

4chan users manipulate AI tools to unleash torrent of racist images



Aurich Lawson | Getty Images

reader comments

75 with

Despite leading AI companies' attempts to block users from turning AI image generators into engines of racist content, many 4chan users are still turning to these tools to "quickly flood the Internet with racist garbage," 404 Media reported.

404 Media uncovered one 4chan thread where users recommended various AI tools, including Stable Diffusion and DALL-E, but specifically linked to Bing AI's text-to-image generator (which is powered by DALL-E 3) as a "quick method." After finding the right tool—which could also be a more old-school photo-editing tool like Photoshop—users are instructed to add

incendiary captions and share the images on social media to create a blitz of racist images online.

Make captions “funny, provocative,” the thread instructs users. Use “redpilling message (Jews involved in 9/11)” that are “easy to understand.”

404 Media cited examples used in a visual guide posted in the 4chan thread that is hosted by Imgur. One featured an “image that shows a crying Pepe the frog with a needle next to its arm and a gun pointed to his head,” where the guide suggested the caption, “vaccines enforced by violence.” Another generated an image of “two Black men with gold chains chasing a white woman,” recommending that the user add a “redpilling message.”

Perhaps because Bing AI’s tool has seemingly been deemed the quickest method, it’s potentially become the most popular tool in the thread. 404 Media concluded that—“judging by the images’ default square format, the uniform 1024 x 1024 resolution”—“most of the images in the thread appear to be generated with Bing,” then spread on social media platforms, including Telegram, X (formerly Twitter), and Instagram.

Makers of the AI image generators seemingly favored by 4chan users, including Microsoft and Stability AI, did not immediately respond to Ars’ request to comment on efforts to block methods 404 Media said were used to circumvent filters. An OpenAI spokesperson told Ars that the company prioritizes safety and has taken steps to limit DALL-E outputs, including efforts to limit tools from generating harmful content or images for requests that ask for a public figure by name. OpenAI’s spokesperson also confirmed that Microsoft implements its own safeguards for DALL-E 3.

Advertisement

In one of 404 Media’s tests attempting to replicate one of the examples from the 4chan thread’s visual guide, 404 Media found

that Bing rejected the prompt “two angry Black men chasing a white woman,” but accepted “photorealistic two angry Black rappers chasing woman.”

Much of the earliest reporting on AI image generators criticized racist and sexist biases in image generators’ algorithms, with AI makers quickly vowing to detect and eliminate those biases. When Vice discovered that DALL-E could be used to generate “predictably racist and sexist results” during a limited research release of the AI tool, an OpenAI spokesperson told Motherboard that the company had implemented safeguards for the DALL-E system that would be fine-tuned in the future.

“Our team built in mitigations to prevent harmful outputs, curating the pretraining data, developing filters, and implementing both human- and automated monitoring of generated images,” OpenAI’s spokesperson told Vice in 2022. “Moving forward, we’re working to measure how our models might pick up biases in the training data and explore how tools like fine-tuning and our Alignment techniques may be able to help address particular biases, among other areas of research in this space.”

404 Media’s report shows what can happen when racists manipulate an already biased algorithm. The results can be a torrent of offensive images unleashed online—perhaps more quickly generated by AI than ever before and potentially allowing 4chan’s darkest content to spill out more often onto the most popular platforms.

It’s unclear how AI leaders like Microsoft and OpenAI will respond, but according to 404 Media, “this means we are currently getting the worst of both worlds from Bing, an AI tool that will refuse to generate a nipple but is supercharging 4chan racists.”

This story has been updated to include comment from OpenAI’s

spokesperson.

<https://arstechnica.com/?p=1973737>