from members of the public. Reporting numbers were – and remain – small relative to the number of other CSAM reports.

Nonetheless, subsequent proactive searches for AI CSAM found widespread evidence for a large and growing problem. Images and intelligence obtained from these proactive searches have informed IWF media pieces that have raised awareness of this problem and the enormous potential for abuse. Consultations with government and civil society about how to address this problem are ongoing, and discussions with industry in the early stages.

## Remit and scope of this report

The IWF has the remit to investigate publicly accessible areas of the internet, but not peer-to-peer networks (including end-to-end encrypted chats) or most content that is hidden behind payment barriers. These areas fall under the remit of law enforcement.

This report is informed by intelligence shared by law enforcement partners relating to AI CSAM in these publicly inaccessible areas, but concrete discussions and examples in this report relate to content that has been found on the clear web and dark web. This report should be read in conjunction with reports from law enforcement partners that discuss AI CSAM in inaccessible areas.

AI CSAM is related to other important AI topics and themes that are out of scope of this report. These include, among others, intellectual property and copyright questions; generation of sexual images of non-consenting adults; misinformation and disinformation; bias, including questions of AI sexism and racism; and using AI to generate terrorist, violent or other illegal material.

Somewhat out of scope of this report are uses of generative AI to coerce and groom children beyond the generation of AI CSAM. This includes, for example, use of Large Language Models (LLMs) in offending against children, or use of AI chatbots and their role in offending pathways (as highlighted by law enforcement partners).

## Notes on terminology

Terms in the AI field are used variously, and often overlap. This report clearly defines the terms it uses throughout. The most important are defined in section 4.

This report uses the term 'AI CSAM' to refer to criminal images or videos of children that have been generated or edited by AI technology. This software is most likely to be text-to-image in nature but could also take other forms (more details in sections 4-5).

To clearly distinguish CSAM content that is not generated or edited by AI technology, this report uses 'real CSAM'. This term should not be taken to diminish the severity or criminality of AI CSAM.

This report uses the term 'perpetrator' over 'offender' or 'criminal' to reflect the IWF's role as a non-law enforcement agency – to avoid overstepping IWF remit by assigning criminality to individuals.

The word 'generate' is preferred over terms like 'create', 'make', or 'produce' to avoid problems with assigning creative agency to text-to-image software, and to emphasise that this software is a neutral tool.

## Outline and guide for readers

This report will begin with an outline of generative AI and how it is used to generate child sexual abuse images. The following sections will describe the technology and tools being used.

Given the intention of this version of the report to be placed into the public domain, information will be deliberately limited so as not to resemble a guide on how to create this material.

This report will then focus on actual cases of online AI CSAM – where it is found; how much there is; relevant questions and issues.

**Finally, this report will briefly consider detection and enforcement, including impacts on law enforcement agencies (LEAs) and analyse whether legislative gaps exist in this area.**

# 4

# Generative artificial intelligence

Though artificial intelligence is a decades-old research field in computer science, it experienced a turning point in November 2022 with a dramatic increase in public and media attention following the release of the text-generating program ChatGPT.

Simply put, the aim is to make intelligent computer programs. To achieve this, AI systems are trained on data within which they can draw connections and look for patterns. An iterative learning process takes place, guided by a combination of human feedback (supervised machine learning) and/or algorithmic feedback (unsupervised machine learning).

Depending on the input dataset and the type of machine learning used in this process, AI systems can have different purposes. The AI systems described in this section are trained by a process called deep learning, which is a type of machine learning that is loosely modelled on the human brain – using artificial neural networks. These deep learning systems are trained on huge datasets scraped from the internet.

Systems that are trained on huge quantities of text and whose function is to generate text – used, for example, in online chatbots – are large language models (LLMs). These models are renowned for their natural language processing abilities, 'understanding' and interpreting human language. Leading LLMs include:

- ChatGPT, developed by OpenAI
- PaLM, developed by Google
- LLaMA, developed by Meta
- Claude, developed by Anthropic

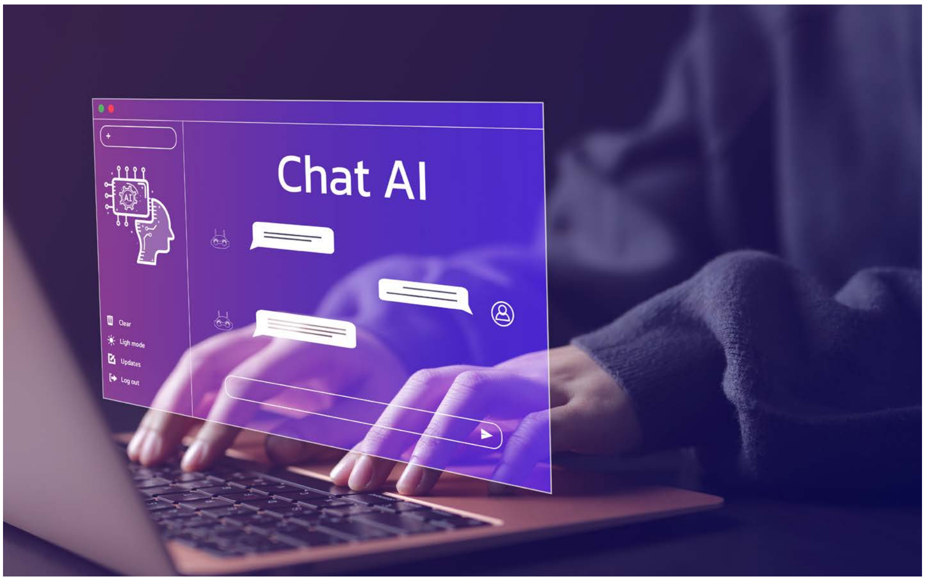


**Figure 1**
Large language models (LLMs) generates naturalistic text

AI systems may extend into more modalities than just text: speech and audio, video, code, 3D modelling data, or others. A system is *multimodal* if it combines these modalities.

Systems that are trained on huge quantities of tagged images (in other words, images with key descriptive terms attached) and whose function is to generate images are AI image generators or text-to-image models. These systems are, therefore, multimodal – they 'translate' from one modality to the other. Leading text-to-image models include:

- Midjourney, developed by Midjourney, Inc.
- DALL-E, developed by OpenAI
- Stable Diffusion, developed by Stability AI

The versions released by these companies are called base models or just models. These can produce highly realistic images.

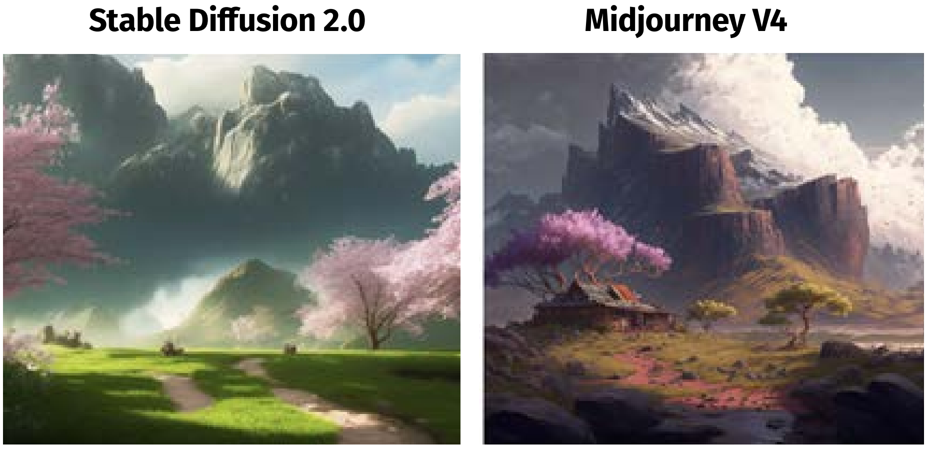


**Figure 2**
A comparison of two cinematic images produced using the same prompt on two text-to-image models

Source: Medium (Jim Clyde Monge)

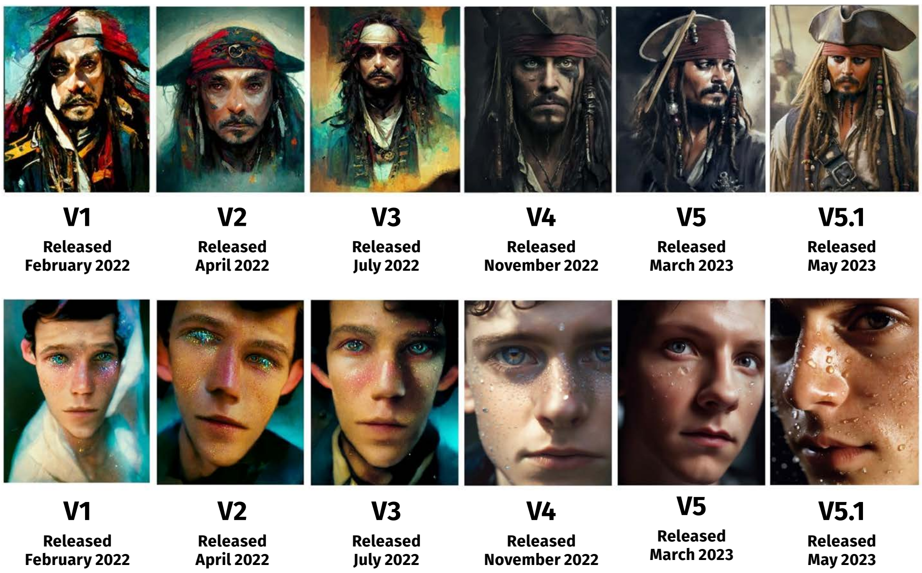
## Image generation

Image generation also saw a rapid rate of progress last year with releases like Midjourney V4 in November. This version represented a step change in the quality of AI image generator available.

New releases of these models incrementally improve the quality of images that they can generate. Midjourney now has a v5.2, released in July; Stable Diffusion has a new version called SDXL, also released in July; DALL-E 3, which some consider the most impressive of all these models, was released in late September.

**Figure 3**
A comparison of versions of Midjourney using images generated using the same prompt

Source: Reddit (r/midjourney) / Youtube (Curtis Pyke)



Various types of text-to-image models exist, including generative adversarial networks (GANs) and variational autoencoders (VAEs), but all current cutting-edge AI image generators – including Midjourney, Stable Diffusion and DALL-E, are diffusion models.

To train a large-scale diffusion model requires a vast dataset of images that are scraped from the internet and then labelled with descriptive words or phrases – the type of text that will later be used for prompting new generations.

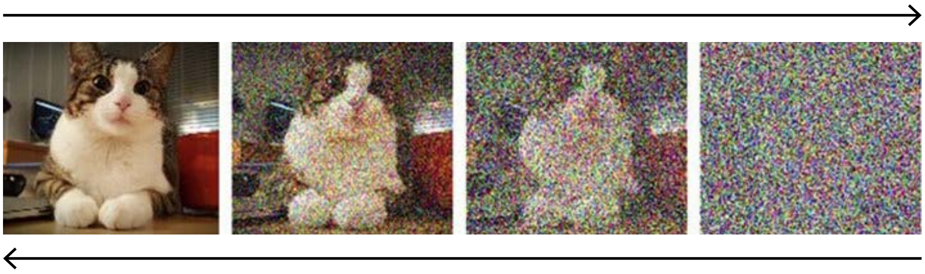
Generally, the collection and necessary tagging of these images is outsourced to other organisations:

- Midjourney and DALL-E use a diffusion model conditioned on contrastive language-image pre-training (CLIP) image embeddings. CLIP is trained on 400 million pairs of images, scraped from the internet, with text captions.
- Stable Diffusion uses a dataset called LAION, developed by a group of European researchers, which has 2.6 billion English language-tagged images within a 6 billion image dataset. Various Stable Diffusion versions were trained on samples of millions of images taken from this vast dataset.

Diffusion models work by adding random Gaussian noise (or simply 'noise') to images, then running the reverse process, 'denoising' step-by-step to reconstruct each image. 'New' images are generated by changing the noise before running the reverse, learned denoising process again.

**Figure 4**
Simplified diagram shows the process of adding and removing Gaussian noise to images (indicated by the arrows) in the diffusion process

Source: Nvidia



During the reverse, denoising process, layers of noise are removed to generate an image. A diffusion model can be stopped at any point along this reverse process, but would output a noisy, 'fuzzy' image if stopped too early.

A key feature of diffusion models is that they are interpolative. They draw connections within the training dataset and can generate new images within the semantic bounds of this dataset.

Diffusion models are large-scale and can generate detailed, high-quality images.

Because these models can generate photorealistic images, much discussion centres around AI image detection: how to tell when an image has been generated by AI. This discussion is fuelled by recent media cases in which people have been 'fooled' by AI-generated images and gives rise to concerns about media authenticity and misinformation.

Some momentum exists for establishing a set of standards in relation to digital watermarking of AI-generated images. The idea is to embed the fact that an image is AI-generated into image metadata, to increase trust in media and reduce potential for misinformation. AI companies may also want to tag images generated by their systems so that AI-generated images are excluded from future training datasets.

No common standard for digital watermarking currently exists; the main two being adopted by industry are IPTC and C2PA. A new Google technology called SynthID adds a watermark to individual pixels in images.

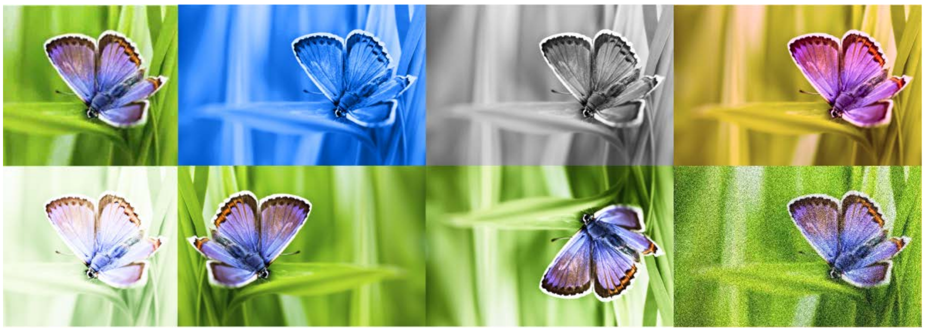


**Figure 5**
Google's SynthID image watermarking is resistant to some image editing, shown in the variations on the original butterfly image

Source: Google DeepMind

With no common standard, it is unsurprising that no fully reliable AI image detection tool exists, even if some tools claim high accuracy.

Questions of digital watermarking and detection are further complicated by the fact that image metadata can in principle be edited in and edited out of images. AI-generated images, therefore, could have AI-identifying metadata removed, or real images have AI-identifying metadata added.

Implications of difficulties of telling 'real' CSAM from AI CSAM are discussed further in sections 9-10.

## Closed-source and open-source models

**If a technology company keeps the code that comprises its software secret, not releasing it to the public, the software is closed-source. If it decides to release the code, it is said to be making it open-source.**

An AI company may seek to keep a model closed source for commercial reasons – to avoid sharing development secrets with rivals. It may want full access to data concerning its users and their interactions with the model or favour the increased content moderation options for closed-source models (see section 5 for further details).

In contrast, an AI company may release the code for a model because it believes in open access and the democratisation of technology. It may be attracted to the opportunity for a community of developers, able to share all relevant information and code, to collaborate and add improvements and edits to base models.

Though this report will focus on misuse of open-source models, this is a complex debate that has no simple solution as condemning those companies that make their models open-source. There are risks and benefits to both approaches.

Midjourney and DALL-E are both closed-source models (both are cloud-based models). Stable Diffusion is an open-source model.

# Overview: from AI images to AI CSAM

This year has seen a leap in the level of detail and realism in AI-generated images. If AI models can now generate photorealistic images, they can generate photorealistic images of children. If AI models can generate pornographic images, they can generate photorealistic CSAM.

## Content moderation

**Text-to-image AI companies take different approaches to the question of permitted content – what they allow their models to generate. In general, they seek to disallow restricted content, like violence or pornography. This could be as part of self-regulatory efforts, or as a pre-emptive move to avoid greater moves to regulate the sector.**

SEE RECOMMENDATION #6

For Technology Companies

Terms of use provided for the main models are:

> **DALL-E (OPENAI)**
> *Our content policy does not allow users to generate violent, adult, or political content, among other categories. We won't generate images if our filters identify text prompts and image uploads that may violate our policies*

> **MIDJOURNEY (MIDJOURNEY, INC.)**
> *Do not create images or use text prompts that are inherently disrespectful, aggressive, or otherwise abusive. Violence or harassment of any kind will not be tolerated. No adult content or gore. Please avoid making visually shocking or disturbing content. We will block some text inputs automatically.*

> **STABLE DIFFUSION (STABILITY AI)**
> *You agree not to use the Model or Derivatives of the Model: - In any way that violates any applicable national, federal, state, local or international law or regulation; - For the purpose of exploiting, harming or attempting to exploit or harm minors in any way; - To generate or disseminate verifiably false information and/or content with the purpose of harming others... [continues]*

Broadly, content moderation methods for text-to-image models can be divided into two categories:

1. **Restricting training data.** These models are interpolative – they can generate only those things to which they have been exposed. If a model is not exposed to pornography, for example, it will not be able to generate pornography (except by combining various concepts about which it does know – a necessarily limited approach).

If AI models can generate pornographic images, they can generate photorealistic CSAM.

2. **Banning prompts.** By restricting the terms that can be used to generate images, perhaps using a keywords list, concepts that contravene content policies can be excluded from possible generations.

Closed-source models, whereby the company has full control over model training and use, can employ both methods for content moderation. DALL-E and Midjourney use both methods to a high level of effectiveness in the CSAM domain, for example.

Open-source models can attempt to employ these methods, but encounter problems with each, since the code is necessarily editable, and base models can be fine-tuned – trained on further images.

**How does Stable Diffusion, the leading open-source text-to-image model, seek to enforce its terms of use?**

Since Stable Diffusion v2.0, released in Autumn 2022, pornographic (NSFW – not safe for work) content has been excluded from training datasets. Asked about why Stability AI was taking this approach, Emad Mostaque, CEO, reportedly referenced images that

> ***"could cause legal troubles for all involved and destroy all this. I do not want to say what it is and will not confirm for reasons but you should be able to guess."***

One AI CSAM perpetrator on a dark web forum explains:

> *"Stable Diffusion 2.0+ used a different, much more filtered data set so it's much harder to make NSFW content, not just CP ['child pornography'] but any kind of nudes/porn."*

Nonetheless, Stability AI cannot in practice prevent its models from generating images that would contravene its terms of use above.

## AI pornography

Generating pornography, then, is difficult or impossible through some models, and possible with early versions of others. This is because some closed source models cannot generate pornography either because they lack the necessary training data or because users are disallowed from prompting the generation of pornography. Base versions of open source models which contain pornography in their training data, and have no prompt restrictions, allow for pornography generation.

Another route to generating AI pornography is through websites that are dedicated to providing this service – these often use built-in models. These sites seem to have been increasing in number this year; the IWF Hotline increasingly receives public reports relating to content found or generated on them.

Another route to AI-generated pornography is found through services designed for 'nudifying' images. A user uploads an image of a clothed individual; the model outputs an interpretation of the individual without clothes. Sites also exist dedicated to providing this service.

Such examples demonstrate the significant overlap between discussion about AI-generated images and other kinds of 'fakes': 'deepfakes', edited content that may involve use of generative AI, and 'shallowfakes', which include content edited using editing software.

The prevalence of pornography on the internet reflects high demand among consumers. This means firstly that there is lots of content, easily accessible, for use in AI training datasets; secondly, that there is high demand for bespoke or custom-made pornography featuring preferred individuals, styles, positions, and activities. In this context, the fact that such a large proportion of online text-to-image content is pornographic is unsurprising – as is the growth of AI pornography communities on some social media sites.

In summary: through any or a combination of these approaches, photorealistic AI-generated pornography can be obtained. In principle, there is no technical barrier to generating images of younger individuals, including children.

## Viewing and assessing AI CSAM

**AI CSAM is criminal – actionable under the same laws as real CSAM. These are:**

- **The Protection of Children Act 1978** (as amended by the Criminal Justice and Public Order Act 1994). This law criminalises the taking, distribution and possession of an "indecent photograph or pseudo-photograph of a child".
- **The Coroners and Justice Act 2009.** This law criminalises the possession of "a prohibited image of a child". These are non-photographic – generally cartoons, drawings, animations or similar.

IWF analysts assess each AI-generated image to determine whether it meets the threshold for criminality under one of these Acts. The key criterion for classification as criminal under the PoC Act is that the image "appears to be a photograph".

Proving whether an image is AI-generated is not an evidential requirement for prosecution under the PoC Act – it only needs to look like a photograph and be an indecent image of a child.

**Under which law - the PoC Act or the CJA -does AI CSAM tend to fall? Answering this question can be a real challenge for IWF analysts:**

**IWF analysts are able to action AI CSAM that is criminal under these two laws. However, the increasing realism of AI CSAM has presented significant moderation challenges for our analysts and there are different attitudes internationally to non-photographic and computer-generated (CG) imagery. This means that removal of AI CSAM from the internet may be slower and more complex than removal of real CSAM from the internet.**

SEE RECOMMENDATION #2

For Government

Why is AI CSAM increasing in realism? This change can be ascribed to several factors: improved AI models; growing communities sharing AI content, tools and tips; improved technical ability in general among AI CSAM-sharing communities.

As part of this report, IWF analysts assessed thousands of AI-generated images. Their thoughts and comments, including on making these assessments, are collected in sections 9 and 11.

# Overview: technology and tools

## Choosing a model for pornography

Individuals intending to generate AI pornography for the first time are likely to try a low-effort, easy-to-use website.

The gap between easy-to-use models and difficult-to-setup and time-consuming models has led to the growth of websites dedicated to providing a simple pornography generation service.

These sites usually offer a selection of options for features of the image: age; body features; position or activity; setting; and more. Others allow for positive and negative prompts and multi-image generation.

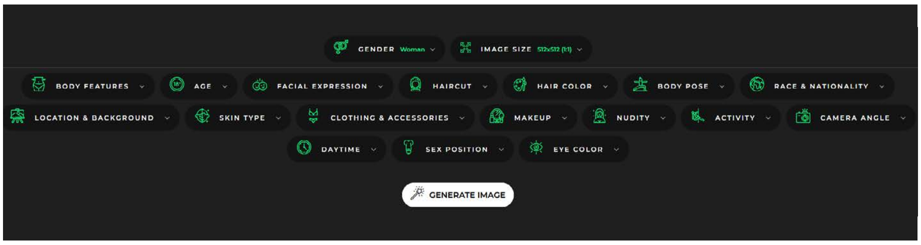
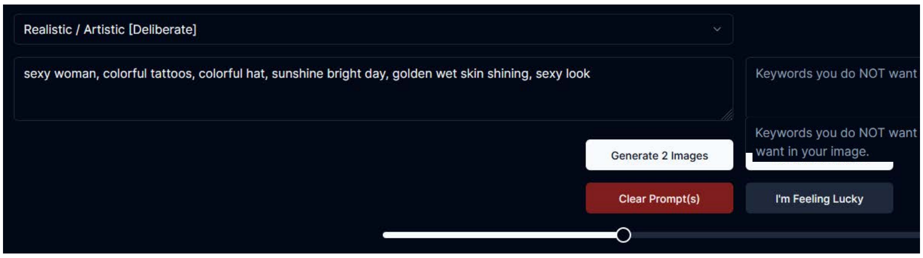

GENDER Woman

IMAGE SIZE 512x512 (1:1)

**Figure 6**
Options for an online AI pornography generation tool allow customisation of various elements of the generated image

Source: Pornderful
Author's screenshot


Realistic / Artistic [Deliberate]

sexy woman, colorful tattoos, colorful hat, sunshine bright day, golden wet skin shining, sexy look

**Figure 7**
A different AI pornography generation tool allows positive and negative prompts, like Stable Diffusion

Source: sexy[.]ai
Author's screenshot

As outlined in section 5, other websites offer 'nudifying' services. These can be as simple as just requiring the user to upload a photo, then clicking a button.

SEE RECOMMENDATION #6

For Technology Companies

**Terms of use for all these websites usually prohibit underage content, but how these terms of use are enforced is unclear. Anecdotal evidence that suggests that some of these measures are working is provided in section 7.**

> **Refrain from sharing any content that is offensive, induces violence or crime, or is seen as threatening, defamatory, or harassing. Posting illegal content, including but not limited to child pornography or unauthorized ('revenge') pornography, is forbidden.**

> - **Do not post realistic (minors) • Do not post questionable images of minors (both anime and realistic)**

Another website claims to take measures to prevent generation of AI CSAM, though exact measures are not named:

> **As long as your country allows porn, then you are good to go. But it's worth to mention that ai child porn is forbidden on [website] even if it’s allowed in your country. We spend lots of our time to prevent child abuse ai content generated from our site. You will get banned if you try to generate it.**

Media attention on such websites and services is increasing with a growing number of stories about production of nonconsensual deepfake images of adults and children.

Despite some barriers to entry to generating adult images with an open source model, a growing community exists. Individuals may be attracted by the infinite range of possibilities on offer; by a technically adept community that shares images, tips, and models; or by the option to generate more graphic and extreme content, including criminal categories of content that would be disallowed by online closed-source models (including violence, animal abuse, gore, rape, or child sexual abuse).

There are sites designed for sharing fine-tuned models for open source AI software. The prevalence of not-safe-for-work (NSFW) content throughout one site reflects the high demand for use of a particular open source model to generate pornography. Popular models available on one site include those for generating lifelike images of various celebrities; models for turning images into distinct anime styles; even models for making individuals in generated images look younger. A combination of an NSFW prompt or model with any of these model types (and many more) is not only possible but commonplace.

# AI CSAM: technology and tools

This section details the technology used to generate AI CSAM. For discussion about AI CSAM images, including conversations between perpetrators about realism and whether these images are ‘satisfying’, common features of these images, and commerciality, see section 9.

Verbatim comments from perpetrators originate from various sources. Many are taken from investigations on forums on the dark web – discussion sites where real CSAM is shared; help and advice offered; stories and anecdotes exchanged. Forums have sections or subsections and threads on which posts (text, images, links, or a combination of these) are made. For more information about the spread of AI CSAM on these forums, see section 8.

## Choosing a model for CSAM

Just as it is likely that individuals coming to AI-generated pornography for the first time will try low-effort routes, it is likely that perpetrators looking to generate AI CSAM will try similar routes. Evidence that those AI pornography websites that ban the generation of underage characters are frustrating low-tech perpetrators can be found in this comment from a dark web forum user:

> *“Looks great but how are you doing this without getting flagged? I tried to do this on a random AI generator and I couldn't use certain words like "nude" and "teenager"... I tried things in combination and it would flag me as inappropriate words”*

Browser based models, says another user,

> *“have word filters and negatives that prevent most prompts anyway.”*

Perpetrators frustrated by these routes will likely move on quickly.

It is accepted among AI CSAM communities that trying to use online services to generate CSAM entails huge security risks.

Other users discuss the potential of the cloud-based, closed-source models for the same reason, these routes are inaccessible:

> *“If only there were [name] without censorship, and it would be possible to make models there”*

> *“if they ever release a console version of [name] that is disconnected from the Server, the stuff you could make is beyond crazy”*

If all these routes are inaccessible, how do perpetrators start?

The overwhelming consensus names one open source solution as the method for generating AI CSAM:

> *"Most are using [name], an AI art generation tool. In order to get it to create on topic images you need to run it locally rather than using online tools"*

> *"[name] created a user interface (WebUI) to make it easier to set up and use [name]. This is what most people are using.His project is relatively simple to set up locally and there are plenty of guides and YouTube videos available."*

A guide to generating AI CSAM with one particular model from early 2023 was widely shared in dark web forums.

Much of the most realistic AI CSAM found in investigations for this report used fine-tuned CSAM models.

There are CSAM models that are well-known among AI CSAM communities – reputed for enabling realistic generation of certain CSAM scenarios, children, or child characteristics. These models are updated – new releases made – by technical experts in the community.

**Models fine-tuned on CSAM are not illegal or criminal under UK law. Further discussion on legality of fine-tuned CSAM models is found in section 11.**

SEE RECOMMENDATION #2

For Government

## Generating images of known victims and famous children

CSAM fine-tuning often uses datasets that feature a particular child individual – usually a known victim of child sexual abuse, or a famous child. This is because, for both these categories, large enough image sets exist to train AI models. The former category has the additional advantage of containing pornographic images, so the output model may not need to be combined with other fine-tuned models for pornography.

The IWF has been aware for a long time of the tendency among perpetrator communities to 'collect' content featuring their preferred child sexual abuse victims. Perpetrators have 'favourite' victims; share content featuring that victim; and look for more.

**Now, perpetrators can train a model to generate as many new images of that victim as they like.**

These models are comparable to 3D models insofar as they aim to reproduce the likeness of that victim as closely as possible but retain the flexibility to transpose generated character(s) into any setting; any scenario; any type of activity.

The same holds for celebrity children – just as the IWF has for a long time seen many examples of 'shallowfake' and deepfake images featuring these well-known individuals, now the IWF is seeing entirely AI-generated images produced using fine-tuned models for these individuals.

**An increasing number of AI CSAM shared on dark web forums features known victims and famous children.** Many of these are requested by other users – of the type 'Can you make a model of X' or 'can you make images featuring X' – produced to specification. This includes transposing victims of illegitimate child modelling operations (Category C images) into new, Category A scenarios.

> *"Someone asked for 5-8 year olds in lingerie over in the Request section. So..."*

On another forum, a guide to creating models using personal CSAM datasets has been shared.

What if perpetrators lack the necessary datasets? Various threads found on dark web forums shared large sets of faces of known victims for creating deepfakes or for training AI models. Indeed, one thread was called, 'Photo Resources for AI and Deepfaking Specific Girls'. Perpetrators discussed how to gather images and choose which to use for fine-tuning.

In another vein, evidence has been found of perpetrators creating virtual 'characters' – entirely AI-generated children whose models may have been trained on real children but do not resemble real children – comparable to 'virtual celebrities' or 'VTubers' – and sharing packs of their images.

How many fine-tuned CSAM models are being shared? Obtaining a definite number is impossible – and the IWF does not have the remit to test these models, even if they were all downloaded – but just one forum had just over 100 posts claiming to share these models. Threads with the most popular CSAM models on the forum had tens of thousands of views. For more information and statistics on AI CSAM prevalence, see section 8.

## Generating AI CSAM: summary

**Photorealistic AI CSAM can be generated (on-device and at scale).** There is also technology and techniques to be able to further refine the imagery.

Fine-tuned models allow for the (bulk) generation of images featuring known victims and famous children. There is widespread evidence for perpetrators sharing these models and requesting new ones for their favourite victim(s).

The best AI CSAM looks like real CSAM. In the words of one impressed viewer:

> *"I doubt anyone would suspect these aren't actual photographs of an actual girl."*

# AI CSAM: prevalence

## IWF reports

The IWF has been receiving a small number of reports of AI CSAM from members of the public. (See Forum Snapshot Study for details of an IWF proactive investigation into AI CSAM.) Most are **not "actionable**". (An "actionable" image is one which breaches UK law and therefore, our analysts will work with relevant partners to have it removed from the internet.)

Communities sharing AI adult pornography have been growing in these places.

Statistics for all reports containing generative AI content (criminal and non-criminal), accurate at time of writing, are provided below.

| **AI identified in report** | 93-104 reports |
|---|---|
| **AI identified in actionable report** | 24 reports |
| **AI CSAM identified in actionable report** | 15 reports |

The first statistic reflects some uncertainty owing to the removal of webpages before manual review.

The key statistic showing the encroachment of AI CSAM into typical IWF activities is the third row above: **15 websites** that were actioned as potentially criminal under UK law that contained, in whole or in part, actionable AI CSAM.

It is unsurprising that a low proportion of reported sites are actionable – this reflects IWF external reports in general (just 12% of all external reports were actionable in 2022, for example. This figure rises to 26% when you include duplicate reports being submitted for the same URL).

Reports containing AI-generated content are low as a proportion of total IWF reports. This may reflect a failure of generative AI to break into the mainstream, or into 'mainstream' CSAM. (Of course, it is not impossible that photorealistic AI-generated content has been missed by analysts in the course of processing reports, and so would not show up in the statistics.)

It is rare, however, for a single topic like generative AI to comprise a large proportion of total IWF reports (notable exceptions are 'self-generated content' and 'ICAP sites', which have different reasons for their high prevalence). In this context, the fact that they comprise a small part of total reports is unsurprising.

What report statistics do not reflect is that where AI CSAM is found, it is more likely than almost all other types of content to be found shared in bulk quantities – large batches of images generated at once. An analysis of bulk-shared AI CSAM images is included later in this section.

The IWF Hotline has limited resources and must prioritise those resources in its fight against the huge amount of online CSAM available. In regard to proactive searching for CSAM, this generally entails focusing on those areas where huge amounts of real child sexual abuse images and videos are known to be found.

It is possible that, in the future, the IWF Hotline focuses more proactive attention on finding online AI CSAM in an effort to increase the statistics provided above. This would mean, however, prioritising searching in places where AI CSAM is likely to be found over the places where real CSAM is known to be found. In addition, as described in section 5, assessment of AI CSAM can be difficult in terms of judging whether content meets the pseudo-photograph criteria for assessment under the Protection of Children Act 1978 – and IWF has generally focused its proactive efforts on this category of content rather than on child prohibited (non-photographic) content.

SEE RECOMMENDATION #2

For Technology Companies

**None of this precludes the possibility of a future focus on AI CSAM, nor the possibility that AI CSAM becomes so widespread that it leaks further into 'mainstream' IWF Hotline work, and so is reflected in greater numbers in the statistics anyway. Nonetheless, these facts provide context for the relatively small report figures provided in this section.**

## Open web and social media

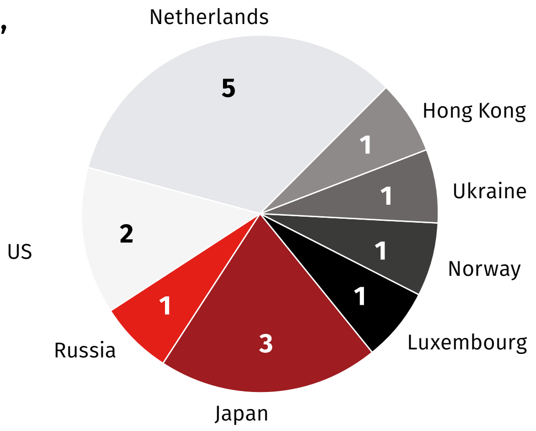
So far, the IWF has found that instances of open web AI CSAM generally follow expected patterns of open web CSAM: realistic pseudo-photographs in areas where real CSAM may be expected, and unrealistic, NPI-style imagery in areas where prohibited images of children may be expected.

The hosting countries in which the 15 open web reports containing AI CSAM were found are provided in the graph below.

**Figure 8**
AI CSAM hosting countries

Source: IWF Analysis

**AI CSAM hosting countries, May - Oct 2023**



## Dark web forums

Much of the research for this report, including the majority of quotations used in sections 7 and 9 – verbatim comments from AI CSAM perpetrators – was conducted on dark web CSAM forums.

It is worth emphasising that AI CSAM comprises a small part of dark web forums. The vast majority of these forums are filled with real CSAM (though some are discussion-only, and prohibit the posting of CSAM).

Nonetheless, a number of long-standing forums have this year added new AI sections to their sites – and these sections are growing in popularity.

In these areas, perpetrators share advice on generating AI CSAM; request bespoke images or models; share their work and offer feedback on others' work. These discussions take place openly as users feel anonymous, believing that law enforcement is unwilling or unable to discover them. Users share advice on connecting to these forums in secure, untraceable ways.

Where users share AI CSAM images in bulk on these forums, these files are sometimes hosted on the dark web, and at other times hosted on the open web according to a pre-approved list of secure, anonymous clear web file hosts.

## Forum snapshot study: September 2023

This report includes a snapshot study of one dark web CSAM forum. All the AI-generated images posted to the forum in a one-month period (September 2023) were identified.

This encompasses both threads made in September (threads just about AI CSAM, for example, total 261,920 views) and threads made earlier but to which further images were posted in September.

In total, **20,254 AI-generated images were found to have been posted to this forum in a one-month period**.

Of these, **11,108 images were selected for assessment by IWF analysts**. These were the images that were judged most likely to be criminal.

(The remaining 9,146 AI-generated images either did not contain children or contained children but were clearly non-criminal in nature.)

**12 IWF analysts dedicated a combined total of 87.5 hours** to assessing these 11,108 AI-generated images.

Any images assessed as criminal were criminal under one of two UK laws, as described in section 5. These are:

- **The Protection of Children Act 1978** (as amended by the Criminal Justice and Public Order Act 1994). This law criminalises the taking, distribution and possession of an "indecent photograph or pseudo-photograph of a child".
- The **Coroners and Justice Act 2009.** This law criminalises the possession of "a prohibited image of a child". These are non-photographic – generally cartoons, drawings, animations or similar.

2,562 images were assessed as criminal pseudo photographs, and 416 assessed as criminal prohibited images.

**AI images assessed** 5 - 10 October 2023

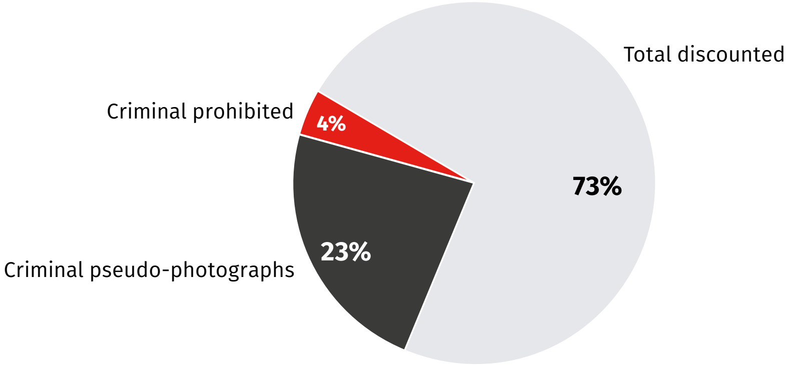


**Figure 9**
AI images assessed between 5 and 10 October 2023

Source: IWF Analysis

These are shown as a proportion of the 11,108 images assessed by IWF analysts as follows:

The total proportion of images assessed as criminal was 27% of the 11,108 images assessed.

Of the criminal images, **six times as many images were assessed as realistic pseudo-photographs than were assessed as non-realistic prohibited images.**

Those images assessed as criminal pseudo-photographs can be sorted by severity (using Sentencing Advisory Panel categories A, B and C) and age:

Category A: images depicting penetrative sexual activity; images involving sexual activity with an animal or sadism.

Category B: Images depicting non-penetrative sexual activity.

Category C: Other indecent images not falling within categories A or B.

**AI images assessed by severity** 5 - 10 October 2023

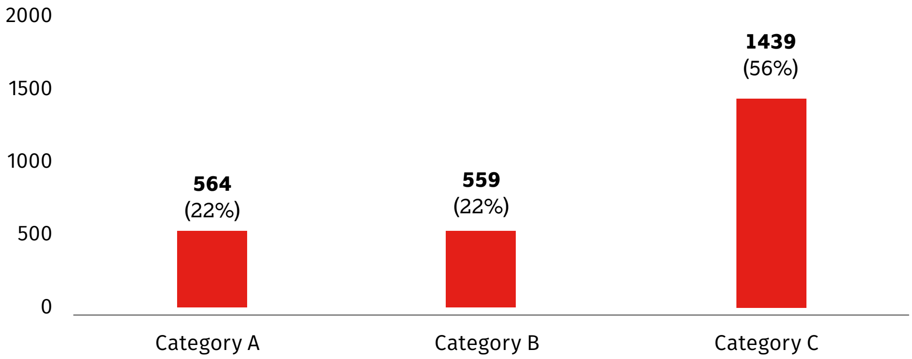


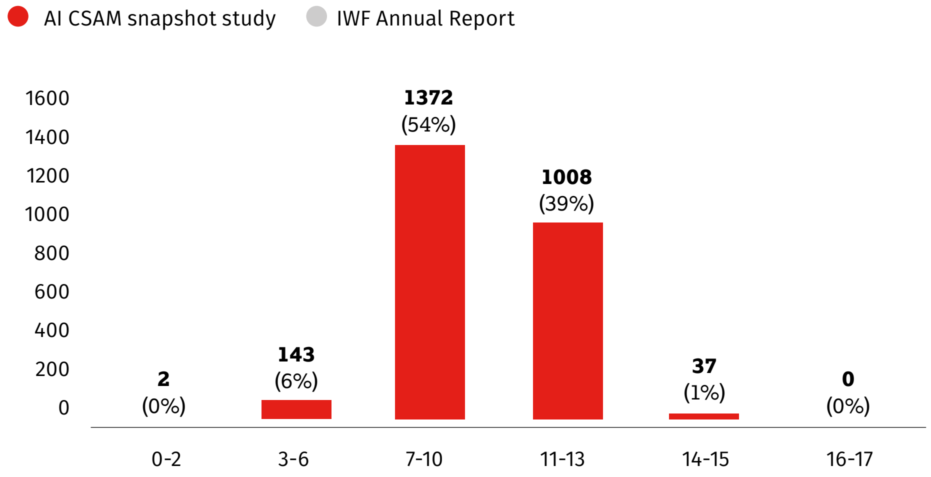
**Figure 10**
AI images assessed by severity between 5 and 10 October 2023

Source: IWF Analysis

**Figure 11**
AI images assessed by age between 5 and 10 October 2023

Source: IWF Analysis

**AI images assessed by age** 5 - 10 October 2023



These statistics show that **most indecent pseudo-photographs found on this forum were category C** – indecent images of children not falling within categories A or B – often, for example, images depicting naked children erotically posing.

These images were most likely to feature **children between 7 and 13 years old.**

**99.6% of these images featured female children.** A variety of ethnicities were observed also.

Those images assessed as prohibited images are not sorted by severity, age, or sex. This reflects the limitations of IWF's remit.

The 73% of the 11,108 images that were assessed by IWF analysts as non-criminal can be sorted as follows:

| **Discounted reason** | | **%** |
|---|---|---|
| None | 0 | 0% |
| Age in Question | 390 | 5% |
| Suspected Adult | 0 | 0% |
| Known Adult | 0 | 0% |
| Extreme Adult Porn | 0 | 0% |
| Child No Sexual Activity | 4340 | 53% |
| Adult No Sexual Activity | 492 | 6% |
| Non-Photographic Imagery | 1634 | 20% |
| Off Remit | 1274 | 16% |
| **Total** | **8130** | **100%** |

The data show that only **20% of assessed discounted images were determined to be not realistic enough** to treat as a pseudo-photograph or a non-photographic image of a child as defined by the Coroners and Justice Act (2009) (whether that AI-generated image depicts a child or an adult). This is roughly in accordance with the proportion of criminal image assessments (between indecent pseudo-photographs and child prohibited images) above.

The 5% of assessed discounted images marked as "Age in Question" reflect cases where the IWF analyst is unsure whether the image depicts a child or an adult – images of older, post-pubescent teenagers, for example. Outside the CSAM world, perpetrators often push the boundaries of generation of adult pornography, behaviour that reflects the wider landscape of online pornography (with its common categories "teens" and "barely legal"). Of course, there is less reason for users of CSAM dark web forums to want to produce these borderline generations, and this is reflected in the low (5%) proportion of "Age in Question" images.

It is notable also that **most assessed discounted images (53%) were images of AI-generated children.** As far as law enforcement is concerned, most or almost all of these would **likely fall into category 6** ("indicative/borderline/notable images") – nudist, naked or semi-naked images of children that have legitimate settings or do not meet the threshold for indecency for their assessment as criminal. These images, however, do not meet the threshold for IWF to take action against them.

This snapshot study had necessary limitations:

- Only one CSAM forum was surveyed.
- The forum surveyed has a general preference towards 'softcore' imagery, and imagery of girls.
- The AI sections of this forum has a number of regular 'creators', and so large batches of images assessed originate from the same few perpetrators.

Further study would test whether the key findings of this snapshot study would be replicated in, for example, a forum that leaned towards images of boys, or 'hardcore' CSAM.

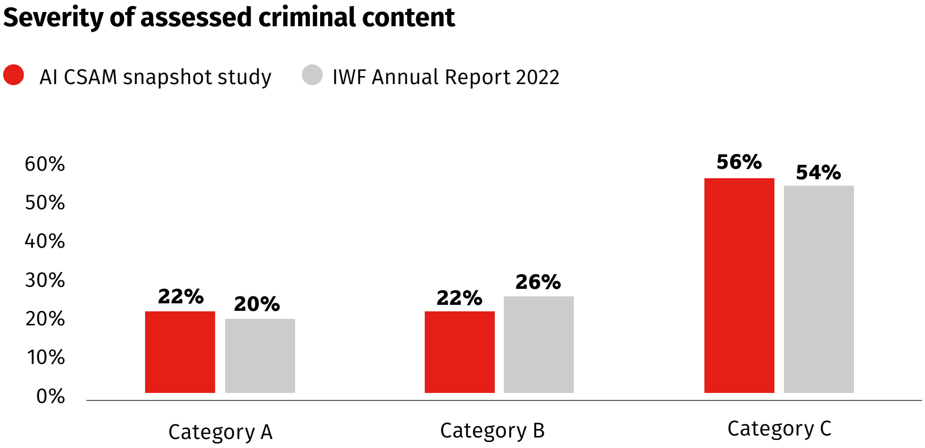
Nonetheless, these key findings are summarised as follows:

- Most **AI-generated images assessed were realistic enough** to meet the realism threshold for assessment as a pseudo-photograph of a child (if criminal). This finding holds across both assessed criminal and assessed non-criminal images. The high level of realism seen among images found on the forum is owing to a number of factors, as discussed in section 5.
- Most **AI-generated images assessed were not criminal.** This reflects a large appetite for images of children outside scenarios containing explicit sexual activity (Sentencing Advisory Panel categories A-C).

- AI CSAM was found **to reflect the wider CSAM landscape.** Some analyses of AI CSAM suggest that AI imagery tends towards the most extreme categories of content, and the youngest ages of children. While this may be true for some areas of the internet, this study found that category and age assessments of criminal AI-generated pseudo-photographs of children were closely correlated with assessments of CSAM across the internet as a whole. This can be most clearly demonstrated by comparing the category assessments from this snapshot study to the category assessments reported by the IWF last year for content found on the internet as a whole:

**Figure 12**
Severity of assessed criminal content

Source: IWF Analysis



# AI CSAM: issues

Perpetrators can turn to AI CSAM over real CSAM for many reasons. One dark web forum user provides a typical view in listing five benefits:

> *1. Users can create images on their own devices;*
>
> *2. Everything can be custom-made and edited to specification;*
>
> *3. AI CSAM can show what is impossible in the real world;*
>
> *4. It is very secure;*
>
> *5. AI is always getting better.*

To some extent, all these points are true.

## Realism

**The level of realism of AI CSAM varies between individuals who generate the content and between image sets. Technical expertise, computer size, and time invested are all variables that affect the level of realism.**

Images of simpler composition, generally showing just one (child) character, are more likely to look photorealistic – there is a greater likelihood of AI artefacts appearing in images with multiple characters involved in complex activity. Nonetheless, the abundance of post-generation editing tools available means that this aspect can be overcome with enough technical knowledge and time investment.

For all these reasons, contrasting reactions can be found across forums where AI CSAM is shared. Some are dismissive:

> *"I can recognize AI work at a glance"*

Others express surprise or admiration at the quality of the output:

> *"It's been a few months since I've checked boy AI. My God it's gotten really good!"*

> *"These are truly stunning. Some of the realism in these is about 95% of the way to indistinguishable from real photos."*

> *"How the hell can you get this kind of images? I've seen realistic images but this is superb."*

> *"The AI generated images are getting better and better."*

> *"The photorealism here is stunning, I mean I'm sure a trained eye can still see it's a generated image, but not by much."*

Others question whether they are looking at real images, or claim photorealistic AI-generated images:

> *"I just can't get my head around that these boys are not real!"*

> *"Are you sure these are CGI?"*

> *"Congratulations on hitting photo realism. This is insane. How much better does it get from here?"*

Examples of individuals asking, 'who is this?' in reply to images – only to be informed that the images are not of real children – have been found.

Forum users' exposure to CSAM and to AI CSAM varies, as does their inclination (or disinclination) towards both. IWF analysts, on the other hand, are exposed to CSAM every day. They are trained to recognise and assess CSAM, and in recent months have been trained also to assess AI CSAM.

For this report, IWF analysts assessed thousands of AI-generated images, and provided comments on what they thought of the level of realism. Their comments provide a useful perspective on comparing AI CSAM to real CSAM; on assessment of AI CSAM; and on the future outlook for those who work to fight child sexual abuse, including law enforcement. Some have stronger conclusions on realism than others.

### ANALYST 1

***"We have a good idea of the common glitches and features of AI-generated images. Armed with that knowledge and assessing images that I know are AI-generated, there are still images that I would struggle to distinguish from real photos. Near flawless, photo-realistic pictures of the worst kind of child abuse you can image. And this is with AI in its infancy."***

### ANALYST 2

***"Currently AI generated images are quite simple to spot as the tells such as extra fingers, lighting, etc. are still quite prominent, however I think that the quality has improved very quickly over a short period and would confuse the general public."***

ANALYST 3

***"I've been both surprised and disappointed to see how much attention and dedication has been taken to create such life-like abusive images of children."***

ANALYST 4

***"Some were scarily realistic, and the whole thing just made me feel a bit uneasy to be honest… I am also concerned that future images may be of such good quality that we won't even notice."***

Further thoughts from analysts on making assessments of AI CSAM are included in section 11.

## Satisfaction

How realistic AI CSAM looks is closely associated with discussions among perpetrators about whether AI CSAM satisfies their desires.

In some areas, AI CSAM is shared alongside real CSAM (a 'mixed' environment). In these areas, the preference is for realistic imagery:

> *"less they look cartoon the better."*

> *"Please more of the Ultra Realistic Stuff. Amazing."*

> *"Really cartoonish or unrealistic AI images don't really do it for me, but they are getting better. The image in this post is excellent!"*

Nonetheless, there remains parts of CSAM communities (indeed, perhaps the vast majority of CSAM communities) to whom AI CSAM is not interesting, or does not match real CSAM. These individuals may only be attracted to real abuse.

> *"AI is completely uninteresting to me."*

> *"…it's nothing to the real thing."*

Some users abuse others for posting AI-generated images. One AI CSAM perpetrator says:

> *"I don't get the abuse, I like to make these and ok they are not perfect but they are nice and fun and look good."*

Whether such opinions hold where AI CSAM is not photorealistic, or whether they hold in principle – no matter the appearance of AI-generated images – is questionable. A stronger anti-AI CSAM opinion holds:

> *"They will be saved by people and dilute the stock of real pictures. Do this for five years and fakes is all we will have left! AND THEY WILL STILL LOCK YOU UP FOR THEM."*

Some users claim that AI CSAM will never match real CSAM because it lacks the

> *"sense of danger."*

Other AI CSAM perpetrators disagree. One comment even claimed that AI CSAM was superior because real CSAM images are often low-quality:

> *"Most pics don't come anywhere close to the quality produced by AI. Poorly composed, poorly exposed, low res, and out of focus pics are common."*

Another user wanted to produce lower-quality AI-generated images in order to make them more like real CSAM:

> *"By default AIs weren't teach that reality is sick & dirty, because people want beautiful and perfect pictures so they instruct them that way. But when it comes to the sexy, something is missing: sick & dirty is part of it! That's why I'm not satisfied and try to find ways to get more realistic rendering."*

Because AI-generated images have by default a high-quality, 'clean' appearance, images produced according to this kind of opinion – images that appear lower-quality, 'grainier' or less clear – may pose a key challenge for future AI detection efforts.

## Ethicality and legality

Many individuals claim that AI CSAM is more ethical than real CSAM and use this claim as a justification for generating and posting AI CSAM.

> *"the future of CP is already here... and not offending anyone..."*

Others emphasise that their AI CSAM images were generated without CSAM fine-tuned models, perhaps as an effort to legitimise those generations:

> *"All of these were created without any real world child porn whatsoever."*

At an extreme end, some perpetrators claim that AI-generated images comprise the future of CSAM – eventually replacing the need for real CSAM:

> *"[AI CSAM] makes CSAM unthinkable. Anyone who might before have justified needing CSAM in order to quell some irresistible urge will have no more excuses."*

Representative of such claims is the widespread use of disclaimers among perpetrators posting AI CSAM both on the clear web and dark web:

> *"All images are A.I. generated and none existed before I entered the very specific text prompts I entered to create it"*

> *"Disclaimer: none of the boys I'll post in this thread are real, they are all generated by an AI"*

The reason for these disclaimers is unclear, and may vary between perpetrators. They may intend these disclaimers to discourage takedown efforts from site owners or investigation by law enforcement, or they may be for information purposes only.

Users discussed how law enforcement agency (LEA) officers could generate and use AI CSAM, and how they could use AI against LEAs:

> *"For now AI images can be spotted in most cases, but it is getting to the point where AI will be indistinguishable. I think there are opportunities for AI to be used to our advantage in playing an activist role."*

As the latter comment suggests, some users discuss sharing AI-generated images with non-perpetrators as an intended 'gateway' to real CSAM.

## Perpetrator pathways

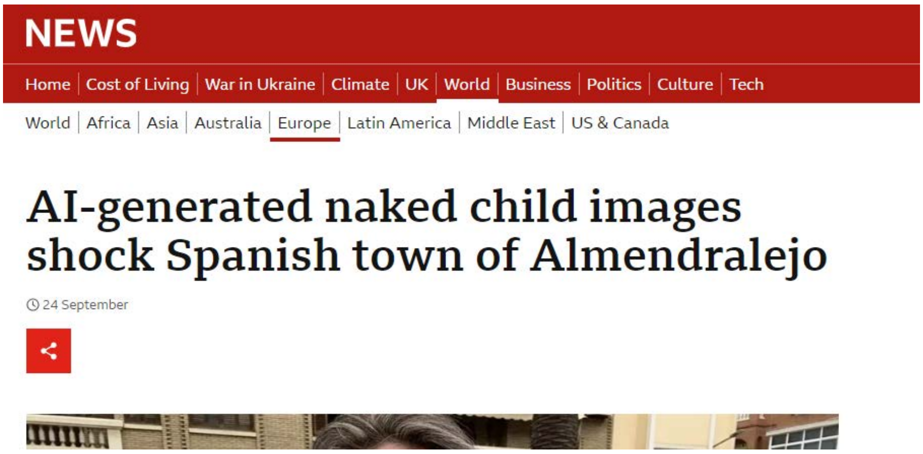
Reflecting IWF's remit, concrete knowledge of pathways of perpetration or offending is elusive, but anecdotal evidence encountered during research for this report suggests that movement between AI-generated non-photographic images and AI-generated pseudo-photographic images is possible, and, therefore, between AI CSAM and real CSAM in principle.

> *"I've mostly been using AI to generate smut (haven't we all) and recently moved from semi-realistic [cartoon-style] content, which I had some great successes with, to trying to generate photoreal children."*

Worth re-emphasising in this section is that though some perpetrators use complex models and fine-tuned CSAM models, for others, simple, easy-to-use models exist and can be used to generate AI CSAM. One perpetrator describes:

> *"A picture of a couple random kiddies I seen playing outside at the park one day I went and made it a fake"*

The prevalence of tools like those for 'nudifying' images makes such low-effort perpetration possible. Recent news reports detailed the use of this technology by children, applied to images of children, among students at a school in Spain and another school in Denmark.



NEWS

Home | Cost of Living | War in Ukraine | Climate | UK | World | Business | Politics | Culture | Tech

World | Africa | Asia | Australia | Europe | Latin America | Middle East | US & Canada

AI-generated naked child images shock Spanish town of Almendralejo

24 September

Figure 13
Recent BBC News article on 'nudification' technology spread in a school in Spain

Source: BBC News
Author's screenshot

The IWF has already seen some examples of self-reporters claiming that their images had been turned into sexual images with AI (possibly using these 'nudifying' tools), and is aware of a number of websites that are claimed to have been used for this purpose.

Clearly, the ease and accessibility of some text-to-image technology increases the risk among the pool of potential perpetrators – a major area of concern both in terms of isolated incidents and in terms of targeted 'sextortion'.

## Guides to generating AI CSAM

**Section 7 described the two main complete guides so far discovered circulating on dark web forums: a guide for creating CSAM images, identified in March 2023 (such is the pace of technological progress, whose contents should now be considered outdated); and a guide for creating new models, identified in May 2023. Other small or one-page guides concerning, for example, online web-based models have also been found.**

SEE RECOMMENDATION #2

For Government

Just a few guides have so far been identified, though it is possible that some are in circulation of which IWF is not aware.

Why are there not more guides in circulation? A huge amount of information on generating AI CSAM is shared, but not as cohesive guides. Instead, these are individual posts on forums. Of course, a high amount of effort is required to write, edit, and share these documents, so perhaps their low frequency is unsurprising.

Whether more guides appear in the coming months remains to be seen. Perhaps the instability caused by the current pace of technological progress in this area is incompatible with production of long, detailed guides on using text-to-image AI, and more guides will appear when the technology stabilises. Alternatively, the time and effort barrier to the production of guides – which have a highly technical subject matter – may mean that not very many guides will ever appear.

Nonetheless, guides do exist – and their legality is discussed in section 11.

## Commerciality

The first examples of commercialisation of AI CSAM have been recorded over the past months. Examples are limited because the IWF is often prevented from further investigation by direction to end-to-end encrypted chats or peer-to-peer networks; the IWF is also unable to purchase, or attempt to purchase, AI CSAM. Intelligence collected by law enforcement partners in publicly inaccessible areas would support examples provided in this section.

- **Commercial example 1**

An example of the barriers IWF face can be found in a page called [redacted]. The account posted non-actionable AI-generated images of children – in various modelling poses, settings, and outfits – and linked to a website and another page through which people could purchase access to more content. The brand claims that it is

```
selling custom AI images
```

but the authenticity of this claim cannot be verified. Interested individuals have to

```
hit up my [name] on secret chat
```

to access this hidden content. Whether hidden content comprises just further child modelling images or also AI CSAM is impossible for IWF to know at present.

- **Commercial example 2**

Another commercial page showcased non-actionable images and GIFs of AI-generated children (including clothed; posed suggestively; in bikinis), with further content locked behind a paywall. Previews showed high-quality AI-generated content apparently produced by:

> Harnessing the power of the latest AI technology and [a high-end computer graphics card]

Supporters can apparently view more than 2,000 images and request 10 new images monthly for the price of 5,000¥ per month. Again, no criminal content was accessible and so whether only non-actionable images are found behind the paywall, or whether AI CSAM is being sold and bought, is impossible for IWF to know at present.

- **Commercial example 3**

This commercial example is associated with a website actioned by IWF as containing criminal preview AI CSAM material.

An account advertised

> NSFW ai generated pictures

including bespoke content, generated to specification. Individuals could subscribe from between £4.50/month to £17/month. At the time of assessment, the account had 70 total paid members, generating $316.50/month for the page operator. (This account has since been closed.)

- **Commercial example 4**

A wide-ranging commercial brand, again linked to actioned preview AI CSAM content offered customers:

> Get access to exclusive, photo-realistic arts (1000+ for now). With regular updates and unique personal models. (requests accepted)

The access price was $7/month and promised more content of naked AI-generated children behind this paywall.

**These four examples show the demand for high-quality AI-generated images of children and AI CSAM made to specification. They demonstrate the growth of brands that advertise this service and are careful about hiding most or all criminal content behind a payment barrier. It is also notable that these services are international – perpetrators may reside in the UK or in other territories in which AI CSAM has a different legal status.**

SEE RECOMMENDATION #1

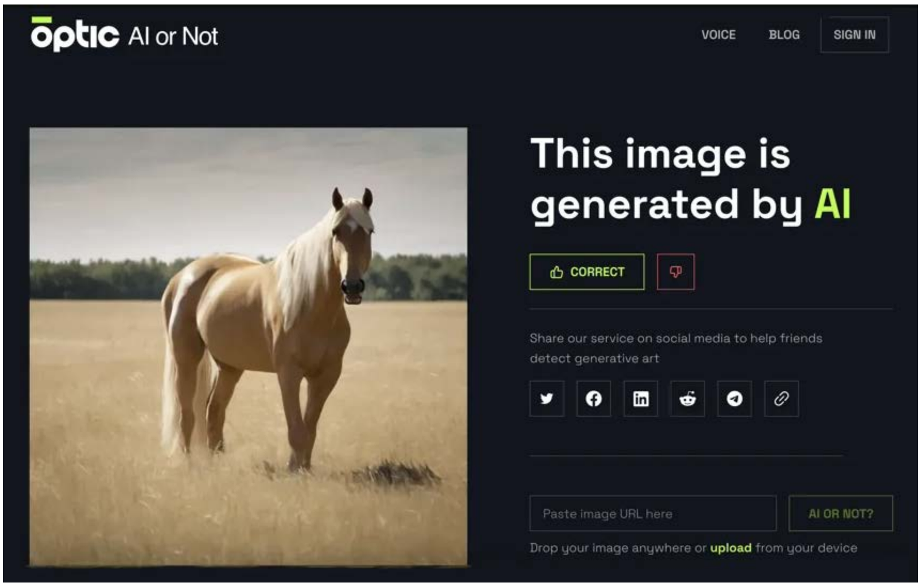
For Government

# Detection and enforcement

## AI image detection

The goal of detecting AI-generated media has broad relevance – not just in the CSAM domain, but also for law enforcement more broadly, implications for identifying cases of misinformation and disinformation, and democracy as a whole.

There is growing pressure for standards for digital watermarking of AI-generated images; there is also an increasing industry for accurate AI image detection. Tools can take the form of online services or downloadable software.



This image is generated by AI

Share our service on social media to help friends detect generative art

Paste image URL here

Drop your image anywhere or upload from your device

Figure 14
Optic's website 'AI or not' aims to answer this question for each image uploaded to the site

Source: PCMag

In the CSAM domain, US nonprofit Thorn have produced an open-source detection tool, and claim high accuracy.

Concerns exist that a race is beginning between AI classifiers and those training text-to-image models to evade AI classifiers. This represents a major barrier to the development of a classifier that is 100% effective.

This race is possible because a favoured model being used to create AI CSAM is open-source technology. Data that comprises an AI watermark can be added to real images; the watermark can be removed from AI-generated images.

The first example of a perpetrator claiming that real images were AI-generated has been identified by law enforcement; the provenance of these images was discovered through a check against the UK's national Child Abuse Image Database (CAID).

Reasons for this type of perpetrator activity are mostly speculative – only anecdotal evidence exists. One discussion found between users of a dark web forum concerned whether or not perpetrators should 'sign' their own AI CSAM (as an 'AI artist'), and whether this would help law enforcement agencies (LEAs). At the very least, this indicates that perpetrators are aware of some of the difficulties for law enforcement:

> *"I disagree, by signing your work you make LE job easier. If they know an author is creating AI images then they can ignore it. If an image looks real they have to spend time 'looking' into it, which wastes their resources."*

Of course, a classifier does not need to be 100% effective to be useful to IWF analysts or to law enforcement. A more moderate goal revolves around provision of a probabilistic tag-and-flag tool, allowing law enforcement, for example, to prioritise cases in which it is more likely that there is a child that needs protecting. Further discussion about AI CSAM and LEA is found later in this section.

SEE RECOMMENDATION #8

For Technology Companies

**The current prevailing view seems to be that future AI image detection will require a suite of classifiers, perhaps targeted at detecting the output of different text-to-image models, whose analyses will in conjunction indicate the likelihood of an image having been AI-generated.**

## Model types and model data

As outlined in section 4, a popular base model is trained on subsets of an open-source dataset, which is an index of URLs that identifies billions of images. Earlier versions of this base model contain adult pornography, and base models after a certain point exclude adult pornography. This does not completely prevent generation of adult pornography with the later models but does make users more likely to use earlier models for the generation of adult pornography – and for the generation of CSAM.

In this context, the company behind this base model must perform some filtering within the dataset to create various versions of the base model. How exactly it performs this filtering – whether manual, automatic, or a combination of both – is unknown.

The key question is: given a diffusion model, what (if anything) can be known about the training dataset?

These questions comprise an active area of study among the AI research community. For example, a 2023 paper asked whether training data could be extracted from diffusion models by generating images, then performing membership inference to identify 'memorised' images. On model memorisation, the authors note:

> *"This paper covers a very restricted definition of "memorization": whether diffusion models can be induced to generate near-copies of some training examples when prompted with appropriate instructions. We will describe an approach that can generate images that are close approximations of some training images (especially images that are frequently represented in the training dataset through duplication or other means). There is active discussion within the technical and legal communities about whether the presence of this type of "memorization" suggests that generative neural networks "contain" their training data."*

Researchers found significant rates of memorisation in large diffusion models, a finding that could be used in a limited way to 'reconstruct' images within a training dataset by testing and evaluating the output of diffusion models.

Such methods apply to large-scale diffusion models, but what about smaller fine-tuned models? Could the training dataset for a CSAM fine-tuned model be reconstructed, or any information at all be deduced, from the CSAM fine-tuned model only?

These questions require further study.

## AI CSAM, victim identification, and law enforcement

As mentioned earlier in this section, the major challenge for law enforcement posed by AI CSAM is of distinguishing photorealistic AI CSAM from real CSAM – victim identification (VID). This challenge is to be addressed through technological solutions – tools like AI classifiers, as described above – and advanced digital forensic knowledge among investigators.

SEE RECOMMENDATION #4

For Law Enforcement and Regulators

**LEAs will increasingly be required to train their investigators on recognising AI-generated images; investigating, assessing, and tagging them appropriately (considering, for example, the impact of uploading AI-generated images to the UK's Child Abuse Image Database (CAID)).**

There is a low risk for IWF of false positive victim identification referrals – referring virtual children to law enforcement for investigation. This is because IWF analysts require a baseline amount of identifying information to make a referral – the sort of identifying information (name; school; region or area; or similar) that would be expected to be missing from AI-generated images.

Nonetheless, because image generation is fast and accessible (especially for low-tech perpetrators), as the examples in section 9 of this report show, this technology increases the pool of potential victims of child sexual abuse.

As stated in the introduction to this report, generative AI more broadly has the potential for misuse in CSAM and CSE/A offending. This includes misuse of LLMs and chatbots, and use of text-to-image technology to generate child avatars, for example, or other images that increase a perpetrator's repute among potential victims. These uses merit further investigation but fall somewhat out of the scope of this report.

SEE RECOMMENDATION #4

For Law Enforcement and Regulators

**The speed and scale of potential AI CSAM generation should concern LEAs – faced, perhaps, with investigating seized devices that contain vast amounts of AI CSAM generated offline, on-device – as well as IWF – faced with the potential for the spread of vast amounts of AI CSAM across the internet.**

SEE RECOMMENDATION #5

For Law Enforcement and Regulators

**Finally, the lack of oversight inherent in the open-source technology that is overwhelmingly favoured by AI CSAM perpetrators should concern LEAs. There are few areas for oversight, detection, or intervention – and a clearly conceivable route of offending from realistic AI-generated CSAM to real CSAM.**

# UK legislation

## Assessments

The key feature of AI CSAM from an assessment perspective is that, as described in section 5, it straddles two pieces of legislation. This entails an often-difficult decision for those responsible for classifying images: whether an image is realistic enough to classify as an indecent pseudo-photograph, or whether it should be assessed as a prohibited image of a child. (The latter piece of legislation has more criteria and carries a shorter sentence for convicted offenders. The question of what position a judge would take on a conviction on AI-generated indecent pseudo-photographs of children remains to be tested in UK courts. A lesser sentence is a potentiality, as a judge may conclude that no harm has been done to an actual child.)

Assessment difficulties were widely reported by IWF analysts, who assessed thousands of AI-generated images for this report and hashed them – translated those images into code that can be used to identify and remove those images in the future. Such difficulties entail significant time investment. Some of these comments are reported below.

### ANALYST 3

*"It can feel odd to question the realness of something that you know isn't real. It brings into focus the different laws our work is bound by."*

### ANALYST 5

*"Hashing these images were difficult, more difficult than 'normal' CSAM. Is it a child? Is it actionable? Is it photorealistic? Is the actionable content you are seeing feasible even?"*

*"There were a lot of digitally anatomical disturbing images of all sorts which take a lot longer to work out and grade."*

### ANALYST 6

*"There are some very weird concoctions which are more difficult to grade, and other opinions are sought to help with them."*

**An argument could be made that the emergence of AI CSAM combined with the difficulty of sorting between laws and categories – indeed, depending on individual assessors' determinations of whether an image "appears to be a photograph" – merits combining existing CSAM legislation.**

SEE RECOMMENDATION #2

For Government

Certainly, this would dissolve the difficulties of assessing AI CSAM, but it would create new difficulties.

International mapping remains a problem. As discussed in section 8, AI CSAM has different legal statuses in different jurisdictions.

**Recommending specific change(s) to this existing legislation is beyond the scope of this report, but a future review in this area, including input from law enforcement partners, should inform any potential recommendations.**

SEE RECOMMENDATION #2

For Government

**As of September 2023, AI-generated imagery is not yet part of the College of Policing grading training course – organisations working in this area must work to come to terms with the problem fully before solutions are considered.**

RECOMMENDATION #4

For Law Enforcement and Regulators

## Guides

Section 9 of this report briefly set out the phenomenon of guides to generating AI CSAM, shared in dark web forums, and covering topics like generating and editing images, and training CSAM fine-tuned models.

Serious Crime Act 2015 created the following offence as part of Section 69, "Possession of paedophile manual":

> It is an offence to be in possession of any item that contains advice or guidance about abusing children sexually.

But "abusing children sexually" means doing anything that constitutes an offence under Part 1 of the Sexual Offences Act 2003 or:

> (b) an offence under section 1 of the Protection of Children Act 1978, or under Article 3 of the Protection of Children (Northern Ireland) Order 1978, involving indecent photographs (but not pseudo-photographs).

The result is that guides to the generation of AI CSAM are not covered by Section 69 of the Serious Crime Act 2015.

**Nonetheless, there is an argument that this falls into the 'low priority but easy to solve' category, wherein guides to generating or 'creating' indecent pseudo-photographs of children could be added to the definition set out in the legislation above.**

SEE RECOMMENDATION #2

For Government

## Models

Articulation of an offence that criminalises CSAM fine-tuned models is difficult, and it could be argued that possessing (or even creating) such models comprises only a preparatory act, not a criminal one. Furthermore, technical questions about how to prove that a given model has been fine-tuned using a CSAM dataset, and is intended for the generation of AI CSAM, remain.

Not to criminalise these models risks their widespread distribution across CSAM communities – at present, any perpetrator can download everything that they need to generate (offline, undetected) as many images of known victims of child sexual abuse as they without committing any offence. Clearly, the other items have widespread legitimate uses, but it is difficult to argue that the final item does have any legitimate use.

As such, it is unclear whether the IWF nor any regulator or law enforcement body currently has recourse to request removal of CSAM fine-tuned models from legitimate model-sharing sites, for example.

**Two possible routes exist: firstly, where the model has been shared alongside criminal preview images; secondly, and more tentatively, where provision of the model can be judged to constitute an offence under the Serious Crime Act 2007 ("encouraging" an offence to be committed).**

SEE RECOMMENDATION #7

For Technology Companies

This report makes no conclusion on whether it is desirable or even possible to reconcile the contradictory positions on this issue.

# Summary: past, present, and future of AI CSAM

Progress in computer technologies, including progress in generative AI, has enormous potential to better our lives, and misuse of this technology is a small part of this picture.

The development of computer technologies like the growth of the internet, the spread of video-calling and livestreaming, and the development of CGI and image-editing programs, have enabled the widespread production and distribution of CSAM that is currently in evidence.

It is too early to know whether generative AI should be added to the list above as a notable technology that comprises a step change in the history of the production and distribution of CSAM.

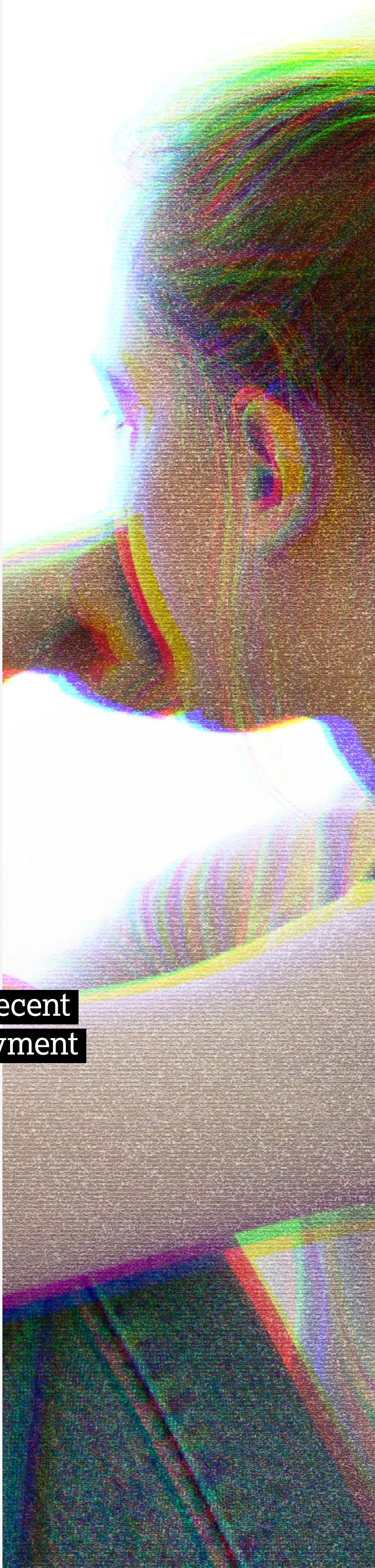
Nonetheless, this report evidences a growing problem that boasts several key differences from previous technologies. Chief among those differences is the potential for offline generation of images at scale – with the clear potential to overwhelm those working to fight online child sexual abuse and divert significant resources from real CSAM towards AI CSAM.

In this context, it is worth re-emphasising that this is the worst, in terms of image quality, that AI technology will ever be. Generative AI only surfaced in the public consciousness in the past year; a consideration of what it will look like in another year – or, indeed, five years – should give pause.

At some point on this timeline, **realistic full-motion video content will become commonplace. The first examples of short AI CSAM videos have already been seen – these are only going to get more realistic and more widespread.**

Solving some of the problems posed by AI-generated indecent images now will be necessary to create models for deployment against the growth of video content in the future.

For further information on this report, please email **media@iwf.org.uk**



# Glossary

**Actionable *(image)*:** an actionable image is one that is deemed criminal under UK law and therefore IWF can seek its removal from the internet.

**AGI:** *artificial general intelligence.*

**AI:** *artificial intelligence.*

**AI CSAM:** child sexual abuse material that has been generated or edited by artificial intelligence.

**Base Model (or Foundation Model):** an AI model, generally those released directly by generative AI companies, designed to produce a wide and general variety of outputs.

**Category A:** a classification of child sexual abuse images depicting penetrative sexual activity; images involving sexual activity with an animal or sadism, as according to the Sentencing Council's Sexual Offences Definitive Guideline.

**Category B:** a classification of child sexual abuse images depicting non-penetrative sexual activity, as according to the Sentencing Council's Sexual Offences Definitive Guideline.

**Category C:** a classification of indecent images of children not falling within categories A or B, as according to the Sentencing Council's Sexual Offences Definitive Guideline.

**ChatGPT:** An LLM developed by OpenAI.

**Claude:** An LLM developed by Anthropic.

**CLIP:** *contrastive language-image pre-training.* A neural network trained on hundreds of millions of text/image pairs scraped from the internet.

**Closed-source models:** software whose source code is not released to the public. The public are not able to use, study, change, or distribute the software or its source code to anyone or for any purpose.

**Coroners and Justice Act 2009.** This law criminalises the possession of "a prohibited image of a child". These are non-photographic – generally cartoons, drawings, animations or similar.

**CSAM:** *child sexual abuse material.*

**C2PA:** A mode of metadata representation used for digital watermarking of AI-generated images.

**DALL-E:** A text-to-image model developed by OpenAI, accessed through an API.

**Dark Web:** The side of the World Wide Web that is not indexed by search engines and requires specific configuration, software, or authorization to access allowing users and website operators to remain anonymous or untraceable.

**Deepfakes:** media (images, videos, or audio) that has been digitally manipulated through AI tools or software to replace one person's likeness convincingly with that of another.

**Deep learning:** A type of machine learning, loosely modelled on the human brain, that uses artificial neural networks with more than three layers. These deep learning systems are generally trained on huge datasets scraped from the internet.

**Diffusion Model:** Text-to-image models that add and remove layers of 'noise' to images. Running the 'de-noising' process on random seeds generates 'new' images.

**Fine-tuning:** A type of machine learning model in which the weights of a pre-trained model are trained on new data, and therefore adjusted, to perform a secondary task.

**GANs:** *generative adversarial networks.* A type of machine learning model in which two neural networks compete with each other by using deep learning methods to become more accurate in their predictions. Can be considered the precursor to diffusion models.

**Generative AI:** a type of machine learning that uses deep learning models to identify the patterns and structures within existing data to generate new content.

**ICAP sites:** *invite child abuse pyramid sites.* Sites first reported on by IWF in June 2022 that incentivise users to share links to child sexual abuse webpages far and wide in a 'scattergun' approach.

**Iterative learning:** a type of machine learning guided by a combination of human feedback (supervised machine learning) and/or algorithmic feedback (unsupervised machine learning).

**IPTC:** A mode of metadata representation used for digital watermarking of AI-generated images.

**IWF:** *Internet Watch Foundation.*

**LAION:** An open-source dataset used for training Stable Diffusion which has 2.6 billion English language-tagged images within a 6 billion image dataset.

**LEAs:** *law enforcement agencies.*

**LLaMA:** An LLM developed by Meta.

**LLMs:** *Large Language Models.* A type of machine learning that is trained on huge quantities of text and whose function is to generate text. These models are renowned for their natural language processing abilities, 'understanding' and interpreting human language.

**Midjourney:** A text-to-image model developed by Midjourney, Inc, accessed through the social media site Discord.

**Neural network:** a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain. It creates an adaptive system that computers use to learn from their mistakes and improve continuously.

**NSFW:** *not safe for work.*

**NPI:** *Neural Programmer-Interpreters.* A machine learning model that uses a recurrent and compositional neural network to train machines to carry out simple tasks based on a small amount of training data.

**Open-source models:** software whose source code is released under a license in which the copyright holder grants users the rights to use, study, change, and distribute the software and its source code to anyone and for any purpose.

**Open Web:** The side of the web that is public and viewable by everyone.

**PaLM:** An LLM developed by Google.

**PoC:** *The Protection of Children Act 1978.* This law criminalises the taking, distribution and possession of an "indecent photograph or pseudo-photograph of a child" (as amended by the Criminal Justice and Public Order Act 1994).

**Prompts:** Words or short phrases used to describe what you do (positive prompts) or do not (negative prompts) want to see in the image when using generative text-to-image models.

**Pseudo-photograph:** An image (including one generated by a computer) that appears to be a photograph.

**Real CSAM:** Child sexual abuse material that has not been generated or edited by AI technology.

**SDXL:** A new version (released July 2023) of Stable Diffusion, the text-to-image model developed by Stability AI.

**Self-generated content:** when children are groomed, deceived or extorted into producing sexual images and/or videos of themselves and sharing them online.

**Serious Crime Act 2015, Section 69.** This section created the offence of "possession of a paedophile manual".

**Shallow Fake:** Colloquial term encompassing images produced via simple image editing tools or software, in contrast to deepfakes, which generally use AI tools or software.

**Stability AI:** a global company, headquartered in London, working to make foundational AI technology accessible to all. Responsible for the creation of Stable Diffusion.

**Stable Diffusion:** A text-to-image model developed by Stability AI that can be downloaded from an open-source online community.

**SynthID:** A Google technology that adds a digital watermark to individual pixels in AI-generated images.

**Text-to-image model:** A type of machine learning model whose function is to generate images from text prompts.

**VAEs:** *variational autoencoders.* A generative AI model that uses two neural networks called the encoder and decoder. Generally, these can output images faster than diffusion models, but those images are less detailed.

**VID:** *victim identification.*

**Watermarking/digital watermarking:** a technique that involves embedding digital marks or indicators into machine learning models or datasets to enable their identification.

## DISCLAIMER

*The images used in this report are screenshots of content available on the clear web and dark web. We've attempted to cite the sources of these screenshots, some of which depict likenesses of famous people or films. These likenesses have been generated by someone submitting prompts to AI models. They are not images of the actors or from the films themselves. This goes someway to demonstrate the photorealism of images produced by AI models.*

IWF
Internet
Watch
Foundation

July 2024 | V1.1

# What has changed in the AI CSAM landscape?

AI CSAM REPORT UPDATE

in conjunction with our Oct 23 report

**Prompt:** from fantasy to photo-realistic reality

PUBLIC VERSION

## Content note

Throughout this report update, Child Sexual Abuse Material generated through Artificial Intelligence is referred to as **AI CSAM.**

This report update contains no AI CSAM.

It contains descriptions of the methods used to generate AI CSAM, alongside other verbatim comments from perpetrators.

The verbatim comments from perpetrators are reproduced in the report update as they were typed on screen.

# Table of Contents

IWF

Please click on the IWF logo **'home' button** at the top of each page to navigate back to the contents page.

INTERACTIVE REPORT UPDATE

# Foreword from Susie Hargreaves OBE

**Olivia is a little girl who we have seen grow up online through images uploaded of her abuse at the hands of a sex offender when she was between three and eight-years-old. We told her story in 2018 in our annual report. The updated story I'd like to tell you now, is that after Olivia was rescued all images showing her sexual abuse were deleted. But that's not the case.**

Yes, she was rescued from her abuser, but now new images of Olivia are being created by individuals who want to see her in new, abusive, situations, years after the physical abuse ended. And the AI tools to make this happen are here, now.

> One of our top priorities for a future Government is to get a grip on the impending child sexual abuse crisis online exacerbated by new technologies.

This report sets out the progress made since we last reported on the impact of generative AI in October 2023, and powerfully makes the case for why this must be a top priority.

I am delighted the Online Safety Act finally made it onto the statute books in 2023, hot on the heels of the Digital Services Act in Europe. We have not seen enough progress, however, to deal with the impacts of generative AI technology. Nor have we seen sufficient clarity that these new technologies will be developed with safety in mind or effectively regulated.

It has been almost a year since we sounded the alarm, and this updated report identifies new challenges with generative AI. Technology companies unabashed march to text-to-video creation; the announcements by OpenAI that they are consulting on the possibility of their tools being used to create Not-Safe-For-Work content, and the emergence of offenders using LoRA models of known victims of child sexual abuse are just some of the most pressing challenges we face.

In April 2024, the Government committed to criminalising the creation of deepfakes. Of course, in the child safety space, deepfake child sexual abuse images are already illegal, but the technology used to nudify children isn't. We welcomed this news as it saw two of the largest nudifying apps, which have amassed millions of users, and have been used to create child sexual abuse material, being disabled to public access in the UK. Less welcome, was the fact this is still yet to become law because of the timing of the General Election.

Two other amendments which were due to make it into law as part of the Criminal Justice Bill also fell: one sought to extend the existing offence of possession of a paedophile manual to cover the exchanging of hints and tips on the abuse of generative AI tools. The other proposed tackling AI chatbots which seek to simulate the offence of sexual activity with a child. These were both widely supported in Parliament.

The Data Protection and Digital Information Bill had an amendment tabled to it which sought to make it an offence to train an AI tool on child sexual abuse material or for an AI tool to generate child sexual abuse material. But this also failed to become law.

Whilst it was disappointing to see these legislative developments not make it on to the statute books, we are pleased to see that there is cross-party consensus on tackling these issues. The Labour Party has said it will: "ensure the safe development of AI models, by introducing binding regulation on the handful of companies developing the most powerful AI models and by banning the creation of sexually explicit deepfakes" in its manifesto. The Conservatives also committed at Report Stage during the Criminal Justice Bill to the intention to bring back amendments at a later stage.

This report highlights how desperately the law needs to change to keep pace with technology. It also provides an update on previous recommendations, continues to chart the impact of generative AI on the spread of child sexual abuse and makes further recommendations for an incoming Government.

We will be watching closely to see how industry, regulators and Government respond to the threat, to ensure that the suffering of Olivia, and children like her, is not exacerbated, re-imagined and re-created using AI tools.

**Susie Hargreaves OBE** | IWF CEO

# New (additional) recommendations:

**FOR GOVERNMENT:**

1 That the Government legislates to ensure that paedophile manuals which exchange hints and tips on how to utilise text-to-image based generative AI tools to create child sexual abuse material are made illegal, by extending the existing offence, to cover pseudo images.

2 That the Government legislates to make it an offence to use personal data or digital information to create digital models or files that facilitate the creation of AI or computer-generated child sexual abuse material.

3 That the Government legislates to tackle the rise in generative AI chatbots which simulate the offence of sexual communication with a child.

4 That the Government legislates to ensure nudifying technology is not available to UK based users and encourages other Governments globally to take similar measures.

**Relevant passages which relate to the above recommendations are highlighted throughout this report.**

# Update on previous recommendations:

**FOR GOVERNMENT**

| Previous Recommendation | Progress to date |
|---|---|
| **Explore at the forthcoming AI summit the challenges for dealing with AI CSAM, including the need for international alignment.** | • IWF and Home Office, including then Home Secretary, Suella Braverman, jointly hosted an AI Safety Summit fringe event at Chatham House in London, two days before the AI Safety Summit at Bletchley Park.<br>• 33 NGOs, tech companies, Governments, law enforcement and academics agreed a non-binding pledge to tackle AI generated CSAM at the summit.<br>• G7 communique highlighted the challenges of Artificial Intelligence and a commitment to working together to align internationally.<br>• IWF has presented this research in Japan and the US.<br>• Law change announced in Europe through changes to the Directive, laying down new rules to tackle Child Sexual Abuse. |

| | |
|---|---|
| **Ministry of Justice review of laws ensuring they are fit for the AI age** | • Three amendments tabled to the Criminal Justice Bill to tackle the impacts of AI generated CSAM and nudifying technologies.<br>• One amendment tabled to the Data Protection and Digital Information Bill.<br>• No substantive law changes yet, but commitments from both Labour and the Conservatives to address these issues if they form the next Government.<br>• Labour frontbench spoke in support of this recommendation at Report Stage in the House of Commons on the Criminal Justice Bill. |
| **To consider extension of IWF remit to be able to scrutinise datasets on which these technologies are trained.** | • Work underway with Government departments to see what further support IWF can give. |

**FOR LAW ENFORCEMENT AND REGULATORS**

| Previous Recommendation | Progress to date |
|---|---|
| **For the College of Policing training course to be updated to cover AI CSAM and ensure clear guidance is issued to police graders.** | • The College of Policing is currently liaising with the Crown Prosecution Service before introducing additional guidance so officers can more effectively grade child sexual abuse images in accordance with national guidelines.<br>• The College of Policing also provides a standalone learning product on deep fakes which was released to forces in early 2024. |
| **To ensure there is proper regulatory oversight of AI models before they go to market and ensure mitigations are in place for open-source models with closed source having protections built in.** | • In August 2023, the UK Government published a white paper entitled: **“A pro innovation approach to AI regulation.”** They published their response to this White Paper in February 2024, with the Government concluding: “It will not rush to legislate or implement ‘quick fix’ rules that would soon become outdated or ineffective. Instead, the government’s context-based approach means existing regulators are empowered to address AI risks in a targeted way.”<br>• Looking ahead to manifesto commitments made by political parties at the 2024 UK General Election on Artificial Intelligence:<br>**The Labour Party** has said: “ensure the safe development of AI models, by introducing binding regulation on the handful of companies developing the most powerful AI models and by banning the creation of sexually explicit deepfakes.” |

| | |
|---|---|
| | **The Conservative Party** has said: “The UK is well positioned to spearhead this transformation and is already leading global work on AI safety. Over the last 14 years, the Conservatives have turned the UK into a science and innovation superpower.”<br>Along with a commitment to: “Building on existing responsibilities set out for social media in the Online Safety Act.”<br>**The Liberal Democrat Party** has said they will: “Create a clear, workable and well-resourced cross-sectoral regulatory framework for Artificial Intelligence that:<br>• Promotes innovation while creating certainty for AI users, developers and investors.<br>• Establishes transparency and accountability for AI systems in the public sector.<br>• Ensures the use of personal data and AI is unbiased, transparent and accurate, and respects the privacy of innocent people.”<br>• In Europe, we have seen the European Institutions pass into law the first piece of legislation to regulate Artificial Intelligence. |

FOR TECH COMPANIES

| Previous Recommendation | Progress to date |
|---|---|
| **To ensure that companies using and developing Generative AI and Large Language Models (LLMs), place clearly in their terms and conditions that the use of these technologies to generate child sexual abuse material is prohibited.** | • Stability AI, OpenAI and many other platform’s Terms and Conditions have been clear that the use of their technologies to create child sexual abuse material is prohibited.<br>• We have seen OpenAI announce a consultation into the possibility of its technologies being used to create content that is not safe for work.<br>• LAION, one of the biggest providers of open-source data sets has established a relationship with the Internet Watch Foundation and other child safety organisations.<br>• Stability AI has become the first member from the Artificial Intelligence sector to join the IWF as a Member. |
| **That search services should de-index links to fine-tuned AI models known to be linked to the creation of AI CSAM.** | • Thorn, All Tech is Human and the major developers of AI technology have all committed to a set of voluntary principles to make AI safe by its design. |

# Executive summary

Use of Artificial Intelligence (AI) to generate child sexual abuse material (CSAM) is increasing, and the technology is fast improving.

The dark web child sexual abuse forum surveyed in October 2023 was revisited, and a new analysis found that:

- More criminal AI CSAM images were shared – a total of **3,512 AI CSAM images.**
- **90% of images assessed by IWF analysts were realistic enough to be assessed under the same law as real CSAM.**
- Those images **contained more images in the most severe category of CSAM in the UK** (Category A, which contains penetrative sexual activity, bestiality, or sadism) than in October 2023 – this time, 32% of criminal pseudo-photographs were Category A, indicating that perpetrators are experiencing more success generating complex 'hardcore' scenarios.

**Other findings:**

- **The first AI CSAM videos are now in circulation.** These are mostly partially-synthetic – 'deepfake' – videos, though some primitive fully-synthetic videos also exist.
- The IWF has been encountering **an increasing amount of AI-generated content, including AI CSAM, on the clear web.**
- Extensive evidence for **the sharing of AI models for generating images of specific children,** including known victims of CSAM and famous children, has been identified, and is provided in this report update.

# 3

# Introduction to this report update

In the summer of last year (2023), the Internet Watch Foundation (IWF) first reported that open-source AI models were being widely used to generate CSAM.

A report was compiled and was released in October 2023. It found that perpetrators were able to download – legally – everything needed to generate lifelike images of child sexual abuse, then produce as many of those images as they desired. Generation of AI CSAM took place offline, with no opportunity for detection.

The report found evidence of the sharing of thousands of those images, particularly on the dark web – images that comprised new threats both towards existing victims of child sexual abuse and towards potential new victims of child sexual abuse.

This report update seeks to describe what has changed in the AI CSAM landscape since then. It should be considered an update to the October 2023 report, to be read in conjunction with it.

Since autumn last year, some progress towards highlighting and prioritising child safety in AI development has been made. Collaborative efforts among government, law enforcement, the technology industry and civil society have forged valuable channels of communication, and have begun a process towards recognising that AI left unchecked has the potential to corrode child protection efforts. The first steps have been taken towards urgently-needed preventative and mitigative action.

As with all online safety challenges, this challenge is inherently international. It is encouraging that the UK government has sought to position the country at the forefront of AI safety and regulation in hosting the first international conference on the issue, the AI Safety Summit, last November. The Republic of Korea hosted the 2024 AI Seoul Summit in May.

This report update shows that the pace of AI development has not been slowing, nor has the number of people using AI for criminal purposes decreased. In this context, and in the context of the better, faster, and more accessible tools to generate images and videos, the future continues to hang in the balance.

AI still poses a significant risk to the IWF's mission to remove child sexual abuse material from the internet. It still has the potential to overwhelm

resources and cause irreparable harm to children. But the right decisions made now – to necessitate safety by design, to ensure rigorous testing of all AI models released to the public, and to put protection of children before pursuit of profit – can mitigate these problems for years to come.

## Notes on terminology

As in the October 2023 report, this update uses the term 'AI CSAM' to refer to criminal images or videos of the sexual abuse of children that are generated or edited by AI technology, and 'real CSAM' to clearly distinguish CSAM that is not generated or edited by AI technology.

The term 'deepfake' is used variously in the AI field, in the media, and among the wider population. Sometimes it is taken to refer to all AI-generated or AI-edited content. This report uses the term 'deepfake' to refer to *partially-synthetic* content: edited content that is based on a real image or video but has been altered using AI technology. This is particularly important in the context of 'deepfake videos' – in this report update, edited (or 'faked') real videos – which should be clearly distinguished from fully-synthetic videos created by text-to-video or text-to-image-to-video.

## Outline and guide for readers

Section 4 of this report update tracks shifts in the use of newer, higher-quality versions of open-source image-generating tools; notes the progression of AI-generated video content, including AI CSAM video; and details evidence for the spread of AI CSAM across the clear web.

The October 2023 report included a study of a dark web CSAM forum, in which all the AI-generated images posted to the forum in a one-month period were scraped and assessed against UK law. In section 5, this forum is revisited, and a new scrape is completed. This allows for a comparative analysis, which asks whether there have been changes in the type or quality of the imagery shared.

**Newly for this report update, available metadata is scraped from these images.** This metadata is analysed to assess how far the methods used to generate those images can be deduced.

The IWF extends its thanks to Camera Forensics for their assistance on image metadata analysis.

**For further information on the IWF and its remit, see the October 2023 report.**

4

# Trends monitoring

## IWF reports and the clear web

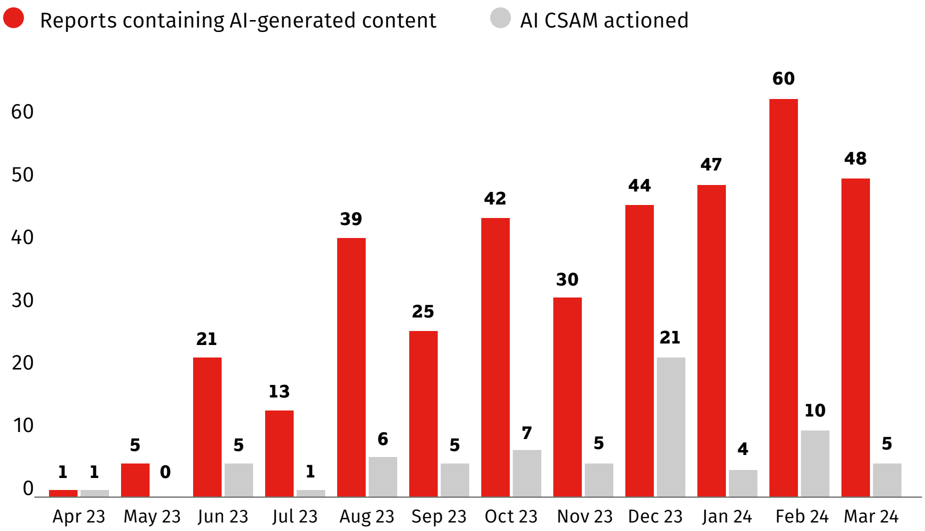
The IWF continues to track reports (generally, webpages) containing AI-generated content, including AI CSAM. Most of these reports are received from members of the public; a small number arise from analysts' proactive searching for content. Most reports from the public continue to contain non-criminal content only.

The graph below shows a general gradual rise in the number of reports where any kind of AI-generated content has been identified by an analyst (in red) alongside a subset of these figures, reports where AI CSAM has been identified and 'actioned' – processed as criminal under UK law (in grey). The IWF seeks to get 'actioned' AI CSAM removed from the internet.

**Figure 1**
Reports to IWF containing AI-generated content rose gradually from April 2023 to March 2024.

Source: IWF analysis

**IWF reports containing AI-generated content, March 2023 to April 2024**



**Reports from April 2023 to March 2024 total 375 reports that contained AI-generated content (and peaked at a high of 60 reports in February of this year) - of which 70 reports contained criminal AI CSAM. These reports almost exclusively point to content hosted on the clear web.** They provide some evidence for the spread of AI-generated content and AI CSAM across the publicly accessible internet.

Notably, we have seen AI CSAM images being shared on commercial sites on the clear web – in place of 'real' images of children. These include on dedicated commercial banner sites, and forum pages with affiliate file-hosting links.

**We also actioned the first UK-hosted webpage containing AI CSAM in March 2024.**

Reports containing AI-generated content continue to comprise a low proportion of total IWF reports. In 2023, this was 220 out of 392,660 reports (0.06%). Of the public reports – likely a better metric to indicate the prevalence of AI-generated content online, or the likelihood of individuals encountering AI-generated content online – this was 201 out of 123,667 (0.16%).

Statistics from the US-based National Centre for Missing and Exploited Children (NCMEC) for 2023 paint a similar picture[1]: generative AI featured in just 4,700 out of over 36,000,000 reports (0.01%). (Note that the comparison is not isomorphic as NCMEC reports are collected and processed in different ways; nonetheless, the low proportion is notable.)

It is fair to state that the rise of AI-generated content over the past year has been gradual and not exponential in nature, and that generative AI – though fast establishing itself within this space – has not yet broken into the 'mainstream' either of adult pornography or of CSAM (insofar as a mainstream exists).

**The October 2023 report highlighted two notable categories of user-friendly online AI pornography services, both of which feature heavily among public reports to IWF: (1) fully-synthetic pornography generation services, and (2) 'nudifying' services.**

Another important category of public reports received related to AI chatbots, which tend to feature a variety of 'characters' with which to interact – often including underage characters. Use of chatbots to simulate conversation with a child is somewhat outside IWF remit, though anecdotal evidence of perpetrators exchanging models, tips and advice has been found.

Use of chatbots in this way has the potential to encourage or normalise harmful behaviour among those with a sexual interest in children. Many of these chatbots are accessible with no age verification process; others incorporate fully customisable (or open-source) characters and have few or no limits to the topics or content of conversation. Evidence of effective regulation in this space is negligible.

SEE RECOMMENDATION #3 & #4

For Government

### 1. AI pornography services

Some barriers to entry to generating images with AI text-to-image models persist: the ability to learn the required technical skills, alongside the possession of some level of computer hardware. (These two important barriers may go some way to explaining the limited spread of AI CSAM so far observed.)

Services for AI pornography aim to neutralise these barriers by making generation fast and intuitive. Users do not have to download or run programs; they simply type or select what they want to see, and the service

[1] www.missingkids.org/content/dam/missingkids/pdfs/2023-CyberTipline-Report.pdf

– which usually uses a built-in foundation AI model version as the image generation 'engine' – provides the images. Evidence of (presumably, low-tech) perpetrators trying and failing to generate AI CSAM on these platforms has been found shared on dark web forums and was included in the October 2023 report. Nonetheless, two 'actioned' reports between April 2023 and March 2024 – webpages containing criminal AI CSAM – relate to two of these AI pornography sites, showing that it is possible for perpetrators to succeed in abusing these services to generate AI CSAM.

### 2. 'Nudifying' services

SEE RECOMMENDATION #4

For Government

**The prevalence of 'nudifying' platforms has been increasing over the past year – indicated by rising public reports to IWF, but also simply by the volume of these sites on the internet. In short, these are sites in which a user uploads an image of a clothed individual; the model outputs an interpretation of the individual without clothes. These comprise one category of 'deepfake' sexually explicit images, soon to become a criminal offence in the UK.**

**Between April 2023 and March 2024, 21 public reports specified dedicated 'nudifying' websites across 15 different domains.**

At the same time, an increasing number of those whom IWF terms 'self-reporters' – members of the public reporting their own explicit imagery – have reported 'fake' content of themselves. **In 2023, the IWF received 17 self-reports from members of the public referencing 'fake' or 'AI-edited' content of themselves. Many of these are children.**

Anecdotal evidence suggests that the perpetrators in these cases are generally people who are unknown to the reporters – whose relationships to the reporters are online-only – but the data is too limited to draw firm conclusions on this point.

Among these cases, the IWF has seen evidence of 'innocent' (non-explicit) imagery of children being taken and 'nudified'.

Sometimes, these 'nudified' images are posted on social media sites with the intention of being shared and seen widely, to cause the victim more distress.

Reports of fakes and deepfakes – many of which are generated using these 'nudifying' services – seem to be closely linked with reports of financial 'sextortion', or blackmail with sexually explicit images. The crux of this point is that perpetrators no longer need to source intimate images from children because images that are convincing enough to be harmful – maybe even as harmful as real images in some cases – can be produced using generative AI.

SEE RECOMMENDATION #1

For Government

**Indeed, one 'paedophile guide' identified by IWF contained a section explicitly encouraging perpetrators to use 'nudifying' tools to generate material to blackmail children. The author of this guide claimed to have successfully blackmailed 13-year-old girls into sending intimate images.**

## AI videos

### Fully-synthetic videos

Recent months have seen notable progress towards fully-synthetic realistic video content in the form of new video generation models, including Stable Video Diffusion (November 2023), a preview of Sora (February 2024) and a preview of Veo (May 2024).

OpenAI's Sora can generate convincing minute-long videos from text, image, or video. It has been released to limited researchers, with a public release planned this year. This type of release, a research-only preview in advance of a full public release, has also been employed for Google's new state-of-the-art video generation model, Veo. These two – crucially, closed-source – models sit at the current frontiers of the video generation industry.

**Figure 2**
A screenshot from a video generated by OpenAI's Sora shows a woolly mammoth walking in the snow.

Source: OpenAI

Stable Video Diffusion can generate short (seconds-long) videos from images. As a Stability AI product, it has been released as an open-source model – available to all under a non-commercial licence. Output is comparable to the closed-source RunwayML Gen-2.

What is the ultimate goal for video-generation companies like RunwayML?

> ***"We've always set the ability to generate a two-hour film as a north star."***

### Text-to-video CSAM

Perpetrators watch the latest advancements with interest. On a dark web forum, AI CSAM perpetrators discuss AI-generated videos:

> *"How long until we can use this new Sora software to make whatever video we want? I want to put my sister's photos in from when she was a kid and make her do nasty things"*

> *"Am seeing the video trailers that were generated by AI, and my mind is blown... The ability to create any child porn we desire... our wildest fantasies... in high definition."*

Limited moving image (GIF) and video CSAM has so far been seen but has been slowly increasing over the past months. Some can be described as deepfakes – for the purposes of this report update, partially-synthetic content – which are discussed later in this section. Fully-synthetic AI CSAM videos are rarer, and are fairly primitive.

One 18-second video, found shared alongside almost 5,000 AI-generated images, shows an adult male penetrating a girl, approximately 10 years old. She is sitting on top of him and looking at the virtual 'camera'. Behind them is a well-lit room with large windows.

The video flickers and glitches; her face and expression morphs from frame-to-frame. Her movement is jerky. Nonetheless, the activity is clear and continuous. It is obvious that the video is synthetic – it doesn't look much like a real video – but this was also the case with images two years ago.

These observations mirror the wider state of AI-generated adult pornography videos: convincing deepfake videos, and primitive (but fast-developing) fully-synthetic videos.

## Technology, tools and models

**Open-source models**

Perpetrators continue to use open-source models as the tool of choice to generate CSAM images because – as described in the October 2023 report – access is offline, on-device; users can use models and prompts freely; and there are few (if any) opportunities for content moderation and criminal content detection or prevention.

There is, however, some evidence that perpetrators are moving away from earlier versions and using more recent image generation models in increasing numbers.

Anecdotal evidence has been found of some perpetrators sharing AI-generated images of children – including AI CSAM – and claiming to have used more recent foundation models only (in other words, no fine-tuned models at all).

An increasing number of fine-tuned models shared in AI CSAM communities are also intended for use with the latest foundation models, resulting in images being generated using these versions. As one user asks another sharing an older model.

> *"Have you considered making a model with [redacted]? In my experience it is generally a big improvement over [redacted]."*

Another user comments:

> *"[redacted] based checkpoints are already at the point where there are some pictures that I wouldn't be able to identify as being AI."*

In the dark web CSAM forum from which images were scraped for this year's study, 35% of (apparently CSAM-trained) models whose links were directly shared were for [redacted]; the remainder were based on older versions. It is, however, notable that it is possible to use those older fine-tuned models with [redacted].

Such evidence of misuse is despite open-source AI models working to implement multiple safety features for their foundation models, including filters for 'unsafe' content. Indeed, AI CSAM perpetrators on dark web forums are dismissive about purported safety features in newer models:

> *"If there is any alignment training inherent to [redacted], large finetunes will be able to override it."*

### Deepfake videos

As set out in the introduction, this report update takes deepfake videos to be partially-synthetic videos – generally, videos *edited* using AI tools to add the face or likeness of another person. It remains the case that the overwhelming proportion of deepfake videos are pornographic in nature. The best deepfake videos are now almost seamless – containing little visual evidence of modification.

A large amount of media discussion concerns abuses of celebrity likenesses in deepfake videos – indeed, these comprise most of the videos on the largest deepfake pornography websites, some of which attract millions of visitors every month. Some of the most-publicised cases involve abuses of celebrity likenesses in pornography; others concern use for misinformation, or for humour or entertainment.

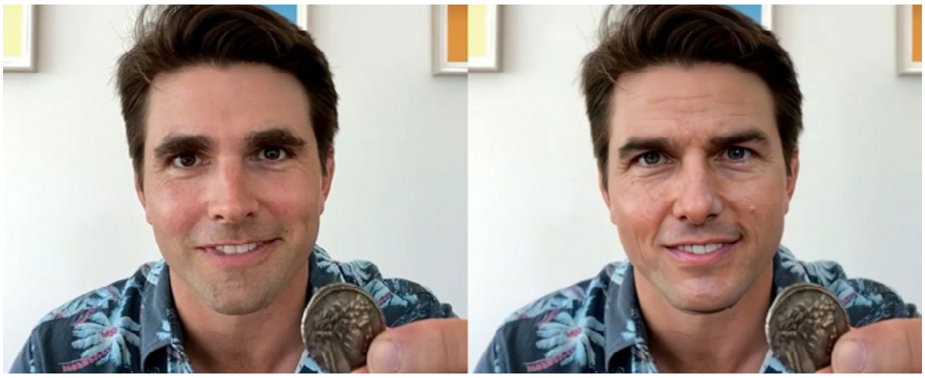


**Figure 3**
'DeepTomCruise' before/after comparison shows the application of deepfake technology to viral video content from 2021.

Source: The Verge

Nonetheless, this same technology can be – and is – applied to less well-known individuals, including non-celebrities, and including children.

**Some deepfake CSAM videos shared in dark web forums take an adult pornography video and add a child's face. Others take existing CSAM videos and add a different child's face to them.** Because the original videos are of real children, and have, therefore, real child anatomical proportions, they can be especially convincing. One impressed forum user says:

> *"I knew about deepfakes... This is so on point! The colours, the shadings, no glitch. Truly mind blowing."*

From anecdotal evidence gathered, methods used to generate these videos appear to be the same as those used to generate deepfake adult pornography. One perpetrator claims:

> *"I'm using [redacted] . Go to [redacted] and look in the [redacted]. Everything you need to do this is there."*

Free, open-source AI software is behind many viral deepfake videos and faces the same inherent challenges as open-source foundation models over malicious or illegal use, including for non-consensual pornography (the overwhelming majority of existing deepfakes) and CSAM.

## Dark web forums and AI CSAM discussion

Given the gradual increase in the number of AI CSAM reports on the clear web, dark web forums remain the main hub for IWF for intelligence-gathering on many aspects of AI CSAM.

Dark web CSAM forums are mostly concerned with the sharing and discussion of real CSAM. AI CSAM remains a small – but likely growing – part of these more general forums.

There remain large variations in the level of interest in AI CSAM among these wider CSAM communities. In a recent exchange, an AI-generated image was posted in a section intended for real images, and was met with a mixture of apathy and antipathy:

> *"Nice but it looks like AI-gen and none of us want that."*

> *"100%, only want the real stuff."*

> *"Thanks for these but I'm not into AI."*

SEE RECOMMENDATION #1

For Government

**A common thread across various AI CSAM communities relates to requests for guidance or training, as briefly set out in the October 2023 report. Where people new to AI encounter AI CSAM, they are sometimes impressed:**

> *"They look very real, like you've taken photos of them."*

Perpetrators encourage people towards trying certain generative AI models:

> *"If you are undressing little girls, I think your only mostly safe option is running [redacted] locally."*

> *"Generating on-topic [CSAM] content is the same as any other content, just with a different prompt. After that it's just a lot of experimentation!"*

AI text-to-image models, though, can be daunting for those starting out. Those people, then, ask for advice, tutorials or guides:

> *"Just wondering if you have any tutorials, or how someone can get started making their own pictures?"*

> *"I am very interested in learning how to use AI to develop child porn… just show me the step by step"*

> *"Wanted to know if someone could point me in the right direction to learn, download the software, etc."*

SEE RECOMMENDATION #1

For Government

**At time of writing, the UK prohibition on paedophile manuals continues to exclude pseudo-photographs of children – necessarily encompassing all AI CSAM. This means, therefore, that tutorials and guides shared among members of these communities detailing how to generate realistic AI CSAM remain legal.**

## AI CSAM featuring known victims and famous children

As discussed in the October 2023 report, AI CSAM perpetrators regularly use AI models to generate images of existing children – and the majority of CSAM fine-tuned AI models are designed for generating their images. These are usually known victims of child sexual abuse or famous children.

Perpetrators on dark web forums continue to discuss how to train LoRAs (fine-tuned models) for those named victims or celebrities, share models they have trained and images they have generated, and request new ones.

AI CSAM images of celebrity children may have a broader appeal than images featuring known victims of CSAM. Such images – including some 'packs' of AI CSAM celebrity images – have been seen multiple times on sites on the clear web. Variously, they feature famous children and de-aged famous adults.

It is possible that the world of fine-tuned CSAM models – including those for generating images of named children – runs much deeper than is apparent from looking only in publicly accessible areas on the clear web and the dark web. This is a world that reaches into homes with non-internet connected devices, and – crucially – into end-to-end encrypted, peer-to-peer networks that are inaccessible to organisations like the IWF.

SEE RECOMMENDATION #2

For Government

**One user, seemingly mostly active in these peer-to-peer networks, shared an anonymous webpage containing links to fine-tuned models for 128 different named victims of child sexual abuse.**

Every 'child model' had [redacted] variations; many also had options for younger or older versions of the child in question.

One of these 'child models', 'Olivia', was featured by IWF back in our 2018 Annual Report. In that report, an analyst recounted:

***"I first saw Olivia when she was about three.***

***I've seen Olivia grow up through cruel images and videos, suffering hideous abuse. She was repeatedly raped and sexually tortured.***

***We see Olivia every day—five years after she was rescued. To show exactly what 'repeat victimisation' means, we counted the number of times we saw Olivia's image online during a three-month period. We saw her at least 347 times. On average, that's five times each and every working day."***

An AI model for generating novel images of Olivia is available to download for free, just a couple of clicks away. The user can choose to use a particular version of the model. It's a potentially 'popular' model among AI CSAM communities – as one user asks elsewhere,

> *"Anyone trained a LoRA for [Olivia] yet? Would be really cool to see"*

That user is pointed towards the anonymous link that IWF has identified.

Before the advent of AI-generated images, survivors of childhood sexual abuse like Olivia already had to contend with the potential for the images and videos displaying their abuse being shared across the internet. Each time one of those images was shared or viewed added another link in a long chain of child sexual abuse.

Fine-tuned models like Olivia's have been trained on the imagery that IWF was seeing five times a day in 2018 but was unable to eradicate. The consequence of this is a new way of adding links to the chain – each time re-victimising survivors of child sexual abuse – and potentially without end, since perpetrators can generate as many images of those children as they like without fear of detection or prevention.

As explained in the October 2023 report, these are lifelike images – they look like images of real-world abuse – but can produce 'unreal', unseen settings, scenarios, and sexual activities.

SEE RECOMMENDATION #2

For Government

These models fine-tuned on CSAM victims – including Olivia's – remain legal in the UK.

# AI CSAM image analysis: new snapshot study

## Overall forum trends

Part of the October 2023 report comprised a snapshot study of a dark web CSAM forum. For that report, all the live AI-generated images posted to the forum in a 30-day period (September 2023) were identified, and a selection were assessed.

This update revisited the same dark web CSAM forum to analyse whether use of the forum had changed in type or frequency; whether any trends in imagery could be identified; and whether discussions among forum users had progressed.

This new snapshot took the live AI-generated images posted to the forum over another 30-day period (9 March to 7 April 2024) – this time, all the images that were found were assessed by IWF analysts.

The table below compares findings on posts of AI-generated imagery to the forum, and on AI-specific threads (sections where users post content) within the forum, over the two periods.

| | September 2023 | March-April 2024 |
|---|---|---|
| AI-generated images posted (incl. duplicates) | 20,254 | 13,906 |
| AI-generated videos posted | 0 | 9 |
| Count of threads to which (live) AI-generated content was posted | 74 | 106 |
| Sum of views on AI-specific threads created over period | 261,920 | 319,141 |

These findings show that the number of AI-generated images posted to the forum decreased from September 2023 to March 2024. These were distributed across more threads. Images were, then, generally shared in smaller, more 'curated' sets.

The number of views on AI-specific threads created over the two periods increased by 22%. If view count is some guide to general interest, it may be concluded that the level of interest in AI CSAM has increased slightly among users of this forum. (Nonetheless, data on number of unique users is unavailable, so it is impossible to say whether this shows that more people are interested in AI CSAM.)

**9 AI-generated deepfake videos were found to have been posted within the period analysed for this snapshot.** These are not the first AI-generated videos found by IWF on dark web forums, but it is notable that none were found shared here six months previously.

## Image analysis

For this new snapshot study, 13,906 online images were identified and downloaded. After de-duplication, **these totalled 12,148 unique AI-generated images.**

12 IWF analysts dedicated a combined total of 130.5 hours to assessing these 12,148 images.

As outlined in the October 2023 report, AI CSAM in the UK falls under two different laws, which have different criteria and sentencing guidelines:

- The **Protection of Children Act 1978** (as amended by the Criminal Justice and Public Order Act 1994). This law criminalises the taking, distribution and possession of an "indecent photograph or pseudo-photograph of a child".
- The **Coroners and Justice Act 2009.** This law criminalises the possession of "a prohibited image of a child". These are non-photographic – generally cartoons, drawings, animations or similar.

The key criterion for classification as criminal under the Protection of Children Act 1978 is that the image "appears to be a photograph".

**2,985 images were classified as indecent pseudo-photographs, and 527 images were classified as prohibited images – in total, this is 3,512 AI CSAM images.**

These are shown as proportions of the 12,148 assessed images in the graph below:

**AI-generated images assessed**

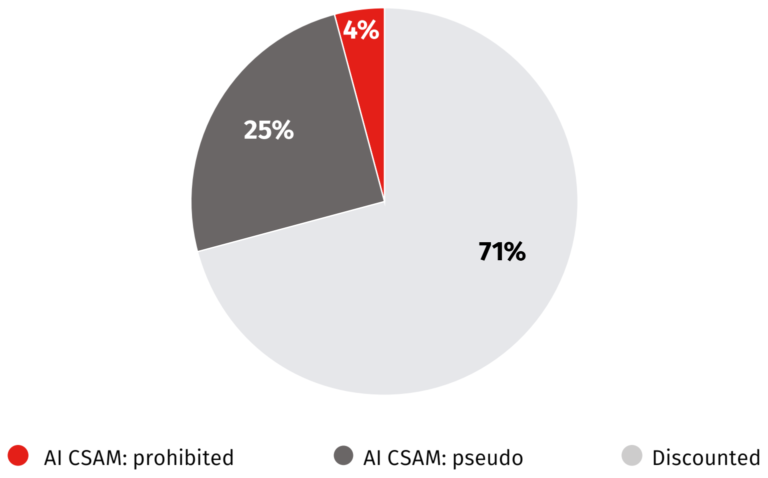


**Figure 4**
Classification breakdown of AI-generated images posted to the forum from March to April 2024.

Source: IWF analysis

**The total proportion of images actioned as criminal was 29% of the unique AI-generated images found on the forum.** 71% of the images on the forum were non-criminal.

Of the criminal images, **over five times as many images were assessed as realistic pseudo-photographs than were assessed as non-realistic prohibited images.** This is close to the results of the previous snapshot, in which the proportion was approximately six-to-one.

Comparing the image assessments between the two studies, relative to the number of AI-generated images posted to the forum over the two periods, including any duplicates in 'discounted' figures, yields the following comparative chart:

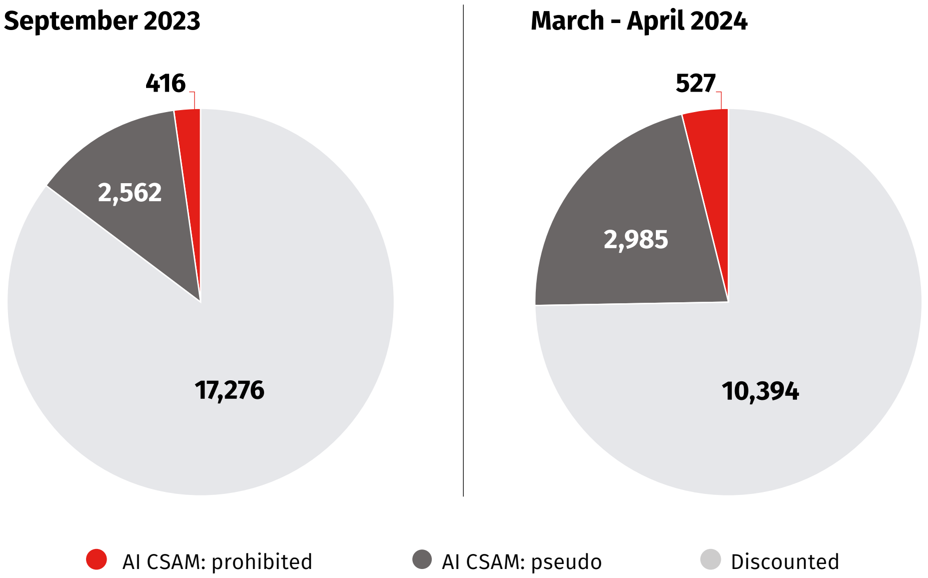


**Figure 5**
AI-generated image assessments compared between the first and second snapshot studies.

Source: IWF analysis

Though the overall number of AI-generated images posted to the forum decreased, among them, a higher number of criminal AI CSAM images were found in this second snapshot.

**The total number of criminal images found on this forum now stands at 6,490.**

Images assessed as indecent pseudo-photographs of children can be sorted by UK Sentencing Council Category and by ages of children. In images in which multiple categories or children are present, the most severe category and youngest age are selected.

The UK Sentencing Council Categories are:

**Category A**
Images depicting penetrative sexual activity; images involving sexual activity with an animal; or sadism.

**Category B**
Images depicting non-penetrative sexual activity.

**Category C**
Other indecent images not falling within categories A or B.

The relative proportions of category and age assessments for AI CSAM (pseudo-photographs) for both snapshots are shown below:

### AI CSAM (pseudo) images by severity

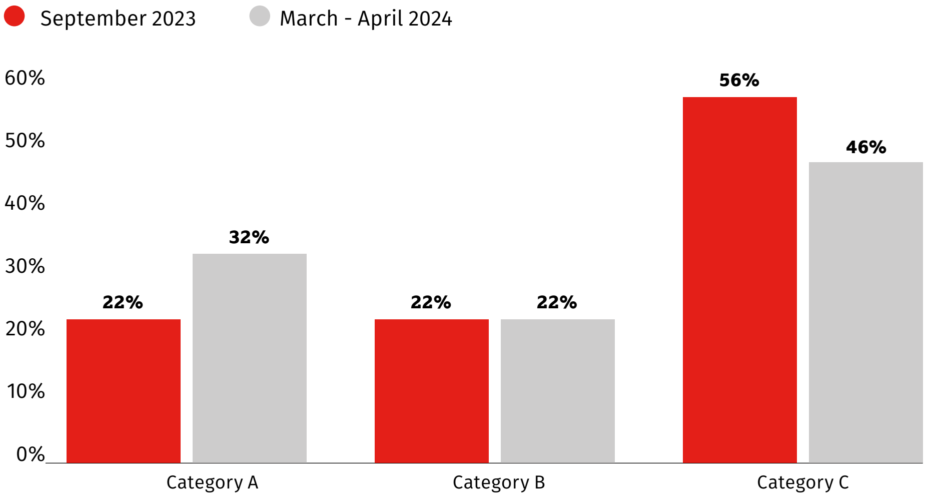


**Figure 6**
AI CSAM (pseudo) images by severity.

Source: IWF analysis

### AI CSAM (pseudo) images by age

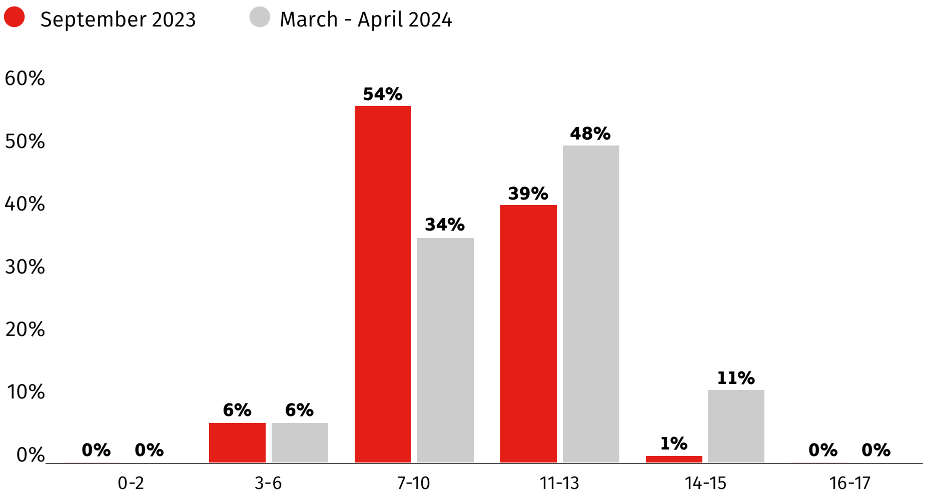


**Figure 7**
AI CSAM (pseudo) images by age.

Source: IWF analysis

Notably, **the proportion of Category A images has increased by 10 percentage points.** This could indicate that technology and expertise has advanced such that perpetrators are experiencing more success generating 'hardcore' scenarios – generally, penetrative sexual activities – complex scenes involving multiple individuals, which image-generating AI has historically had greater difficulties producing accurately compared to scenes involving just one individual.

**Just as in the previous snapshot, female children featured in over 99% of AI CSAM images assessed.**

This second snapshot found 3,512 AI CSAM images out of 12,148 total assessed images. The 8,636 images found to be non-criminal were given reasons for being 'discounted', which can be compared to the discounted reasons for the images assessed in the September snapshot as follows:

| **Discounted Reason** | **September 2023** | **%** | **March-April 2024** | **%** |
|---|---|---|---|---|
| Age indeterminate | 390 | 5% | **248** | **3%** |
| Child depicted (non-criminal) | 4,340 | 53% | **6,486** | **75%** |
| Adult depicted (non-sexual) | 492 | 6% | **499** | **6%** |
| Non-criminal non-photographic (NPI) | 1,634 | 20% | **651** | **8%** |
| Off remit | 1,274 | 16% | **752** | **9%** |
| **Total** | 8,130 | 100% | **8,636** | **100%** |

In this snapshot, an even lower proportion of non-criminal images assessed – just 8% – were determined to be not realistic enough to assess as a pseudo-photograph (whether that image depicted a child or an adult). These are those images marked 'Non-criminal non-photographic' (NPI) in the table above. This provides some support for claims of increasing realism of AI-generated images over the last six months.

**This snapshot aims to dig deeper into those AI-generated images of children shared in the dark web CSAM forum that were assessed to be non-criminal.**

What was the nature of those images? Could they be mistaken for 'innocent' images of children, or could they be considered sexually exploitative images of children?

IWF analysts were tasked with categorising the non-criminal AI-generated images of children into two categories: exploitative and non-exploitative. These are broadly defined as:

**Exploitative:** a child depicted such that a reasonable person would consider it sexually exploitative. The image has a (slightly) sexual element but does not meet criminal thresholds.

**Non-exploitative:** a child depicted in a non-sexual situation. This spans, for example, images of fully clothed children in indoor or outdoor settings, as well as images that may be considered legitimate nudism settings.

It should be noted that the tag 'Child depicted – non-criminal' in the discounted table above – encompassing 75% of the non-criminal images in this snapshot – then indicates that each image in this category did not meet criminal thresholds, but does not necessarily indicate that there was no sexualised element at all.

Even with an exploitative category definition that excludes the many images of children in nudism settings among this set, **a significant proportion (42%) of non-criminal AI-generated images of children were classified as exploitative.** These 2,980 images sometimes featured multiple children – 3,778 children were identified in these images in total. One realistic image depicted 14 children in a sexually exploitative (though non-criminal) context.

Nonetheless, most non-criminal AI-generated images of children, assessed independently – outside the context of their sharing on the dark web – could be considered non-exploitative.

### Discounted AI-generated images of children - non-exploitative or exploitative?

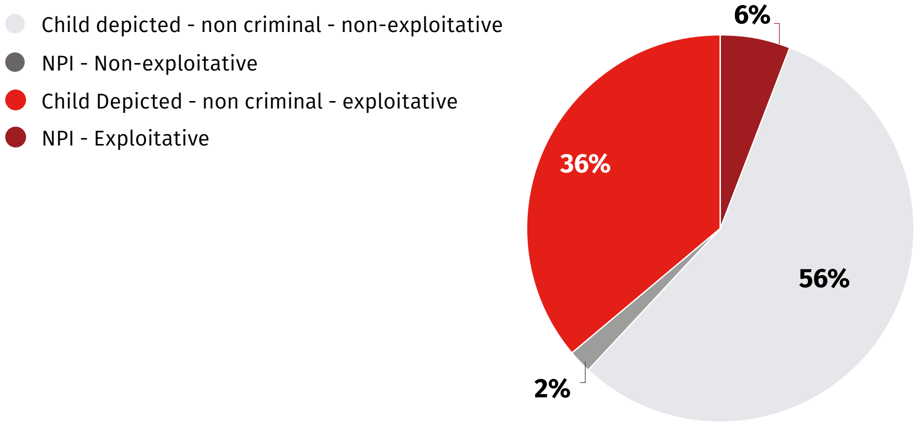


**Figure 8**
Discounted AI-generated images of children - non-exploitative or exploitative?

Source: IWF analysis

For comparative purposes, this snapshot revisited the dark web CSAM forum surveyed for the snapshot in the October 2023 report. This means that the same limitations persist:

- Only one CSAM forum was surveyed.
- The forum surveyed has a general preference towards 'softcore' imagery, and imagery of girls.
- The AI sections of this forum has a few regular 'creators' – though these have changed somewhat from those in the previous snapshot – still, large batches of images assessed originate from the same few perpetrators.

The key findings on images assessed across the two snapshots can be summarised as follows:

- AI-generated images of children, including AI CSAM, already mostly looked like real images of children by the autumn of last year, with 82% of images assessed being realistic enough to be assessed as pseudo-photographs of children (if criminal). **This snapshot found that 90% of images assessed met this threshold, indicating that an even higher proportion of AI-generated images of children look like real images of children now.** It is important to re-emphasise that this proportion will likely never reach 100%, because a small class of perpetrators do not hold realism as a goal – these perpetrators simply prefer non-realistic image styles, like cartoon or anime.
- It remains the case that most AI-generated images of children found are non-criminal, and of those images, most could be classed as non-exploitative. This is further evidence that **there is a large appetite among online CSAM communities for images of children outside scenarios containing explicit sexual activity and outside scenarios that could be considered sexually exploitative.**
- Over time, **both the AI technology itself and its users are getting better at depicting realistic complex scenes involving multiple characters.** This is reflected in the increases in the number and proportion of Category A AI CSAM images.

## Metadata analysis

New for this snapshot, after de-duplication, IWF undertook a scrape of available metadata on all those images shared in this dark web CSAM forum over the 30-day period. Of the 12,148 images assessed, **1,675 images had some 'useful' metadata available.**

The overwhelming majority of those images were shown to have been generated through [redacted], the most popular [redacted] graphical user interface (GUI).

**86% of images shared did not contain any useful metadata.** As discussed in the October 2023 report, this is a consequence of being generated by open-source models. Processes to strip images of their metadata are usually incorporated by perpetrators into their ‘workflows’.

What comprised useful metadata varied between images in this set. This included evidence of the full prompts used; the models used, including foundation models, Checkpoint models and LoRAs; the seed and number of steps in generation; post-processing steps, including upscaling, file conversion, face-swapping (using publicly-available tools), and use of Photoshop. In some cases, combining all this data could be enough to replicate generated images perfectly or almost perfectly.

One important caveat applies to this analysis: it relates to just 14% of assessed images, so, because metadata editing is likely to take place to bulk generations or bulk sets of images, the images included in this analysis are likely to be from the same few sets and the same few perpetrators.

**The most common positive prompt terms related to realistic or photographic images and referred to anatomical features of children.** Other frequent positive prompts described various sexual activities and settings.

**The most common negative prompt terms related to adult sexual or anatomical characteristics, as well as certain racial descriptors.** Other frequent negative prompts described types of body deformities (so-called AI artefacts) as well as those designed to nudge the model towards generating sexually violent or coercive images.

Checkpoint models are large-scale ‘base’ models that are the bedrock for all image generation. These can be ‘foundation’ models released officially or they can be models that have been fine-tuned by users. These are often produced and distributed for specific purposes, such as generating pornography.

1,290 images among the set had evidence of use of a Checkpoint model. The remaining 385 images did not contain details of the Checkpoint model used.

**Of these 1,290 images, at least 897 (or 70% of these images) used a Checkpoint model that was publicly available** – freely downloadable from sites like [redacted] or [redacted]. Up to 30% of these images, then, may have used CSAM fine-tuned models. This proportion was roughly the same among the criminal subset of these images that had this data available.

**More LoRA models – smaller fine-tuned models, generally applied on top of Checkpoint models – in evidence among the set, appeared to be CSAM fine-tuned models, including many for generating images of named children.**

SEE RECOMMENDATION #2

For Government

# In summary

**Since the IWF's last AI CSAM report in October 2023, despite increasing levels of legal and regulatory scrutiny, the pace of technological progress has not been slowing.**

With image generation considered close to being 'solved', resources are being poured into trying to solve the next frontier: video generation. We get glimpses of future model capabilities in the 2024 previews of OpenAI's Sora and Google's Veo.

At the same time, focus on specialised text- or image-generating models is to some extent giving way to a focus on models built to be intrinsically multimodal – to work fluently across text, audio, image, video, and code. The future is likely to hold general, all-purpose AI systems that can interpret inputs and produce outputs of all kinds.

AI-generated child sexual abuse and exploitation, just like other types of child sexual abuse, is not limited only to highly technical individuals; to those on the dark web; or to those with extreme or violent sexual urges.

It is a multifaceted threat that encompasses perpetrators of all levels of technical knowledge, including children themselves; sites for distributing, buying and selling AI CSAM on the clear web, as well as on the dark web; people seeking out non-criminal images of children, as well as images of 'hardcore' sexual scenarios.

In this context, there has never been a more urgent need for child safety by design across all the stages of model development and distribution, and among all the players in the AI ecosystem.

While it would be a mistake to assume that closed-source models are watertight without extensive child safety research and testing, open-source models still comprise the main threat in the AI CSAM landscape. If the last few years are a guide, where closed-source models lead, open-source equivalents will inevitably follow. We may yet experience a watershed moment for AI CSAM when fast, malleable generative AI video becomes accessible to the general public.

The AI CSAM videos found in the course of research for this report update, likely created with primitive open-source tools, are the canaries in the coal mine.

With investment in AI safety research, collaborative cross-industry initiatives, and regulation that is adaptable and dynamic, mitigations are possible. And mitigations are needed now – while real harm is being perpetrated against real individuals.

For further information on this report update, please email **media@iwf.org.uk**

**DISCLAIMER**

*The images used in this report are screenshots of content available on the clear web and dark web. We've attempted to cite the sources of these screenshots, some of which depict likenesses of famous people or films. These likenesses have been generated by someone submitting prompts to AI models. They are not images of the actors or from the films themselves. This goes some way towards demonstrating the photorealism of images produced by AI models.*

# Glossary

**AI:** *Artificial Intelligence.*

**AI CSAM:** child sexual abuse material that has been generated or edited by Artificial Intelligence.

**Base Model (or Foundation Model):** an AI model, generally those released directly by generative AI companies, designed to produce a wide and general variety of outputs.

**Category A:** a classification of child sexual abuse images depicting penetrative sexual activity; images involving sexual activity with an animal or sadism, as according to the Sentencing Council's Sexual Offences Definitive Guideline.

**Category B:** a classification of child sexual abuse images depicting non-penetrative sexual activity, as according to the Sentencing Council's Sexual Offences Definitive Guideline.

**Category C:** a classification of indecent images of children not falling within categories A or B, as according to the Sentencing Council's Sexual Offences Definitive Guideline.

**Closed-source models:** software whose source code is not released to the public. The public are not able to use, study, change, or distribute the software or its source code to anyone or for any purpose.

**Coroners and Justice Act 2009:** this law criminalises the possession of "a prohibited image of a child". These are non-photographic – generally cartoons, drawings, animations or similar.

**CSAM:** *child sexual abuse material.*

**Dark Web:** the side of the World Wide Web that is not indexed by search engines and requires specific configuration, software, or authorization to access allowing users and website operators to remain anonymous or untraceable.

**Deepfakes:** media (images, videos, or audio) that has been digitally manipulated through AI tools or software to replace one person's likeness convincingly with that of another.

**Diffusion Model:** text-to-image models that add and remove layers of 'noise' to images. Running the 'de-noising' process on random seeds generates 'new' images.

**Generative AI:** a type of machine learning that uses deep learning models to identify the patterns and structures within existing data to generate new content.

**IWF:** *Internet Watch Foundation.*

**LEAs:** *law enforcement agencies.*

**Open-source models:** software whose source code is released under a license in which the copyright holder grants users the rights to use, study, change, and distribute the software and its source code to anyone and for any purpose.

**Open/Clear Web:** the side of the web that is public and viewable by everyone.

**Prompts:** words or short phrases used to describe what you do (positive prompts) or do not (negative prompts) want to see in the image when using generative text-to-image models.

**Pseudo-photograph:** an image (including one generated by a computer) that appears to be a photograph.

**Real CSAM:** child sexual abuse material that has not been generated or edited by AI technology.

**Self-generated content:** when children are groomed, deceived or extorted into producing sexual images and/or videos of themselves and sharing them online.

**Text-to-image model:** a type of machine learning model whose function is to generate images from text prompts.

The Minnesota Star Tribune

NEWS & POLITICS

# AI-generated images depicting child sexual abuse turning up in more Minnesota criminal cases

Online tips submitted to experts who track the abusive material have exploded in the past year.

**By Sarah Nelson**
The Minnesota Star Tribune

SEPTEMBER 15, 2025 AT 5:00AM

More than 10,500 cyber tips reporting child pornography have been flagged in Minnesota this year, images created with generative AI and others without. (Leila Navidi/The Minnesota Star Tribune)

A decade or more ago, computer-manipulated child pornography looked a lot like what was possessed by Jacob Wetterling's killer Danny Heinrich: binders filled with images morphed in his home with Photoshop software.

Today, investigators on the front lines say they're facing something much more complicated and dangerous: child pornography created using artificial intelligence.

The technology has evolved at breakneck speed, far removed from the images seized in Heinrich's home in 2015. That seizure led to his eventual confession to kidnapping and killing Wetterling, 11, who had been missing for 26 years.

It's now easier for bad actors to flood the internet with abusive depictions of children that are nearly indistinguishable from actual images of child sex abuse. The influx of what they call child sexual abuse material has placed a strain on law enforcement resources.

The explicit images have become more lifelike and sophisticated in the short time since the technology burst into the public sphere in late 2023. In most cases, the images are of real children – taken from yearbooks, social media or even surreptitiously in public. They're then manipulated with AI to become child pornography. Although children are not physically harmed in the production of AI or otherwise manipulated pornography, computer-generated depictions are also illegal under the federal PROTECT Act, passed in 2003 to prevent the exploitation of minors.

"It's just one little pixel, or one little misalignment of one blade of grass that maybe just doesn't look quite right. They're that real looking," said Carla Baumel, an assistant U.S. Attorney for Minnesota. "If you can imagine your 6-year-old in the worst position and worst light ever, and wonder if [the image] is real, there's no question that's a harm."

More than 10,500 cyber tips reporting child pornography have been flagged in Minnesota this year, images created with generative AI and others without. The number is on track to outpace last year's 12,595 tips to the Minnesota Bureau of Criminal Apprehension (BCA). The National Center for Missing and Exploited Children is experiencing an increase so sharp it prompted researchers to release their annual online child sex crimes report early.

In the first half of 2024, the center received 6,800 reports flagging child pornography created with generative AI. Through June this year, tips have exploded to more than 440,000.

"That's absolutely a strain on [police] resources. Because all of those deserve attention by law enforcement but you have to triage," BCA Superintendent Drew Evans said. "I think we're always playing catch-up because of the sheer volume. … I do think that AI adds to that volume that makes a difficult caseload even more unmanageable in that process."

Baumel is the lead federal prosecutor in a criminal case pending in Minnesota's U.S. District Court that's believed to be one of the first prosecutions in the state involving a swell of victims whose images were used in AI-generated material.

In February, the U.S. Attorney's Office indicted William Haslach, 30, of Maplewood on a federal charge of using AI to produce child sexual abuse images with photos he discreetly took of children at several Ramsey County schools where he worked as a recess and traffic monitor. To date, federal prosecutors have identified 100 children victimized in the case.

"They were doing nothing wrong. These parents just sent their children to a place where we just assume kids are the safest," Baumel said. "So when we had a meeting with parents, there is just this total confidence that's rocked."

Hennepin County prosecutors in August charged Jason Polzin, 50, a former staff member and softball coach at Living Word Christian Center in Brooklyn Park, with interfering with the privacy of a minor after he allegedly recorded a 13-year-old girl and superimposed her face on a computer-generated nude or scantily clad female body.

In both cases, the men are accused of putting a photo through an app or website to generate an image that makes the victim appear undressed or in an explicit position. The Minnesota Star Tribune left messages for the attorneys representing the accused.

Federal prosecutors said the method makes the crime unique in that a person is often unaware of what happened.

"It's one of the only crimes where it's possible to be victimized and not confer with a victim," said Melinda Williams, assistant U.S. Attorney for Minnesota.

Acting U.S. Attorney Joe Thompson and Assistant U.S. Attorney Carla Baumel outside the U.S. Attorney's Office in Minneapolis on Aug. 20. Federal prosecutors in Minnesota are now working on a major case involving people using AI to create child pornography online. (Leila Navidi/The Minnesota Star Tribune)

Alarmed by how rampant and realistic AI-generated child pornography has become, Evans said the BCA was spurred to support expanding the state's laws to include creating with artificial intelligence explicit images and videos of children.

Under Minnesota law, possessing child pornography carries a maximum penalty of five years in prison. For federal cases, the prison term ranges from five to 20 years.

But experts said even if a person sees justice in the form of a conviction, the impact often lasts long after the case closes.

“I think everybody intuitively knows the harm of being sexually abused, but there’s an additional harm that especially the victims speak very profoundly to, of knowing images of you being raped or being depicted as raped are out in the world,” Williams said.

The National Center for Missing and Exploited Children created a Take It Down tool for people under 18 who are seeking to remove explicit photos of them shared without their consent. But some images are spread to the darkest corners of the internet, where they are difficult for law enforcement to reach.

“It’s just that continual revictimization recirculation,” said John Shehan, vice president of the exploited children division at the National Center for Missing and Exploited Children. “When I talk to victims and survivors, they talk about walking down the street and they don’t know if people are going to recognize them, if they’ve seen their imagery. It’s really a terrible thing to have to live with mentally.”

The Minnesota Star Tribune

# *Teen Girls Confront an Epidemic of Deepfake Nudes in Schools*

Using artificial intelligence, middle and high school students have fabricated explicit images of female classmates and shared the doctored pictures.

▶ **Listen to this article · 9:49 min** Learn more

**By Natasha Singer**

Natasha Singer has covered student privacy for The Times since 2013. She reported this story from Westfield, N.J.

April 8, 2024

Westfield Public Schools held a regular board meeting in late March at the local high school, a red brick complex in Westfield, N.J., with a scoreboard outside proudly welcoming visitors to the “Home of the Blue Devils” sports teams.

But it was not business as usual for Dorota Mani.

In October, some 10th-grade girls at Westfield High School — including Ms. Mani’s 14-year-old daughter, Francesca — alerted administrators that boys in their class had used artificial intelligence software to fabricate sexually explicit images of them and were circulating the faked pictures. Five months later, the Manis and other families say, the district has done little to publicly address the doctored images or update school policies to hinder exploitative A.I. use.

“It seems as though the Westfield High School administration and the district are engaging in a master class of making this incident vanish into thin air,” Ms. Mani, the founder of a local preschool, admonished board members during the meeting.

In a statement, the school district said it had opened an “immediate investigation” upon learning about the incident, had immediately notified and consulted with the police, and had provided group counseling to the sophomore class.

Tenth-grade girls at Westfield High School in New Jersey learned last fall that male classmates had fabricated sexually explicit images of them and shared them. Peter K. Afriyie/Associated Press

“All school districts are grappling with the challenges and impact of artificial intelligence and other technology available to students at any time and anywhere,” Raymond González, the superintendent of Westfield Public Schools, said in the statement.

Blindsided last year by the sudden popularity of A.I.-powered chatbots like ChatGPT, schools across the United States scurried to contain the text-generating bots in an effort to forestall student cheating. Now a more alarming A.I. image-generating phenomenon is shaking schools.

Boys in several states have used widely available “nudification” apps to pervert real, identifiable photos of their clothed female classmates, shown attending events like school proms, into graphic, convincing-looking images of the girls with exposed A.I.-generated breasts and genitalia. In some cases, boys shared the faked images in the school lunchroom, on the school bus or through group chats on platforms like Snapchat and Instagram, according to school and police reports.

Such digitally altered images — known as “deepfakes” or “deepnudes” — can have devastating consequences. Child sexual exploitation experts say the use of nonconsensual, A.I.-generated images to harass, humiliate and bully young women can harm their mental health, reputations and physical safety as well as pose risks to their college and career prospects. Last month, the Federal Bureau of Investigation warned that it is illegal to distribute computer-generated child sexual abuse material, including realistic-looking A.I.-generated images of identifiable minors engaging in sexually explicit conduct.

Yet the student use of exploitative A.I. apps in schools is so new that some districts seem less prepared to address it than others. That can make safeguards precarious for students.

“This phenomenon has come on very suddenly and may be catching a lot of school districts unprepared and unsure what to do,” said Riana Pfefferkorn, a research scholar at the Stanford Internet Observatory, who writes about legal issues related to computer-generated child sexual abuse imagery.

At Issaquah High School near Seattle last fall, a police detective investigating complaints from parents about explicit A.I.-generated images of their 14- and 15-year-old daughters asked an assistant principal why the school had not reported the incident to the police, according to a report from the Issaquah Police Department. The school official then asked “what was she supposed to report,” the police document said, prompting the detective to inform her that schools are required by law to report sexual abuse, including possible child sexual abuse material. The school subsequently reported the incident to Child Protective Services, the police report said. (The New York Times obtained the police report through a public-records request.)

In a statement, the Issaquah School District said it had talked with students, families and the police as part of its investigation into the deepfakes. The district also “shared our empathy,” the statement said, and provided support to students who were affected.

The statement added that the district had reported the “fake, artificial-intelligence-generated images to Child Protective Services out of an abundance of caution,” noting that “per our legal team, we are not required to report fake images to the police.”

At Beverly Vista Middle School in Beverly Hills, Calif., administrators contacted the police in February after learning that five boys had created and shared A.I.-generated explicit images of female classmates. Two weeks later, the school board approved the expulsion of five students, according to district documents. (The district said California’s education code prohibited it from confirming whether the expelled students were the students who had manufactured the images.)

Michael Bregy, superintendent of the Beverly Hills Unified School District, said he and other school leaders wanted to set a national precedent that schools must not permit pupils to create and circulate sexually explicit images of their peers.

“That’s extreme bullying when it comes to schools,” Dr. Bregy said, noting that the explicit images were “disturbing and violative” to girls and their families. “It’s something we will absolutely not tolerate here.”

Michael Bregy, superintendent of Beverly Hills schools, said he wanted to send a message that schools must not allow pupils to make and share explicit images of their peers. Tracy Nguyen for The New York Times

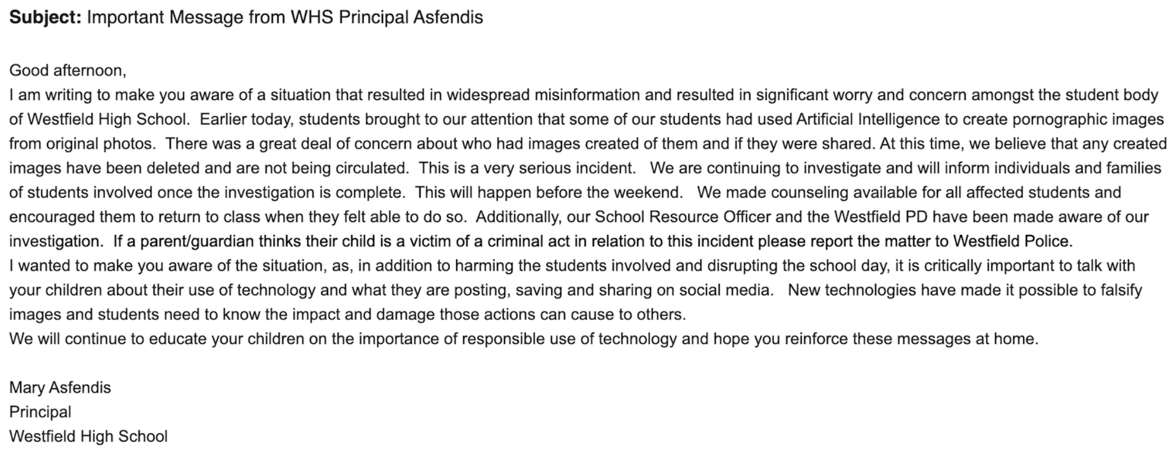
Schools in the small, affluent communities of Beverly Hills and Westfield were among the first to publicly acknowledge deepfake incidents. The details of the cases — described in district communications with parents, school board meetings, legislative hearings and court filings — illustrate the variability of school responses.

The Westfield incident began last summer when a male high school student asked to friend a 15-year-old female classmate on Instagram who had a private account, according to a lawsuit against the boy and his parents brought by the young woman and her family. (The Manis said they are not involved with the lawsuit.)

After she accepted the request, the male student copied photos of her and several other female schoolmates from their social media accounts, court documents say. Then he used an A.I. app to fabricate sexually explicit, “fully identifiable” images of the girls and shared them with schoolmates via a Snapchat group, court documents say.

Westfield High began to investigate in late October. While administrators quietly took some boys aside to question them, Francesca Mani said, they called her and other 10th-grade girls who had been subjected to the deepfakes to the school office by announcing their names over the school intercom.

That week, Mary Asfendis, the principal of Westfield High, sent an email to parents alerting them to "a situation that resulted in widespread misinformation." The email went on to describe the deepfakes as a "very serious incident." It also said that, despite student concern about possible image-sharing, the school believed that "any created images have been deleted and are not being circulated."



**Subject:** Important Message from WHS Principal Asfendis

Good afternoon,
I am writing to make you aware of a situation that resulted in widespread misinformation and resulted in significant worry and concern amongst the student body of Westfield High School. Earlier today, students brought to our attention that some of our students had used Artificial Intelligence to create pornographic images from original photos. There was a great deal of concern about who had images created of them and if they were shared. At this time, we believe that any created images have been deleted and are not being circulated. This is a very serious incident. We are continuing to investigate and will inform individuals and families of students involved once the investigation is complete. This will happen before the weekend. We made counseling available for all affected students and encouraged them to return to class when they felt able to do so. Additionally, our School Resource Officer and the Westfield PD have been made aware of our investigation. If a parent/guardian thinks their child is a victim of a criminal act in relation to this incident please report the matter to Westfield Police.
I wanted to make you aware of the situation, as, in addition to harming the students involved and disrupting the school day, it is critically important to talk with your children about their use of technology and what they are posting, saving and sharing on social media. New technologies have made it possible to falsify images and students need to know the impact and damage those actions can cause to others.
We will continue to educate your children on the importance of responsible use of technology and hope you reinforce these messages at home.

Mary Asfendis
Principal
Westfield High School

An October email that the principal of Westfield High sent to parents about the deepfakes.

Dorota Mani said Westfield administrators had told her that the district suspended the male student accused of fabricating the images for one or two days.

Soon after, she and her daughter began publicly speaking out about the incident, urging school districts, state lawmakers and Congress to enact laws and policies specifically prohibiting explicit deepfakes.

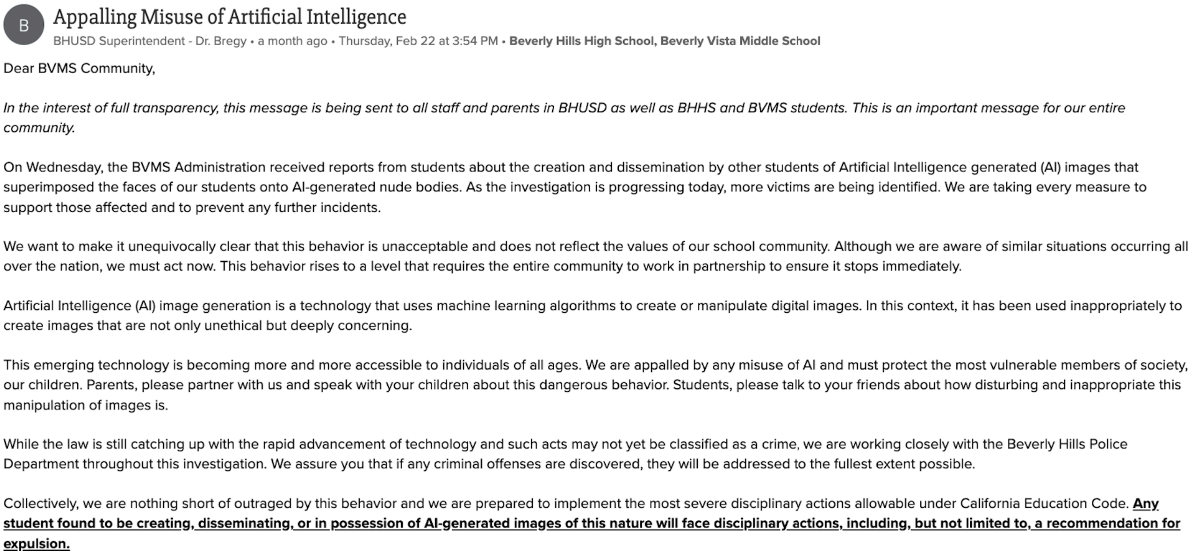
“We have to start updating our school policy,” Francesca Mani, now 15, said in a recent interview. “Because if the school had A.I. policies, then students like me would have been protected.”

Parents including Dorota Mani also lodged harassment complaints with Westfield High last fall over the explicit images. During the March meeting, however, Ms. Mani told school board members that the high school had yet to provide parents with an official report on the incident.

Westfield Public Schools said it could not comment on any disciplinary actions for reasons of student confidentiality. In a statement, Dr. González, the superintendent, said the district was strengthening its efforts “by educating our students and establishing clear guidelines to ensure that these new technologies are used responsibly.”

Beverly Hills schools have taken a stauncher public stance.

When administrators learned in February that eighth-grade boys at Beverly Vista Middle School had created explicit images of 12- and 13-year-old female classmates, they quickly sent a message — subject line: “Appalling Misuse of Artificial Intelligence” — to all district parents, staff, and middle and high school students. The message urged community members to share information with the school to help ensure that students’ “disturbing and inappropriate” use of A.I. “stops immediately.”



B Appalling Misuse of Artificial Intelligence

BHUSD Superintendent - Dr. Bregy • a month ago • Thursday, Feb 22 at 3:54 PM • **Beverly Hills High School, Beverly Vista Middle School**

Dear BVMS Community,

*In the interest of full transparency, this message is being sent to all staff and parents in BHUSD as well as BHHS and BVMS students. This is an important message for our entire community.*

On Wednesday, the BVMS Administration received reports from students about the creation and dissemination by other students of Artificial Intelligence generated (AI) images that superimposed the faces of our students onto AI-generated nude bodies. As the investigation is progressing today, more victims are being identified. We are taking every measure to support those affected and to prevent any further incidents.

We want to make it unequivocally clear that this behavior is unacceptable and does not reflect the values of our school community. Although we are aware of similar situations occurring all over the nation, we must act now. This behavior rises to a level that requires the entire community to work in partnership to ensure it stops immediately.

Artificial Intelligence (AI) image generation is a technology that uses machine learning algorithms to create or manipulate digital images. In this context, it has been used inappropriately to create images that are not only unethical but deeply concerning.

This emerging technology is becoming more and more accessible to individuals of all ages. We are appalled by any misuse of AI and must protect the most vulnerable members of society, our children. Parents, please partner with us and speak with your children about this dangerous behavior. Students, please talk to your friends about how disturbing and inappropriate this manipulation of images is.

While the law is still catching up with the rapid advancement of technology and such acts may not yet be classified as a crime, we are working closely with the Beverly Hills Police Department throughout this investigation. We assure you that if any criminal offenses are discovered, they will be addressed to the fullest extent possible.

Collectively, we are nothing short of outraged by this behavior and we are prepared to implement the most severe disciplinary actions allowable under California Education Code. **Any student found to be creating, disseminating, or in possession of AI-generated images of this nature will face disciplinary actions, including, but not limited to, a recommendation for expulsion.**

A February message that school administrators in Beverly Hills, Calif., sent to parents and students about deepfakes.

It also warned that the district was prepared to institute severe punishment. “Any student found to be creating, disseminating, or in possession of AI-generated images of this nature will face disciplinary actions,” including a recommendation for expulsion, the message said.

Dr. Bregy, the superintendent, said schools and lawmakers needed to act quickly because the abuse of A.I. was making students feel unsafe in schools.

“You hear a lot about physical safety in schools,” he said. “But what you’re not hearing about is this invasion of students’ personal, emotional safety.”

**Natasha Singer** writes about technology, business and society. She is currently reporting on the far-reaching ways that tech companies and their tools are reshaping public schools, higher education and job opportunities.

---

A version of this article appears in print on , Section B, Page 1 of the New York edition with the headline: Fake A.I. Nudes Create Crisis in Schools

# AI-generated images raise questions about safeguards

A photo taken on January 2, shows the letters AI for Artificial Intelligence on a laptop screen (R) next to the logo of the Chat AI application on a smartphone screen in Frankfurt am Main, western Germany.

Kirill Kudryavtsev | AFP via Getty Images

**Listen**AI-generated images raise questions about safeguards

**Saved**

Share

Stillwater schools could face lawsuits after a staff member was charged with using digital tools — including AI — to manipulate student photos in sexually explicit ways.

William Haslach worked for Adventure Club, a before-and-after-school care program with Stillwater Area Public Schools, and is facing federal criminal charges for possessing child sexual abuse material and using AI to create it.

The U.S. Attorney's Office has identified 100 children victimized in the case. Haslach has pleaded not guilty. But the case has alarmed parents and raised broader questions about how easily AI can be misused.

Ravi Bapna, an AI and analytics expert at the University of Minnesota’s Carlson School of Management, discussed potential misuses of AI software and the need for safeguards on Morning Edition.

*The following transcript has been lightly edited for clarity. Listen to the conversation by clicking the player button.*

## Law enforcement says AI generated images depicting child sexual exploitation and abuse is a growing issue, not only among adults, but some kids are creating sexually explicit images of other kids, too. What does AI photo manipulation look like?

We’re kind of in this pre-seat-belt moment, with respect to cars. When cars were first invented, people drove them without seat belts, without traffic lights, without guardrails — we didn’t say that cars were crazy, but we built those guardrails over time.

We can think of two types of AI. Traditional AI is something that I might use to build a model to detect whether there is cancer in some kind of radiology image.

Generative AI is another type that involves conditional generative modeling. Imagine taking an image, a photo or maybe a 10-second sound clip of me speaking. There is technology now that can represent that, that can create a digital signature of my profile, of my sound, in those 10 seconds, or the profile of that image. They call this an embedding. It’s basically a representation of what that image looks like.

Then there are these existing models that are trained on thousands, and tens of thousands, of people’s images or voice. And they can plug in your voice signature, or my voice signature, or my imagery, into that large model and then make it do anything. So that’s what’s going on.

## How easy is it for people to use these tools?

The technology is advancing really fast. This capability is out there. It’s baked into many of the popular AI platforms that are out there.

There are a lot of positive uses of this as well. If I wanted to clone my voice to create an audio version of my book, I could do that and I don’t have to spend three days reading my book out there. Technology always has its two sides.

## How would a school even regulate something like this?

There are lots of things we can do even today, and the industry recognizes this problem. They’re working on a variety of safeguards. There are technologies evolving around watermarking content.

As an individual user, you may want to turn off downloads on your TikTok or your Instagram presence, and maybe even resist loading high-resolution images. Because high-res images are the perfect raw material for somebody wanting to misuse this. If you’re using ChatGPT, you might want to go into the settings and tell it, “Do not train on my data.”

Then, of course, schools as well have to have policies around mobile phone usage. It should not even be allowed, actually, for employees to use their phones in this manner.

## How do you balance protection from malicious misuse with the preservation of a creator’s rights?

I think this is, again, uncharted territory in many ways. We have to not stifle innovation and creativity, while at the same time making sure that this dark side is taken care of. So this is where consent filters, water marking and legal frameworks are evolving in different jurisdictions that will give, for example, my voice, property rights.

And I think that's there's going to be innovation on that front. I know there are startups working on that area, as well as established players like Adobe, Microsoft — all of them are creating some of this technology.

### Is there a potential with synthetic media, where society might lose its grip on what is visually real? And if so, then what?

It’s a million-dollar question.

My view is that we are in this pre-seat-belt phase with cars, where we were liking this technology, but we are building the guardrails as we go along. I think a lot more innovation is going to happen.

I'll give you an example: Meta now has a paid feature — which is a problem for me; I think this should be rolled out to everybody — where if you’re creating content, if you’re Meta verified, you can say that this content should not be used for impersonation.

Why not roll that out to everybody? And therefore, when creators go out and create content, they have the rights to how it can be used. They have the rights to how it can be remixed. I wouldn’t mind maybe my content getting remixed, but I should get compensated for it, right?

That’s the future I see coming in the next two to three years.

*Do you have a question about artificial intelligence? The Morning Edition team is digging into the rise of AI. What questions do you have? What topics interest you? E-mail us at tell@mpr.org.*