www.theatlantic.com /technology/archive/2024/10/muah-ai-hack-child-abuse/680300/

People Are Asking AI for Child Pornography

Caroline Mimbs Nyce ⋮ 8-10 minutes ⋮ 10/18/2024

The Age of AI Child Abuse Is Here

For maybe the first time, the scale of the problem is coming into view.

Illustration by The Atlantic. Source: Getty.

Produced by ElevenLabs and News Over Audio (NOA) using AI narration.

Muah.AI is a website where people can make AI girlfriends—chatbots that will talk via text or voice and send images of themselves by request. Nearly 2 million users have registered for the service, which describes its technology as "uncensored." And, judging by data purportedly lifted from the site, people may be using its tools in their attempts to create child-sexual-abuse material, or CSAM.

Last week, Joseph Cox, at *404 Media*, was the first to report on the data set, after an anonymous hacker brought it to his attention. What Cox found was profoundly disturbing: He reviewed one prompt that included language about orgies involving "newborn babies" and "young kids." This indicates that a user had asked Muah.AI to respond to such scenarios, although whether the program did so is unclear. Major AI platforms, including ChatGPT, employ filters and other moderation tools intended to block generation of content in response to such prompts, but less prominent services tend to have fewer scruples.

People have used AI software to generate sexually exploitative images of real individuals. Earlier this year, pornographic deepfakes of Taylor Swift circulated on X and Facebook. And child-safety advocates have warned repeatedly that generative AI is now being widely used to create sexually abusive imagery of real children, a problem that has surfaced in schools across the country.

The Muah.AI hack is one of the clearest—and most public—illustrations of the broader issue yet: For maybe the first time, the scale of the problem is being demonstrated in very clear terms.

I spoke with Troy Hunt, a well-known security consultant and the creator of the data-breach-tracking site HaveIBeenPwned.com, after seeing a thread he posted on X about the hack. Hunt had also been sent the Muah.AI data by an anonymous source: In reviewing it, he found many examples of users prompting the program for child-sexual-abuse material. When he searched the data for *13-year-old*, he received more than 30,000 results, "many alongside prompts describing sex acts." When he tried *prepubescent*, he got 26,000 results.

He estimates that there are tens of thousands, if not hundreds of thousands, of prompts to create CSAM within the data set.

Hunt was surprised to find that some Muah.AI users didn’t even try to conceal their identity. In one case, he matched an email address from the breach to a LinkedIn profile belonging to a C-suite executive at a “very normal” company. “I looked at his email address, and it’s literally, like, his first name dot last name at gmail.com,” Hunt told me. “There are lots of cases where people make an attempt to obfuscate their identity, and if you can pull the right strings, you’ll figure out who they are. But this guy just didn’t even try.” Hunt said that CSAM is traditionally associated with fringe corners of the internet. “The fact that this is sitting on a mainstream website is what probably surprised me a little bit more.”

Recommended Reading

-
-
-

Last Friday, I reached out to Muah.AI to ask about the hack. A person who runs the company’s Discord server and goes by the name Harvard Han confirmed to me that the website had been breached by a hacker. I asked him about Hunt’s estimate that as many as hundreds of thousands of prompts to create CSAM may be in the data set. “That’s impossible,” he told me. “How is that possible? Think about it. We have 2 million users. There’s no way 5 percent is fucking pedophiles.” (It is possible, though, that a relatively small number of users are responsible for a large number of prompts.)

When I asked him whether the data Hunt has are real, he initially said, “Maybe it is possible. I am not denying.” But later in the same conversation, he said that he wasn’t sure. Han said that he had been traveling, but that his team would look into it.

The site’s staff is small, Han stressed over and over, and has limited resources to monitor what users are doing. Fewer than five people work there, he told me. But the site seems to have built a modest user base: Data provided to me from Similarweb, a traffic-analytics company, suggest that Muah.AI has averaged 1.2 million visits a month over the past year or so.

Han told me that last year, his team put a filtering system in place that automatically blocked accounts using certain words—such as *teenagers* and *children*—in their prompts. But, he told me, users complained that they were being banned unfairly. After that, the site adjusted the filter to stop automatically blocking accounts, but to still prevent images from being generated based on those keywords, he said.

At the same time, however, Han told me that his team does not check whether his company is generating child-sexual-abuse images for its users. He assumes that a lot of the requests

to do so are “probably denied, denied, denied,” he said. But Han acknowledged that savvy users could likely find ways to bypass the filters.

He also offered a kind of justification for why users might be trying to generate images depicting children in the first place: Some Muah.AI users who are grieving the deaths of family members come to the service to create AI versions of their lost loved ones. When I pointed out that Hunt, the cybersecurity consultant, had seen the phrase *13-year-old* used alongside sexually explicit acts, Han replied, “The problem is that we don’t have the resources to look at every prompt.” (After Cox’s article about Muah.AI, the company said in a post on its Discord that it plans to experiment with new automated methods for banning people.)

In sum, not even the people running Muah.AI know what their service is doing. At one point, Han suggested that Hunt might know more than he did about what’s in the data set. That sites like this one can operate with such little regard for the harm they may be causing raises the bigger question of whether they should exist at all, when there’s so much potential for abuse.

Meanwhile, Han took a familiar argument about censorship in the online age and stretched it to its logical extreme. “I’m American,” he told me. “I believe in freedom of speech. I believe America is different. And we believe that, hey, AI should not be trained with censorship.” He went on: “In America, we can buy a gun. And this gun can be used to protect life, your family, people that you love—or it can be used for mass shooting.”

Federal law prohibits computer-generated images of child pornography when such images feature real children. In 2002, the Supreme Court ruled that a total ban on computer-generated child pornography violated the First Amendment. How exactly existing law will apply to generative AI is an area of active debate. When I asked Han about federal laws regarding CSAM, Han said that Muah.AI only provides the AI processing, and compared his service to Google. He also reiterated that his company’s word filter could be blocking some images, though he is not sure.

Whatever happens to Muah.AI, these problems will certainly persist. Hunt told me he’d never even heard of the company before the breach. “And I’m sure that there are dozens and dozens more out there.” Muah.AI just happened to have its contents turned inside out by a data hack. The age of cheap AI-generated child abuse is very much here. What was once hidden in the darkest corners of the internet now seems quite easily accessible—and, equally worrisome, very difficult to stamp out.

About the Author