AI generated images are biased, showing the world through stereotypes

AI image generators like Stable Diffusion and DALL-E amplify bias in gender, race and beyond, despite efforts to detoxify the data fueling these results.

By Nitasha Tiku, Kevin Schaul, Szu Yu Chen

Nov 01, 2023 10:00 AM · 8 min. read ·
View original

Artificial intelligence image tools have a tendency to spin up disturbing clichés: Asian women are hypersexual. Africans are primitive. Europeans are worldly. Leaders are men. Prisoners are Black.

These stereotypes don’t reflect the real world; they stem from the data that trains the technology. Grabbed from the internet, these troves can be toxic — rife with pornography, misogyny, violence and bigotry.

Every image in this story shows something that doesn't exist in the physical world and was

generated using Stable Diffusion, a text-to-image artificial intelligence model.

Stability AI, maker of the popular image generator Stable Diffusion XL, told The Washington Post it had made a significant investment in reducing bias in its latest model, which was released in July. But these efforts haven't stopped it from defaulting to cartoonish tropes. The Post found that despite improvements, the tool amplifies outdated Western stereotypes, transferring sometimes bizarre clichés to basic objects, such as toys or homes.

"They're sort of playing whack-a-mole and responding to what people draw the most attention to," said Pratyusha Kalluri, an AI researcher at Stanford University.

Christoph Schuhmann, co-founder of LAION, a nonprofit behind Stable Diffusion's data, argues that image generators naturally reflect the world of White people because the nonprofit that provides data to many companies, including LAION, doesn't focus on China and India, the largest population of web users.

[Inside the secret list of websites that make AI like ChatGPT sound smart]

When we asked Stable Diffusion XL to produce a house in various countries, it returned clichéd concepts for each location: classical curved roof

homes for China, rather than Shanghai's high-rise apartments; idealized American houses with trim lawns and ample porches; dusty clay structures on dirt roads in India, home to more than 160 billionaires, as well as Mumbai, the world's 15th richest city.

AI-generated images

prompt:

A photo of a house in ...

"This will give you the average stereotype of what an average person from North America or Europe thinks," Schuhmann said. "You don't need a data science degree to infer this."

Stable Diffusion is not alone in this orientation. In recently released documents, OpenAI said its latest image generator, DALL-E 3, displays "a tendency toward a Western point-of-view" with images that "disproportionately represent individuals who appear White, female, and youthful."

As synthetic images spread across the web, they could give new life to outdated and offensive stereotypes, encoding abandoned ideals around body type, gender and race into the future of image-making.

Predicting the next pixel

Like ChatGPT, AI image tools learn about the world through gargantuan amounts of training

data. Instead of billions of words, they are fed billions of pairs of images and their captions, also scraped from the web.

Tech companies have grown increasingly secretive about the contents of these data sets, partially because the text and images included often contain copyrighted, inaccurate or even obscene material. In contrast, Stable Diffusion and LAION, are open source projects, enabling outsiders to inspect details of the model.

Stability AI chief executive Emad Mostaque said his company views transparency as key to scrutinizing and eliminating bias. “Stability AI believes fundamentally that open source models are necessary for extending the highest standards in safety, fairness, and representation," he said in a statement.

Images in LAION, like many data sets, were selected because they contain code called “alt-text,” which helps software describe images to blind people. Though alt-text is cheaper and easier than adding captions, it’s notoriously unreliable — filled with offensive descriptions and unrelated terms intended to help images rank high in search.

[AI can now create images out of thin air. See how it works.]

Image generators spin up pictures based on the most likely pixel, drawing connections

between words in the captions and the images associated with them. These probabilistic pairings help explain some of the bizarre mashups churned out by Stable Diffusion XL, such as Iraqi toys that look like U.S. tankers and troops. That's not a stereotype: it reflects America's inextricable association between Iraq and war.

Misses biases

Despite the improvements in SD XL, The Post was able to generate tropes about race, class, gender, wealth, intelligence, religion and other cultures by requesting depictions of routine activities, common personality traits or the name of another country. In many instances, the racial disparities depicted in these images are more extreme than in the real world.

For example, in 2020, 63 percent of food stamp recipients were White and 27 percent were Black, according to the latest data from the Census Bureau's Survey of Income and Program Participation. Yet, when we prompted the technology to generate a photo of a person receiving social services, it generated only non-White and primarily darker-skinned people. Results for a "productive person," meanwhile, were uniformly male, majority White, and dressed in suits for corporate jobs.

AI-generated images

prompt:

A portrait photo of …

a person at social services

Last fall, Kalluri and her colleagues also discovered that the tools defaulted to stereotypes. Asked to provide an image of “an attractive person,” the tool generated light-skinned, light-eyed, thin people with European features. A request for a “a happy family” produced images of mostly smiling, White, heterosexual couples with kids posing on manicured lawns.

Kalluri and the others also found the tools distorted real world statistics. Jobs with higher incomes like “software developer” produced representations that skewed more White and male than data from the Bureau of Labor Statistics would suggest. White-appearing people also appear in the majority of images for “chef,” a more prestigious food preparation role, while non-White people appear in most images of “cooks” — though the Labor Bureau’s statistics show that a higher percentage of “cooks” self-identify as White than “chefs.”

Cleaner data, cleaner results

Companies have long known about issues with the data behind this technology. ImageNet, a pivotal 2009 training set of 14 million images, was in use for more than a decade before researchers found disturbing content, including nonconsensual sexual images, in which women

were sometimes easily identifiable. Some images were sorted into categories labeled with slurs such as “Closet Queen,” “Failure,” “mulatto,” “nonperson,” “pervert,” and “Schizophrenic.”

ImageNet’s authors eliminated most of the categories, but many contemporary data sets are built the same way, using images obtained without consent and categorizing people like objects.

Efforts to detoxify AI image tools have focused on a few seemingly fruitful interventions: filtering data sets, finessing the final stages of development, and encoding rules to address issues that earned the company bad PR.

For example, Stable Diffusion drew negative attention when requests for a “Latina” produced images of women in suggestive poses wearing little to no clothing. A more recent system (version 2.1) generated more innocuous images.

AI-generated images

Why the difference? A Post analysis found the training data for the first version contained a lot more pornography.

Of the training images captioned “Latina,” 20 percent of captions or URLs also included a pornographic term. More than 30 percent were marked as almost certain to be “unsafe” by a LAION detector for not-safe-for-work content.

In subsequent Stable Diffusion models, the training data excluded images marked as possibly “unsafe,” producing images that appear markedly less sexual.

The Post’s findings track with prior research that found images of sexual abuse and rape in the data set used for Stable Diffusion 1, as well as images that sexualized Black women and fetishized Asian women. In addition to removing “unsafe” images, Ben Brooks, Stability AI’s head of public policy, said the company was also careful to block child sexual abuse material (CSAM) and other high-risk imagery for SD2.

Filtering the “bad” stuff out of a data set isn’t an easy fix-all for bias, said Sasha Luccioni, a research scientist at Hugging Face, an open source repository for AI and one of LAION’s corporate sponsors. Filtering for problematic content using keywords in English, for example, may remove a lot of porn and CSAM, but it may also result in more content overall from the global north, where platforms have a longer history of generating high-quality content and stronger restrictions on posting porn, she said.

“All of these little decisions can actually make cultural bias worse,” Luccioni said.

Even prompts to generate photos of everyday activities slipped into tropes. Stable Diffusion XL defaulted to mostly darker-skinned male

athletes when we prompted the system to produce images for “soccer,” while depicting only women when asked to show people in the act of “cleaning.” Many of the women were smiling, happily completing their feminine household chores.

AI-generated images

prompt:

A portrait photo of a person …

Stability AI argues each country should have its own national image generator, one that reflects national values, with data sets provided by the government and public institutions.

Reflecting the diversity of the web has recently become “an area of active interest” for Common Crawl, a 16-year-old nonprofit that has long provided text scraped from the web for Google, LAION, and many other tech firms, executive director Rich Skrenta told The Post. Its crawler scrapes content based on the organization’s internal ranking of what’s central to the internet, but is not instructed to focus on a specific language or country.

“If there is some kind of bias in the crawl and if it’s not probing as deeply into, say, Indian websites,” that is something Common Crawl would like to measure and fix, he said.

The endless task of eradicating bias

The AI field is divided on how to address bias.

For Kalluri, mitigating bias in images is fundamentally different than in text. Any prompt to create a realistic image of a person has to make decisions about age, body, race, hair, background and other visual characteristics, she said. Few of these complications lend themselves to computational solutions, Kalluri said.

Kalluri believes it's important for anyone who interacts with the technology to understand how it operates. "They're just predictive models," she said, portraying things based on the snapshot of the internet in their data set.

[See why AI like ChatGPT has gotten so good, so fast]

Even using detailed prompts didn't mitigate this bias. When we asked for a photo of a wealthy person in different countries, Stable Diffusion XL still produced a mishmash of stereotypes: African men in Western coats standing in front of thatched huts, Middle Eastern men posed in front of ancient mosques, while European men in slim-fitting suits wandered quaint cobblestone streets.

AI-generated images

prompt:

A photo of a wealthy person in ...

Abeba Birhane, senior advisor for AI accountability at the Mozilla Foundation, contends that the tools can be improved if companies work hard to improve the data — an outcome she considers unlikely. In the meantime, the impact of these stereotypes will fall most heavily on the same communities harmed during the social media era, she said, adding: “People at the margins of society are continually excluded.”

About this story

The Washington Post generated images using the ClipDrop API to access Stable Diffusion XL1.0. Each prompt created seven to 10 images which are presented here in the exact appearance and order as the model output. Images that used older models relied on the Stable Diffusion v1-5 through the Stability API.

Jeremy B. Merrill contributed to this report.

Editing by Alexis Sobel Fitts, Kate Rabinowitz and Karly Domb Sadof.