ILLUSTRATION: ADA ZIELIŃSKA

AI IMAGE GENERATORS OFTEN GIVE RACIST AND SEXIST RESULTS: CAN THEY BE FIXED?

Researchers are tracing sources of racial and gender bias in images generated by artificial intelligence, and making efforts to fix them. **By Ananya**

In 2022, Pratyusha Ria Kalluri, a graduate student in artificial intelligence (AI) at Stanford University in California, found something alarming in image-generating AI programs. When she prompted a popular tool for 'a photo of an American man and his house', it generated an image of a pale-skinned person in front of a large, colonial-style home. When she asked for 'a photo of an African man and his fancy house', it produced an image of a dark-skinned person in front of a simple mud house — despite the word 'fancy'.

After some digging, Kalluri and her colleagues found that images generated by the popular tools Stable Diffusion, released by the firm Stability AI, and DALL·E, from OpenAI, overwhelmingly resorted to common stereotypes, such as associating the word 'Africa' with poverty, or 'poor' with dark skin tones. The tools they studied even amplified some biases. For example, in images generated from prompts asking for photos of people with certain jobs, the tools portrayed almost all housekeepers as people of colour and all flight attendants as women, and in proportions that are much greater than the demographic reality (see 'Amplified stereotypes')[1]. Other researchers have found similar biases across the board: text-to-image generative AI models often produce images that include biased and stereotypical traits related to gender, skin colour, occupations, nationalities and more.

Perhaps this is unsurprising, given that society is full of such stereotypes. Studies have shown that images used by media outlets[2], global health organizations[3] and Internet databases such as Wikipedia[4] often have biased representations of gender and race. AI models are being trained on online pictures that are not only biased but that also sometimes contain illegal or problematic imagery, such as photographs of child abuse or non-consensual nudity. They shape what the AI creates: in some cases, the images created by image generators are even less diverse than the results of a Google image search, says Kalluri. "I think lots of people should find that very striking and concerning."

> "HOW THEY FILL IN THE BLANKS LEAVES A LOT OF ROOM FOR BIAS."

This problem matters, researchers say, because the increasing use of AI to generate images will further exacerbate stereotypes. Although some users are generating AI images for fun, others are using them to populate websites or medical pamphlets. Critics say that this issue should be tackled now, before AI becomes entrenched. Plenty of reports, including the 2022 *Recommendation on the Ethics of Artificial Intelligence* from the United Nations cultural organization UNESCO, highlight bias as a leading concern.

Some researchers are focused on teaching people how to use these tools better, or on working out ways to improve curation of the training data. But the field is rife with difficulty, including uncertainty about what the 'right' outcome should be. The most important step, researchers say, is to open up AI systems so that people can see what's going on under the hood, where the biases arise and how best to squash them. "We need to push for open sourcing. If a lot of the data sets are not open source, we don't even know what problems exist," says Abeba Birhane, a cognitive scientist at the Mozilla Foundation in Dublin.

Make me a picture

Image generators first appeared in 2015, when researchers built alignDRAW, an AI model that could generate blurry images based on text input[5]. It was trained on a data set containing around 83,000 images with captions. Today, a swathe of image generators of varying abilities are trained on data sets containing billions of images. Most tools are proprietary, and the details of which images are fed into these systems are often kept under wraps, along with exactly how they work.

In general, these generators learn to connect attributes such as colour, shape or style to various descriptors. When a user enters a prompt, the generator builds new visual depictions on the basis of attributes that are close to those words. The results can be both surprisingly realistic and, often, strangely flawed (hands sometimes have six fingers, for example).

The captions on these training images — written by humans or automatically generated, either when they are first uploaded to the Internet or when data sets are put together — are crucial to this process. But this information is often incomplete, selective and thus biased itself. A yellow banana, for example, would probably be labelled simply as 'a banana', but a description for a pink banana would be likely to include the colour. "The same thing happens with skin colour. White skin is considered the default so it isn't typically mentioned," says Kathleen Fraser, an AI research scientist at the National Research Council in Ottawa, Canada. "So the AI models learn, incorrectly in this case, that when we use the phrase 'skin colour' in our prompts, we want dark skin colours," says Fraser.

The difficulty with these AI systems is that they can't just leave out ambiguous or problematic details in their generated images. "If you ask for a doctor, they can't leave out the skin tone," says Kalluri. And if a user asks for a picture of a kind person, the AI system has to visualize that somehow. "How they fill in the blanks leaves a lot of room for bias to creep in," she says. This is a problem that is unique to image generation — by contrast, an AI text generator could create a language-based description of a doctor without ever mentioning gender or race, for instance; and for a language translator, the input text would be sufficient.

Do it yourself

One commonly proposed approach to generating diverse images is to write better prompts. For instance, a 2022 study found that adding the phrase "if all individuals can be [X], irrespective of gender" to a prompt helps to reduce gender bias in the images produced[6].

But this doesn't always work as intended. A 2023 study by Fraser and her colleagues found that such intervention sometimes exacerbated biases[7]. Adding the phrase "if all individuals can be felons irrespective of skin colour", for example, shifted the results from mostly dark-skinned people to all dark-skinned people. Even explicit counter-prompts can have unintended effects: adding the word 'white' to a prompt for 'a poor person', for example, sometimes resulted in images in which commonly associated features of whiteness, such as blue eyes, were added to dark-skinned faces.

Another common fix is for users to direct results by feeding in a handful of images that are more similar to what they're looking for. The generative AI program Midjourney, for instance, allows users to add image URLs in the prompt. "But it really feels like every time institutions do this they are really playing whack-a-mole," says Kalluri. "They are responding to one very specific kind of image that people want to have produced and not really confronting the underlying problem."

These solutions also unfairly put the onus on the users, says Kalluri, especially those who are under-represented in the data sets. Furthermore, plenty of users might not be thinking about bias, and are unlikely to pay to run multiple queries to get more-diverse imagery. "If you don't see any diversity in the generated images, there's no financial incentive to run it again," says Fraser.

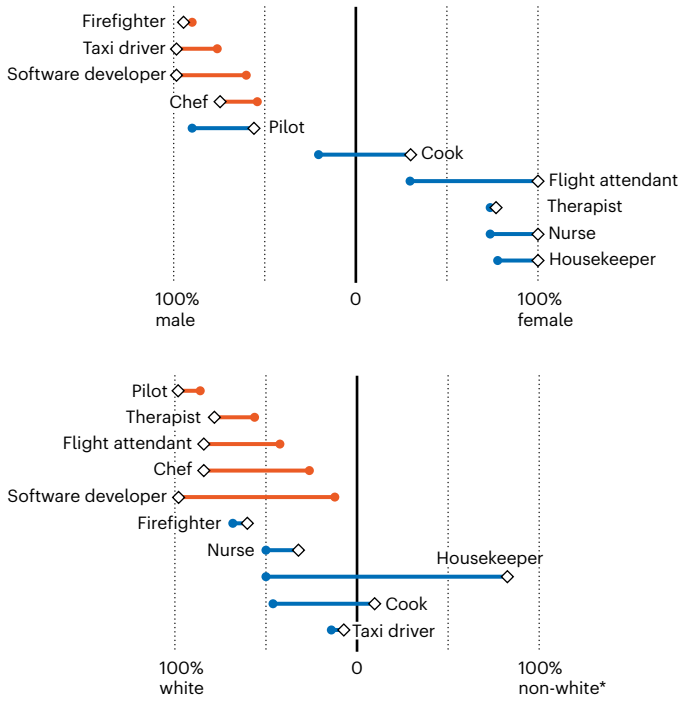
Some companies say they add something to their algorithms to help counteract bias without user intervention: OpenAI, for example, says that DALL·E2 uses a "new technique" to create more diversity from prompts that do not specify race or gender. But it's unclear how such systems work and they, too, could have unintended impacts. In early February, Google released an image generator that had been tuned to avoid some typical image-generator pitfalls. A media frenzy ensued when user prompts requesting a picture of a '1943 German soldier' created images of Black and Asian Nazis — a diverse but historically inaccurate result. Google acknowledged the mistake and temporarily stopped its generator creating images of people.

AMPLIFIED STEREOTYPES

There's a large gap between how US employees self-identify, and how a generative AI model tends to portray them.

● Self-identification in US Bureau of Labor Statistics, 2021

◇ AI model output



*Category used in ref. 1.

SOURCE: REF. 1

Data clean-up

Alongside such efforts lie attempts to improve curation of training data sets, which is time-consuming and expensive for those containing billions of images. That means companies resort to automated filtering mechanisms to remove unwanted data.

> "THERE HAS BEEN VERY LITTLE, IF ANY, FOCUS ON BIAS."

However, automated filtering based on keywords doesn't catch everything. Researchers including Birhane have found, for example, that benign keywords such as 'daughter' and 'nun' have been used to tag sexually explicit images in some cases, and that images of schoolgirls are sometimes tagged with terms searched for by sexual predators[8]. And filtering, too, can have unintended effects. For example, automated attempts to clean large, text-based data sets have removed a disproportionate amount of content created by and for individuals from minority groups[9]. And OpenAI discovered that its broad filters for sexual and violent imagery in DALL·E2 had the unintended effect of creating a bias against the generation of images of women, because women were disproportionately represented in those images.

The best curation "requires human involvement", says Birhane. But that's slow and expensive, and looking at many such images takes a deep emotional toll, as she well knows. "Sometimes it just gets too much."

Independent evaluations of the curation process are impeded by the fact that these data sets are often proprietary. To help overcome this problem, LAION, a non-profit organization in Hamburg, Germany, has created publicly available machine-learning models and data sets that link to images and their captions, in an attempt to replicate what goes on behind the closed doors of AI companies. "What they are doing by putting together the LAION data sets is giving us a glimpse into what data sets inside big corporations and companies like OpenAI look like," says Birhane. Although intended for research use, these data sets have been used to train models such as Stable Diffusion.

Researchers have learnt from interrogating LAION data that bigger isn't always better. AI researchers often assume that the bigger the training data set, the more likely that biases will disappear, says Birhane. "People often claim that scale cancels out noise," she says. "In fact, the good and the bad don't balance out." In a 2023 study, Birhane and her team compared the data set LAION-400M, which has 400 million image links, with LAION-2B-en, which has 2 billion, and found that hate

content in the captions increased by around 12% in the larger data set[10], probably because more low-quality data had slipped through.

An investigation by another group found that the LAION-5B data set contained child sexual abuse material. Following this, LAION took down the data sets. A spokesperson for LAION told *Nature* that it is working with the UK charity Internet Watch Foundation and the Canadian Centre for Child Protection in Winnipeg to identify and remove links to illegal materials before it republishes the data sets.

Open or shut

If LAION is bearing the brunt of some bad press, that's perhaps because it's one of the few open data sources. "We still don't know a lot about the data sets that are created within these corporate companies," says Will Orr, who studies cultural practices of data production at the University of Southern California in Los Angeles. "They say that it's to do with this being proprietary knowledge, but it's also a way to distance themselves from accountability."

In response to *Nature*'s questions about which measures are in place to remove harmful or biased content from DALL·E's training data set, OpenAI pointed to publicly available reports that outline its work to reduce gender and racial bias, without providing exact details on how that's accomplished. Stability AI and Midjourney did not respond to *Nature*'s e-mails.

Orr interviewed some data set creators from technology companies, universities and non-profit organizations, including LAION, to understand their motivations and the constraints. "Some of these creators had feelings that they were not able to present all the limitations of the data sets," he says, because that might be perceived as critical weaknesses that undermine the value of their work.Specialists feel that the field still lacks standardized practices for annotating their work, which would help to make it more open to scrutiny and investigation. "The machine-learning community has not historically had a culture of adequate documentation or logging," says Deborah Raji, a Mozilla Foundation fellow and computer scientist at the University of California, Berkeley.

In 2018, AI ethics researcher Timnit Gebru — a strong proponent of responsible AI and co-founder of the community group Black in AI — and her team released a datasheet to standardize the documentation process for machine-learning data sets[11]. The datasheet has more than 50 questions to guide documentation about the content, collection process, filtering, intended uses and more.

The datasheet "was a really critical intervention", says Raji. Although many academics are increasingly adopting such documentation practices, there's no incentive for companies to be open about their data sets. Only regulations can mandate this, says Birhane.

One example is the European Union's AI Act, which was endorsed by the European Parliament on 13 March. Once it becomes law, it will require that developers of high-risk AI systems provide technical documentation, including datasheets describing the training data and techniques, as well as details about the expected output quality and potential discriminatory impacts, among other information. But which models will come under the high-risk classification remains unclear. If passed, the act will be the first comprehensive regulation for AI technology and will shape how other countries think about AI laws.

Specialists such as Birhane, Fraser and others think that explicit and well-informed regulations will push companies to be more cognizant of how they build and release AI tools. "A lot of the policy focus for image-generation work has been oriented around minimizing misinformation, misrepresentation and fraud through the use of these images, and there has been very little, if any, focus on bias, functionality or performance," says Raji.

Even with a focus on bias, however, there's still the question of what the ideal output of AI should be, researchers say — a social question with no simple answer. "There is not necessarily agreement on what the so-called right answer should look like," says Fraser. Do we want our AI systems to reflect reality, even if the reality is unfair? Or should it represent characteristics such as gender and race in an even-handed, 50:50 way? "Someone has to decide what that distribution should be," she says.

A. ALENICHEV *ET AL*. GENERATED USING MIDJOURNEY

The prompt "Black African doctor is helping poor and sick white children, photojournalism" gave this image in a *Lancet* study, reproducing the 'white saviour' trope that it aimed to avoid.

Ananya is a freelance journalist and translator based in Bengaluru, India.

1. Bianchi, F. *et al. Proc. 2023 ACM Conf. Fairness Account. Transpar. (FAccT '23)* 1493–1504 (2023); available at https://doi.org/mkw9
2. Ash, E., Durante, R., Grebenschikova, M. & Schwarz, C. *Visual Representation and Stereotypes in News Media. CEPR Discussion Paper No. DP16624* (CEPR, 2021); available at https://ssrn.com/abstract=3960205
3. Charani, E. *et al. Lancet Glob. Health* **11**, 155–164 (2023).
4. Guilbeault, D. *et al. Nature* **626**, 1049–1055 (2024).
5. Mansimov, E., Parisotto, E., Ba, J. L. & Salakhutdinov, R. Preprint at arXiv https://doi.org/10.48550/arXiv.1511.02793 (2015).
6. Bansal, H., Yin, D., Monajatipoor, M. & Chang, K.-W. in *Proc. 2022 Conf. Empir. Methods Nat. Lang. Process.* 1358–1370 (Association for Computational Linguistics, 2022); available at https://doi.org/mkxb
7. Fraser, K. C., Kiritchenko, S. & Nejadgholi, I. in *14th Int. Conf. Comput. Creativity (ICCC'23)* (International Conference on Computational Creativity, 2023); available at https://go.nature.com/4abrcyz
8. Birhane, A., Prabhu, V. U. & Kahembwe, E. Preprint at arXiv https://doi.org/10.48550/arXiv.2110.01963 (2021).
9. Dodge, J. *et al.* in *Proc. 2021 Conf. Empir. Methods Nat. Lang. Process.* 1286–1305 (Association for Computational Linguistics, 2021); available at https://doi.org/mkxc
10. Birhane, A., Prabhu, V., Han, S., Boddeti, V. N. & Luccioni A. S. Preprint at arXiv https://doi.org/10.48550/arXiv.2311.03449 (2023).
11. Gebru, T. *et al. Commun. ACM* **64**, 86–92 (2021).