Artificial intelligence (AI)

This article is more than **7 months old**

AI is overpowering efforts to catch child predators, experts warn

Safety groups say images are so lifelike that it can be hard to see if real children were subject to harms in production

A single AI model can generate tens of thousands of new images in a short amount of time, flooding the internet. Composite: The Guardian/Getty Images

Katie McQue

Thu 18 Jul 2024 12.00 EDT

The volume of sexually explicit images of children being generated by predators using artificial intelligence is overwhelming law enforcement's capabilities to identify and rescue real-life victims, child safety experts warn.

Prosecutors and child safety groups working to combat crimes against children say AI-generated images have become so lifelike that in some cases it is difficult to determine whether real children have been subjected to real harms for their production. A single AI model can generate tens of thousands of new images in a short amount of time, and this content has begun to flood both the dark web and seep into the mainstream internet.

"We are starting to see reports of images that are of a real child but have been AI-generated, but that child was not sexually abused. But now their face is on a child that was abused," said Kristina Korobov, senior attorney at the Zero Abuse Project, a Minnesota-based child safety non-profit. "Sometimes, we recognize the bedding or background in a

video or image, the perpetrator, or the series it comes from, but now there is another child's face put on to it."

There are already tens of millions of reports made each year of real-life child sexual abuse material (CSAM) created and shared online each year, which safety groups and law enforcement struggle to investigate.

"We're just drowning in this stuff already," said a Department of Justice prosecutor, who spoke on the condition of anonymity because they were not authorized to speak publicly. "From a law enforcement perspective, crimes against children are one of the more resource-strapped areas, and there is going to be an explosion of content from AI."

Last year, the National Center for Missing and Exploited Children (NCMEC) received reports of predators using AI in several different ways, such as entering text prompts to generate child abuse imagery, altering previously uploaded files to make them sexually explicit and abusive, and uploading known CSAM and generating new images based on those images. In some reports, offenders referred to chatbots to instruct them on how to find children for sex or harm them.

Experts and prosecutors are concerned about offenders trying to evade detection by using generative AI to alter images of a child victim of sexual abuse.

"When charging cases in the federal system, AI doesn't change what we can prosecute, but there are many states where you have to be able to prove it's a real child. Quibbling over the legitimacy of images will cause problems at trials. If I was a defense attorney, that's exactly what I'd argue," said the DoJ prosecutor.

Possessing depictions of child sexual abuse is criminalized under US federal law, and several arrests have been made in the US this year of alleged perpetrators possessing CSAM that has been identified as AI-generated. In most states, however, there aren't laws that prohibit the possession of AI-generated sexually explicit material depicting minors. The act of creating the images in the first place is not covered by existing laws.

In March, though, Washington state's legislature passed a bill banning the possession of AI-generated CSAM and knowingly disclosing AI-generated intimate imagery of other people. In April, a bipartisan bill aimed at criminalizing the production of AI-generated CSAM was introduced in Congress, which has been endorsed by the National Association of Attorneys General (NAAG).

Child safety experts warn the influx of AI content will drain the resources of the NCMEC CyberTipline, which acts as a clearinghouse for reports of child abuse from around the world. The organization forwards these reports on to law enforcement agencies for investigation, after determining their geographical location, priority status and whether the victims are already known.

“Police now have a larger volume of content to deal with. And how do they know if this is a real child in need of rescuing? You don’t know. It’s a huge problem,” said Jacques Marcoux, director of research and analytics at the Canadian Centre for Child Protection.

Known images of child sexual abuse can be identified by the digital fingerprints of the images, known as hash values. The NCMEC maintains a database of more than 5m hash values that images can be matched against, a crucial tool for law enforcement.

When a known image of child sexual abuse is uploaded, tech companies that are running software to monitor this activity have the capabilities to intercept and block them based on their hash value and report the user to law enforcement.

Material that doesn’t have a known hash value, such as freshly created content, is unrecognizable to this type of scanning software. Any edit or alteration to an image using AI also changes its hash value.

“Hash matching is the front line of defense,” said Marcoux. “With AI, every image that’s been generated is regarded as a brand-new image and has a different hash value. It erodes the efficiency of the existing front line of defense. It could collapse the system of hash matching.”

Child safety experts trace the escalation in AI-generated CSAM back to late 2022, coinciding with OpenAI’s release of ChatGPT and the introduction of generative AI to the public. Earlier that year, the LAION-5B database was launched, an open-source catalog of more than 5bn images that anyone can use to train AI models.

Images of child sexual abuse that had been detected previously are included in the database, which meant that AI models trained on that database could produce CSAM, Stanford researchers discovered in late 2023. Child safety experts have highlighted that children have been harmed during the process of producing most, if not all, CSAM created using AI.

“Every time a CSAM image is fed into an AI machine, it learns a new skill,” said Korobov of the Zero Abuse Project.

When users upload known CSAM to its image tools, OpenAI reviews and reports it to the NCMEC, a spokesperson for the company said.

“We have made significant effort to minimize the potential for our models to generate content that harms children,” the spokesperson said.

In 2023, the NCMEC received 36.2m reports of child abuse online, a 12% rise from the previous year. Most of the tips received were related to the circulation of real-life photos and videos of sexually abused children. However, it also received 4,700 reports of images or videos of the sexual exploitation of children made by generative AI.

The NCMEC has accused AI companies of not actively trying to prevent or detect the production of CSAM. Only five generative AI platforms sent reports to the organization last year. More than 70% of the reports of AI-generated CSAM came from social media platforms, which had been used to share the material, rather than the AI companies.

“There are numerous sites and apps that can be accessed to create this type of content, including open-source models, who are not engaging with the CyberTipline and are not employing other safety measures, to our knowledge,” said Fallon McNulty, director of the NCMEC’s CyberTipline.

Considering AI allows predators to create thousands of new CSAM images with little time and minimal effort, child safety experts anticipate an increasing burden on their resources for trying to combat child exploitation. The NCMEC said it anticipates AI fueling an increase in reports to its CyberTipline.

This expected surge in reports will affect the identification and rescue of victims, threatening an already under-resourced and overwhelmed area of law enforcement, child safety experts said.

Predators habitually share CSAM with their communities on peer-to-peer platforms, using encrypted messaging apps to evade detection.

Meta's move to encrypt Facebook Messenger in December and plans to encrypt messages on Instagram have faced backlash from child safety groups, who fear that many of the millions of cases taking place on its platforms each year will now go undetected.

Meta has also introduced a host of generative AI features into its social networks over the past year. AI-generated pictures have become some of the most popular content on the social network.

In a statement to the Guardian, a Meta spokesperson said: "We have detailed and robust policies against child nudity, abuse and exploitation, including child sexual abuse material (CSAM) and child sexualization, and those created using GenAI. We report all apparent instances of CSAM to NCMEC, in line with our legal obligations."

Child safety experts said that the companies developing AI platforms and lawmakers should be largely responsible for stopping the proliferation of AI-generated CSAM.

"It's imperative to design tools safely before they are launched to ensure they can't be used to create CSAM," said McNulty. "Unfortunately, as we've seen with some of the open-source generative AI models, when companies don't follow safety by design, there can be huge downstream effects that can't be rolled back."

Additionally, said Korobov, platforms that may be used to exchange AI-generated CSAM need to allocate more resources to detection and reporting.

"It's going to require more human moderators to be looking at images or to be going into chat rooms and other servers where people are trading this material and seeing what's out there rather than relying on automated systems to do it," she said. "You're going to have to lay eyes on it and recognize this is also child sexual abuse material; it's just newly created."

Meanwhile, major social media companies have cut the resources deployed towards scanning and reporting child exploitation by slashing jobs among their child and safety moderator teams.

"If major companies are unwilling to do the basics with CSAM detection, why would we think they would take all these extra steps in this AI world without regulation?" said

Sarah Gardner, CEO of the Heat Initiative, a Los Angeles-based child safety group. “We’ve witnessed that purely voluntary does not work.”

Related stories

Elon Musk says he'll drop his $97bn bid for OpenAI if it remains a non-profit

13 Feb 2025

Google parent Alphabet's revenues disappoint Wall Street amid stiff AI competition

4 Feb 2025

A man stalked a professor for six years. Then he used AI chatbots to lure strangers to her home

1 Feb 2025

More from Headlines

Trump administration

Trump removes Ice chief amid apparent frustration over rate of deportations

2h ago

Ukraine

US envoy to Ukraine hails Zelenskyy as 'embattled and courageous leader'

1h ago

Amazon rainforest

Brazilian city in Amazon declares emergency after huge sinkholes appear

2h ago