Revealed: US police prevented from viewing many online child sexual abuse reports, lawyers say

Publication Date: 2024-01-17

Author: Katie McQue

Section: Technology

Tags: Technology, Children, Social media, Meta, Facebook, news

Article URL: https://www.theguardian.com/technology/2024/jan/17/child-sexual-abuse-ai-moderator-police-meta-alphabet

Social media companies relying on artificial intelligence software to moderate their platforms are generating unviable reports on cases of child sexual abuse, preventing US police from seeing potential leads and delaying investigations of alleged predators, the Guardian can reveal. By law, US-based social media companies are required to report any child sexual abuse material detected on their platforms to the National Center for Missing & Exploited Children (NCMEC). NCMEC acts as a nationwide clearinghouse for leads about child abuse, which it forwards to the relevant law enforcement departments in the US and around the world. The organization said in its annual report that it received more than 32m reports of suspected child sexual exploitation from companies and the public in 2022, roughly 88m images, videos and other files. Meta is the largest reporter of these tips, with more than 27m, or 84%, generated by its Facebook, Instagram and WhatsApp platforms in 2022. NCMEC is partly funded by the Department of Justice, but it also receives private and corporate donations, including from Meta. NCMEC and Meta do not disclose the size of this donation. Social media companies, Meta included, use AI to detect and report suspicious material on their sites and employ human moderators to review some of the flagged content before sending it to law enforcement. However, US law enforcement agencies can only open AI-generated reports of child sexual abuse material (CSAM) by serving a search warrant to the company that sent them. Petitioning a judge for a warrant and waiting to receive one can add days or even weeks to the investigation process. “If the company has not indicated when they report the file to NCMEC that they have viewed the file prior to making the report, we cannot open it,” said Staca Shehan, vice-president of the analytical services division of NCMEC. “When we send it along to law enforcement, they cannot view it or open it without first serving legal process on the [social media company].” Due to US privacy protections under the fourth amendment, which prohibits unreasonable searches and seizures by the government, neither law enforcement officers nor NCMEC – which receives federal funding – are permitted to open reports of potential abuse without a search warrant unless the contents of a report have been reviewed first by a person at the social media company. These practices were adopted more than a decade ago, when a 2013 ruling in the US district court for Massachusetts stated NCMEC was acting as a government agent in investigations of the alleged spread of child abuse material online. Several federal courts have come to the same conclusion since then. A 2021 case in the ninth circuit court, which covers states on the west coast, held the position that law enforcement officers’ warrantless review of child abuse reports generated by Google’s AI was a violation of the fourth amendment. Since NCMEC personnel and law enforcement agents can’t legally look at the contents of a tip if an AI generated it but a human never reviewed it, investigations of alleged predators are stalling for up

to several weeks, which can lead to evidence being lost, according to child safety experts and attorneys. “Any delays [in viewing the evidence] are detrimental to ensuring the safety of the community,” as offenders go undetected for longer, said a California-based assistant US attorney who spoke on the condition of anonymity. “They are a risk to every child.” In some cases, after submitting a report, some social media companies disable the user’s account to prevent their continued activity on their platforms. This can result in the removal of evidence related to their suspected crimes from the platform’s servers. “This is frustrating,” said the California-based assistant US attorney. “By the time you have an account identified and have a warrant, there may be nothing there.” In response to a request from the Guardian, NCMEC said it does not keep a record of the number of tips it receives that are AI-generated. However, two federal prosecutors interviewed said they were not permitted to view most tips they receive from the major social media companies because AI has generated them. The AI-generated tips are often not investigated by law enforcement because they lack the specific information needed to obtain an initial probable cause affidavit that would persuade a judge to issue the search warrant, said a Massachusetts-based federal prosecutor, who requested not to be named. These tips require additional investigative legwork – extra time that prosecutors and their staff don’t have, he said. “Departments are underwater and don’t have the resources given the volume of tips, so we triage. And thus, unviable tips that aren’t immediately actionable sit in a drawer to be worked when there is time and resources,” said the attorney. “They don’t get acted on because the men and women doing this work will never be adequately resourced enough to have the time to make these tips actionable.” NCMEC’s Shehan called the potential delays “concerning”. “When you’re providing information to law enforcement about a possible crime of child sexual exploitation, everyone takes that seriously and they want to take action. If there are additional steps because of those types of barriers, it’s obviously concerning,” she said. Relying on AI for moderation puts the onus on a relatively small group of overworked law enforcement investigating these cases, who “drown in this heartbreaking work”, said the Massachusetts-based prosecutor. “AI may be a solution to treat their employees better, but social media companies will not find new child abuse material because the AI will only be ingesting old data points,” he said. “It’s not the solution for the betterment of the world of exploitation. You still need people to put their eyes on it.” In December, New Mexico’s attorney general’s office filed a lawsuit against Meta, alleging the company’s social networks had become marketplaces for child predators and that it had repeatedly failed to report illegal activity on its platforms. In response, Meta said it prioritizes fighting child sexual abuse material. The state attorney general laid the blame at Meta’s feet for the struggle to send viable tips. “Reports indicating the ineffectiveness of the company’s AI-generated cyber tip system demonstrate what we have laid out in our complaint – Mark Zuckerberg and Meta executives have purposefully prioritized profits over child safety,” Raúl Torrez said in a statement to the Guardian. “It is long past time the company reform and implement changes to its staffing levels, policies and algorithms to ensure children are safe, parents are informed and that law enforcement can effectively investigate and prosecute online sex crimes against children,” Torrez added. Despite the legal limitations on moderation AI, social media companies may increase its use in the near future. In 2023, OpenAI, the makers of ChatGPT, announced its GPT-4 engine could do the work of human content moderators, claiming the large language model was almost as accurate. The chatbot would, according to the company, help avoid the psychological trauma people can experience by viewing violent and abusive content for their jobs. However, child safety experts argue the AI software social media companies are utilizing for content moderation is only effective in identifying known images of child sexual abuse because the digital fingerprints of the images, known as hash values, are already known. For newly created images, or altered known images and videos, AI is not effective, said the attorneys interviewed. “There’s always a concern with cases of newly identified victims, and the material doesn’t have a hash value because they’re new,” said Kristina Korobov, senior attorney with Zero Abuse Project, a non-profit organization focused on combatting child abuse. “You’d see an increase in detection of newly discovered victims if humans were doing the work.” In the last year, the major tech companies Meta, Twitter and Alphabet have all slashed jobs from their teams responsible for moderation. Such cuts ultimately result in fewer reports of child abuse material being reviewed by humans at the companies, placing an even more significant burden on law enforcement as the potential number of search warrants needed increases, said Korobov. “Investigators say they are already drowning in cyber tips. The reality is these officers don’t have time,” said Korobov. “The bigger problem is the volume of cyber tips coming in and this is an extra step. We’re dealing with thousands of cyber tips that come into any state per year.” She argues that expanding the number of human content moderators would help ease the workload of law enforcement and said the recent moderator job cuts are “frustrating”. “It becomes a sickening realization that there are human beings at those companies who most likely have children they love in their lives, who have decided they can make more money by using a computer to do this,” she said. In a statement to the Guardian after the publication of this story, a spokesperson for Meta said: “It’s unfortunate that court rulings have increased the burden on law enforcement by requiring search warrants to open identical copies of content we’ve already reviewed and reported. Conflicting court decisions and constitutional interpretations add to the confusion. Our image-matching system finds copies of known child exploitation at a scale that would be impossible to do manually, and we work to detect new child exploitation content through technology, reports from our community, and investigations by our specialist child safety teams. We also continue to support NCMEC and law enforcement in prioritising reports, including by helping build NCMEC’s case management tool and labelling cybertips so they know which are urgent.” • This article was amended on 18 January 2024 to remove a reference to Meta holding a board seat at NCMEC. The company no longer holds that seat as of 2022. • In the US, call or text the Childhelp abuse hotline on 800-422-4453 or visit their website for more

resources and to report child abuse or DM for help. For adult survivors of child abuse, help is available at ascasupport.org. In the UK, the NSPCC offers support to children on 0800 1111, and adults concerned about a child on 0808 800 5000. The National Association for People Abused in Childhood (Napac) offers support for adult survivors on 0808 801 0331. In Australia, children, young adults, parents and teachers can contact the Kids Helpline on 1800 55 1800, or Bravehearts on 1800 272 831, and adult survivors can contact Blue Knot Foundation on 1300 657 380. Other sources of help can be found at Child Helplines International