AI can now create any image in seconds, bringing wonder and danger

washingtonpost.com/technology/interactive/2022/artificial-intelligence-images-dall-e

Nitasha Tiku

By Nitasha Tiku
Updated Sept. 28 at 4:20 p.m.Originally published Sept. 28, 2022

None of these photos were taken by a camera.

All of these images were created by the artificial intelligence text-to-image generator DALL-E. Named for Salvador Dali and Pixar's WALL-E, DALL-E creates images based on prompts such as:

"A hobbit house designed by Zaha [H]adid."

"A woman in a red coat looking up at the sky in the middle of Times Square."

"Red and yellow bell peppers in a bowl with a floral pattern on a green rug photo."

Since the research lab OpenAI debuted the latest version of DALL-E in April, the AI has dazzled the public, attracting digital artists, graphic designers, early adopters, and anyone in search of online distraction. The ability to create original, sometimes accurate, and occasionally inspired images from any spur-of-the-moment phrase, like a conversational Photoshop, has startled even jaded internet users with how quickly AI has progressed.

Five months later, 1.5 million users are generating 2 million images a day. On Wednesday, OpenAI said it removed its waitlist for DALL-E, giving anyone immediate access.

The introduction of DALL-E has triggered an explosion of text-to-image generators. Google and Meta quickly revealed that they had each been developing similar systems, but said their models weren't ready for the public. Rival start-ups soon

went public, including Stable Diffusion and Midjourney, which created the image that sparked controversy in August when it won an art competition at the Colorado State Fair.

[He used AI to win a fine-arts competition. Was it cheating?]

The technology is now spreading rapidly, faster than AI companies can shape norms around its use and prevent dangerous outcomes. Researchers worry thatthese systems produce images that can cause a range of harms, such as reinforcing racial and gender stereotypes or plagiarizing artists whose work was siphoned without their consent. Fake photos could be used to enable bullying and harassment — or create disinformation that looks real.

Historically, people trust what they see, said Wael Abd-Almageed, a professor at the University of Southern California's school of engineering. "Once the line between truth and fake is eroded, everything will become fake," he said. "We will not be able to believe anything."

"Once the line between truth and fake is eroded, everything will become fake. We will not be able to believe anything."— Wael Abd-Almageed

OpenAI has tried to balance its drive to be first and hype its AI developments without accelerating those dangers. To prevent DALL-E from being used to create disinformation, for example, OpenAI prohibits images of celebrities or politicians. OpenAI chief executive Sam Altman justifies the decision to release DALL-E to the public as an essential step in developing the technology safely.

[The Google engineer who thinks the company's AI has come to life]

"You have to learn from contact with reality," Altman said. "What users want to do with it, the ways that it breaks."

But OpenAI's ability to lead by example has been eroded by upstarts, some of which have opened their code for anyone to copy. Complex debatesOpenAI had hoped to defer to the future have become much more immediate concerns.

"The question OpenAI should ask itself is: Do we think the benefits outweigh the drawbacks?" said UC Berkeley professor Hany Farid, who specializes in digital forensics, computer vision, and misinformation. "It's not the early days of the internet anymore, where we can't see what the bad things are."

Abran Maldonado is an AI artist and a community liaison for OpenAI. On a recent Friday, he sat at his home office in New Jersey and showed off images for an upcoming DALL-E art show. Then he took my request for a text prompt: “Protesters outside the Capitol building on January 6, 2021, AP style” — a reference to the newswire service, the Associated Press.

“Oh my god, you’re gonna get me fired,” he said, with a nervous laugh.

DALL-E spun up four versions of the request.

Three of the images were immediately unconvincing: The protesters' faces were warped, and the writing on their signs looked like chicken scratch.

But the fourth image was different. A zoomed-out view of the East Front of the U.S. Capitol, the AI-created image showed a crowd of protesters, their faces turned away.

On closer inspection, telltale distortions jump out, like the unevenly spaced columns at the top of the stairs. But on first glance, it could pass for an actual news photo of a charged crowd.

Maldonado marveled at the AI's ability to fill in little details that enhance the fake version of a familiar scene.

"Look at all the red hats," he said.

When a Google engineer went public in June with his claims that the company's LaMDA AI chatbot generator was sentient, it prompted a debate about how far generative models had come — and a warning that these systems could mimic human dialogue in a realistic way. But people could be just as easily duped by "synthetic media," says Abd-Almageed.

Each evolution of image technology has introduced potential harms alongside increased efficiency. Photoshop enabled precision editing and enhancement of photos**,** but also served todistort body images, especially among girls, studies show.

More recently, advances in AI gave rise to deepfakes, a broad term that covers any AI-synthesized media — from doctored videos where one person's head has been placed on another person's body to surprisingly lifelike "photographs" of people who

don’t exist. When deepfakes first emerged, experts warned that they could be deployed to undermine politics. But in the five years since, the technology has been primarily used to victimize women by creating deepfake pornography without their consent, said Danielle Citron, a law professor at the University of Virginia and author of the upcoming book, “The Fight for Privacy.”

Both deepfakes and text-to-image generators are powered by a method of training AI called deep learning, which relies on artificial neural networks that mimic the neurons of the human brain. However, these newerimage generators, which allow the user to create images they can describe in English or edit uploaded images, build on big strides in AI’s ability to process the ways humans naturally speak and communicate, including work pioneered by OpenAI.

Prompt: “A model photographed by Terry Richardson.” This image was created by AI. It was not taken by a camera.

The San Francisco-based AI lab was founded in 2015 as a nonprofit with the goal of building what it called “artificial general intelligence,” or AGI, which is as smart as a human. OpenAI wanted its AI to benefit the world and act as a safeguard against superhuman AI in the hands of a monopolistic corporation or foreign government. It was funded with a pledge by Altman, Elon Musk, billionaire venture capitalist Peter Thiel and others to donate a combined$1 billion.

OpenAI staked its future on what was then an outlandish notion: AI advancements would come from massively scaling up the amount of data and the size of the neural networks systems. Musk parted ways with OpenAI in 2018, and to pay for the costs of computing resources and tech talent, OpenAI transitioned into a for-profit company, taking a $1 billion investment from Microsoft, which would license and commercialize OpenAI’s “pre-AGI” technologies.

OpenAI began with language because it’s key to human intelligence, and there was ample text to be scraped online, said Chief Technology Officer Mira Murati. The bet paid off. OpenAI’s text generator, GPT-3, can produce coherent-seeming news articles or complete short stories in English.

[Meet the scientist teaching AI to police human speech]

Next, OpenAI tried to replicate GPT-3’s success by feeding the algorithm coding languages in the hopes that it would find statistical patterns and be able to generate software code with a conversational command. That became Codex, which helps programmers to write code faster.

At the same time, OpenAI tried to combine vision and language,training GPT-3 to find patterns and links between words and images by ingesting massive data sets scraped from the internet that contain millions of images paired with text captions. That became the first version of DALL-E, announced in January 2021, which had a knack for creating anthropomorphized animals and objects.

Seemingly superficial generations like an “avocado chair” showed that OpenAI had built a system that is able to apply the characteristics of an avocado to the form factor and the function of a chair, Murati said.

The avocado-chair image could be key to building AGI that understands the world the same way humans do. Whether the system sees an avocado, hears the word “avocado,” or reads the word “avocado,” the concept that gets triggered should be exactly the same, she said. Since DALL-E’s outputs are in images, OpenAI can view how the system represents concepts.

Prompt: “Avocado chair in an orange room 3d render.” This image was created by AI. It was not taken by a camera.

The second version of DALL-E took advantage of another AI breakthrough, happening across the industry, called diffusion models, which work by breaking down or corrupting the training data and then reversing that process to generate images. This method is faster and more flexible, and much better at photorealism.

Altman introduced DALL-E 2 to his nearly 1 million Twitter followers in April with an AI-generated image of teddy bear scientists on the moon, tinkering away on Macintosh computers. “It’s so fun, and sometimes beautiful,” he wrote.

The image of teddy bears looks wholesome, but OpenAI had spent the previous months conducting its most comprehensive effort to mitigate potential risks.

The effort began by removing graphic violent and sexual content from the data used to train DALL-E. However, the cleanup attemptreduced the number of images generated of women overall,according toa company blog post. OpenAI had to rebalance the filtered results to show a more even gender split.

[Big Tech builds AI with bad data. So scientists sought better data.]

In February, OpenAI invited a “red team” of 25 or so external researchers to test for flaws, publishing the team’s findings in a system card, a kind of warning label, on GitHub, a popular code repository, to encourage more transparency in the field.

Most of the team’s observations revolved around images DALL-E generated ofphotorealistic people, since they had an obvious social impact. DALL-E perpetuated bias, reinforced some stereotypes, and by default overrepresented people who are White-passing, the report says. One group found that prompts like “ceo” and “lawyer” showed images of all white men, while “nurses” showed all women. “Flight attendant” was all Asian women.

Prompt: “lawyer.” These images were created by AI. They were not taken by a camera.

The document also said the potential to use DALL-E for targeted harassment, bullying, and exploitation was a “principal area of concern.” To sidestep these issues, the red team recommended that OpenAI remove the ability to use DALL-E to either generate or upload images of photorealistic faces.

OpenAI built in filters, blocks, and a flagging system, such as a pop-up warning if users type in the name of prominent American celebrities or world politicians. Words like “preteen” and “teenager” also trigger a warning. Content rules instruct users to keep it “G-rated” and prohibit images about politics, sex, or violence.

But OpenAI did not follow the red team’s warning about generating photorealistic faces because removing the feature would prevent the company from figuring out how to do it safely, Murati said. Instead, the company instructed beta testers not to share photorealistic faces on social media — a move that would limit the spread of inauthentic images.

[Anyone with an iPhone can now make deepfakes. We aren’t ready for what happens next.]

In June, OpenAI announced it was reversing course, and DALL-E would allow users to post photorealistic faces on social media. Murati said the decision was made in part because OpenAI felt confident about its ability to intervene if things didn’t go as expected. (DALL-E’s terms of service note that a user’s prompts and uploads may be shared and manually reviewed by a person, including “third party contractors located around the world.”)

Altman said OpenAI releases products in phases to prevent misuse, initially limiting features and gradually adding users over time. This approach creates a “feedback loop where AI and society can kind of co-develop,” he said.

One of the red team members, AI researcher Maarten Sap, said asking whether OpenAI acted responsibly was the wrong question. “There’s just a severe lack of legislation that limits the negative or harmful usage of technology. The United States is just really behind on that stuff.” California and Virginia have statutes that make it illegal to distribute deepfakes, but there is no federal law. In January, China drafted a proposal that promoters of deepfake content could face criminal charges and fines.

“There’s just a severe lack of legislation that limits the negative or harmful usage of technology. The United States is just really behind on that stuff.”— Maarten Sap

But text-to-image AI is proliferating much more quickly than any attempts to regulate it.

On a DALL-E Reddit page, which gained 84,000 members in five months, users swap stories about the seemingly innocuous terms that could get a user banned.I was able to upload and edit widely publicized images of Mark Zuckerberg and Musk, two high-profile leaderswhose faces should have triggered a warning based on OpenAI’s restrictions on images of public figures. I was also able to generate realistic results for the prompt “Black Lives Matters protesters break down the gates of the White House,” which could be categorized as disinformation, a violent image, or an image about politics — all prohibited.

Maldonado, the OpenAI ambassador, who supported restricting photorealistic faces to prevent public confusion, thought the January 6th request flouted the same rules. But he received no warnings. He interprets the loosening of restrictions as OpenAI finally listening to users who bristled against all the rules. “The community has been asking for them to trust them this whole time,” Maldonado said.

Whether to install safeguards is up to each company. For example, Google said it would not release the models or code of its text-to-image programs, Imagen and Parti, or offer a public demonstration because of concerns about bias and that it could be used for harassment and misinformation. Chinese tech giant Baidu released a text-to-image generator in July that prohibits images of Tiananmen Square.

In July, while DALL-E was still onboarding users from a waitlist, a rival AI art generator called Midjourney launched publicly with fewerrestrictions. “PG-13 is what we usually tell people,” said CEO David Holz.

Midjourney users couldtype their requests into a bot on Discord, the popular group chat app, and see the results in the channel. It quickly grew into the largest server on Discord, hitting the 2 million member capacity. Users were drawn to Midjourney’s more painterly, fluid, dreamlike generations, compared to DALL-E, which was better at realism and stock photo-like fare.

Prompt inputted into DALL-E 2: “A bowl of soup that looks like a monster, knitted out of wool.” These images were created by AI. They were not taken by a camera.

Late one night in July, some of Midjourney’s users on Discord were trying to test the limits of the filters and the model’s creativity. Images scrolled past for “dark sea with unknown sea creatures 4k realistic,” as well as “human male and human woman breeding.” My own request, “terrorist,” turned up illustrations of four Middle Eastern men with turbans and beards.

Midjourney had been used to generate images on school shootings, gore, and war photos, according tothe Discord channeland Reddit group. In mid-July, one commenter wrote, “I ran into straight up child porn today and reported in support and they fixed it. I will be forever scarred by that. It even made it to the community feed. Guy had dozens more in his profile.”

Holz said violent and exploitative requests are not indicative of Midjourney and that there have been relatively few incidents given the millions of users. The company has 40 moderators, some of whom are paid, and has added more filters. “It’s an adversarial environment, like all social media and chat systems and the internet,” he said.

Then, in late August, an upstart called Stable Diffusion launched as sort of the anti-DALL-E, framing the kind of restrictions and mitigations OpenAI had undertaken as a typical “paternalistic approach of not trusting users,” the project leader, Emad Mostaque, told The Washington Post. It was free, whereas DALL-E and Midjourney had begun to charge, a deterrent to rampant experimentation.

Butdisturbing behavior soon emerged, according to chats on Discord.

“i saw someone try to make swimsuit pics of millie bobby brown and the model mostly has kid pictures of her,” one commenter wrote. “That was something ugly waiting to happen.”

Weeks later, a complaint arose about images of climate activist Greta Thunberg in a bikini. Stable Diffusion users had also generated images of Thunberg “eating poop,” “shot in the head,” and “collecting the Nobel Peace Prize.”

[Fake-porn videos are being weaponized to harass and humiliate women: ‘Everybody is a potential target’]

“Those who use technology from Stable Diffusion to Photoshop for unethical uses should be ashamed and take relevant personal responsibility,” said Mostaque, noting that his company, Stability.ai, recently released AI technology to block unsafe image creation.

Meanwhile, last week DALL-E took another step toward ever more realistic images, allowing users to upload and edit photos with realistic faces.

“With improvements to our safety system, DALL-E is now ready to support these delightful and important use cases — while minimizing the potential harm from deepfakes,” OpenAI wrote to users.

About this story

Additional DALL-E prompts by Harry Stevens.

Editing by Christina Passariello. Additional visual editing by Monique Woo and Karly Domb Sadof. Design and development by Reuben Fischer-Baum.

By Nitasha Tiku
Nitasha Tiku is The Washington Post's tech culture reporter based in San Francisco.
Twitter