TechXplore

Study shows AI image-generators being trained on explicit photos of children

December 20 2023, by MATT O'BRIEN and HALELUYA HADERO

David Thiel, chief technologist at the Stanford Internet Observatory and author of its report that discovered images of child sexual abuse in the data used to train artificial intelligence image-generators, poses for a photo on Wednesday, Dec. 20, 2023 in Obidos, Portugal. Credit: Camilla Mendes dos Santos via AP

Hidden inside the foundation of popular artificial intelligence image-generators are thousands of images of child sexual abuse, according to a new report that urges companies to take action to address a harmful flaw in the technology they built.

Those same images have made it easier for AI systems to produce realistic and explicit imagery of fake children as well as transform social media photos of fully clothed real teens into nudes, much to the alarm of schools and law enforcement around the world.

Until recently, anti-abuse researchers thought the only way that some unchecked AI tools produced abusive imagery of children was by essentially combining what they've learned from two separate buckets of online images—adult pornography and benign photos of kids.

But the Stanford Internet Observatory found more than 3,200 images of suspected child sexual abuse in the giant AI database LAION, an index of online images and captions that's been used to train leading AI image-makers such as Stable Diffusion. The watchdog group based at Stanford University worked with the Canadian Centre for Child Protection and other anti-abuse charities to identify the illegal material and report the original photo links to law enforcement. It said roughly 1,000 of the images it found were externally validated.

The response was immediate. On the eve of the Wednesday release of the Stanford Internet Observatory's report, LAION told The Associated Press it was temporarily removing its datasets.

LAION, which stands for the nonprofit Large-scale Artificial Intelligence Open Network, said in a statement that it "has a zero tolerance policy for illegal content and in an abundance of caution, we have taken down the LAION datasets to ensure they are safe before republishing them."

While the images account for just a fraction of LAION's index of some 5.8 billion images, the Stanford group says it is likely influencing the ability of AI tools to generate harmful outputs and reinforcing the prior abuse of real victims who appear multiple times.

It's not an easy problem to fix, and traces back to many generative AI projects being "effectively rushed to market" and made widely accessible because the field is so competitive, said Stanford Internet Observatory's chief technologist David Thiel, who authored the report.

"Taking an entire internet-wide scrape and making that dataset to train models is something that should have been confined to a research operation, if anything, and is not something that should have been open-sourced without a lot more rigorous attention," Thiel said in an interview.

Students walk on the Stanford University campus on March 14, 2019, in Stanford, Calif. Hidden inside the foundation of popular artificial intelligence image-generators are thousands of images of child sexual abuse, according to a new report from the Stanford Internet Observatory that urges technology companies to take action to address a harmful flaw in the technology they built. Credit: AP Photo/Ben Margot, File

A prominent LAION user that helped shape the dataset's development is London-based startup Stability AI, maker of the Stable Diffusion text-to-image models. New versions of Stable Diffusion have made it much harder to create harmful content, but an older version introduced last year—which Stability AI says it didn't release—is still baked into other applications and tools and remains "the most popular model for generating explicit imagery," according to the Stanford report.

"We can't take that back. That model is in the hands of many people on their local machines," said Lloyd Richardson, director of information technology at the Canadian Centre for Child Protection, which runs Canada's hotline for reporting online sexual exploitation.

Stability AI on Wednesday said it only hosts filtered versions of Stable Diffusion and that "since taking over the exclusive development of Stable Diffusion, Stability AI has taken proactive steps to mitigate the risk of misuse."

"Those filters remove unsafe content from reaching the models," the company said in a prepared statement. "By removing that content before it ever reaches the model, we can help to prevent the model from generating unsafe content."

LAION was the brainchild of a German researcher and teacher, Christoph Schuhmann, who told the AP earlier this year that part of the reason to make such a huge visual database publicly accessible was to ensure that the future of AI development isn't controlled by a handful of powerful companies.

"It will be much safer and much more fair if we can democratize it so that the whole research community and the whole general public can benefit from it," he said.

Much of LAION's data comes from another source, Common Crawl, a repository of data constantly trawled from the open internet, but Common Crawl's executive director, Rich Skrenta, said it was "incumbent on" LAION to scan and filter what it took before making use of it.

LAION said this week it developed "rigorous filters" to detect and remove illegal content before releasing its datasets and is still working to improve those filters. The Stanford report acknowledged LAION's developers made some attempts to filter out "underage" explicit content but might have done a better job had they consulted earlier with child safety experts.

Many text-to-image generators are derived in some way from the LAION database, though it's not always clear which ones. OpenAI, maker of DALL-E and ChatGPT, said it doesn't use LAION and has fine-tuned its models to refuse requests for sexual content involving minors.

David Thiel, chief technologist at the Stanford Internet Observatory and author of its report that discovered images of child sexual abuse in the data used to train artificial intelligence image-generators, poses for a photo on Wednesday, Dec. 20, 2023, in Óbidos, Portugal. Credit: Camilla Mendes dos Santos via AP

Google built its text-to-image Imagen model based on a LAION dataset but decided against making it public in 2022 after an audit of the database "uncovered a wide range of inappropriate content including pornographic imagery, racist slurs, and harmful social stereotypes."

Trying to clean up the data retroactively is difficult, so the Stanford Internet Observatory is calling for more drastic measures. One is for anyone who's built training sets off of LAION-5B—named for the more than 5 billion image-text pairs it contains—to "delete them or work with intermediaries to clean the material." Another is to effectively make an older version of Stable Diffusion disappear from all but the darkest corners of the internet.

"Legitimate platforms can stop offering versions of it for download," particularly if they are frequently used to generate abusive images and have no safeguards to block them, Thiel said.

As an example, Thiel called out CivitAI, a platform that's favored by people making AI-generated pornography but which he said lacks safety measures to weigh it against making images of children. The report also calls on AI company Hugging Face, which distributes the training data for models, to implement better methods to report and remove links to abusive material.

Hugging Face said it is regularly working with regulators and child safety groups to identify and remove abusive material. Meanwhile, CivitAI said it has "strict policies" on the generation of images depicting children and has rolled out updates to provide more safeguards. The company also said it is working to ensure its policies are "adapting and growing" as the technology evolves.

The Stanford report also questions whether any photos of children—even the most benign—should be fed into AI systems without their family's consent due to protections in the federal Children's Online Privacy Protection Act.

Rebecca Portnoff, the director of data science at the anti-child sexual abuse organization Thorn, said her organization has conducted research that shows the prevalence of AI-generated images among abusers is small, but growing consistently.

Developers can mitigate these harms by making sure the datasets they use to develop AI models are clean of abuse materials. Portnoff said there are also opportunities to mitigate harmful uses down the line after models are already in circulation.

Tech companies and child safety groups currently assign videos and images a "hash"—unique digital signatures—to track and take down child abuse materials. According to Portnoff, the same concept can be applied to AI models that are being misused.

"It's not currently happening," she said. "But it's something that in my opinion can and should be done."

Citation: Study shows AI image-generators being trained on explicit photos of children (2023, December 20) retrieved 14 May 2024 from https://techxplore.com/news/2023-12-ai-image-generators-explicit-photos-children.html