Child abuse images removed from AI image-generator training source, researchers say

August 31 2024, by Matt O'brien

Students walk on the Stanford University campus on March 14, 2019, in Stanford, Calif. Credit: AP Photo/Ben Margot, File

Artificial intelligence researchers said Friday they have deleted more than 2,000 web links to suspected child sexual abuse imagery from a dataset used to train popular AI image-generator tools.

The LAION research dataset is a huge index of online images and captions that's been a source for leading AI image-makers such as Stable Diffusion and Midjourney.

But a report last year by the Stanford Internet Observatory found it contained links to sexually explicit images of children, contributing to the ease with which some AI tools have been able to produce photorealistic deepfakes that depict children.

That December report led LAION, which stands for the nonprofit Large-scale Artificial Intelligence Open Network, to immediately remove its dataset. Eight months later, LAION said in a blog post that it worked with the Stanford University watchdog group and anti-abuse organizations in Canada and the United Kingdom to fix the problem and release a cleaned-up dataset for future AI research.

Stanford researcher David Thiel, author of the December report, commended LAION for significant improvements but said the next step is to withdraw from distribution the "tainted models" that are still able to produce child abuse imagery.

One of the LAION-based tools that Stanford identified as the "most popular model for generating explicit imagery" — an older and lightly filtered version of Stable Diffusion — remained easily accessible until Thursday, when the New York-based company Runway ML removed it from the AI model repository Hugging Face. Runway said in a statement Friday it was a "planned deprecation of research models and code that have not been actively maintained."

The cleaned-up version of the LAION dataset comes as governments around the world are taking a closer look at how some tech tools are being used to make or distribute illegal images of children.

San Francisco's city attorney earlier this month filed a lawsuit seeking to shut down a group of websites that enable the creation of AI-generated nudes of women and girls. The alleged distribution of child sexual abuse images on the messaging app Telegram is part of what led French authorities to bring charges on Wednesday against the platform's founder and CEO, Pavel Durov.

Durov's arrest "signals a really big change in the whole tech industry that the founders of these platforms can be held personally responsible," said David Evan Harris, a researcher at the University of California, Berkeley who recently reached out to Runway asking about why the problematic AI image-generator was still publicly accessible. It was taken down days later.

Citation: Child abuse images removed from AI image-generator training source, researchers say (2024, August 31) retrieved 2 September 2024 from https://techxplore.com/news/2024-08-child-abuse-images-ai-image.html