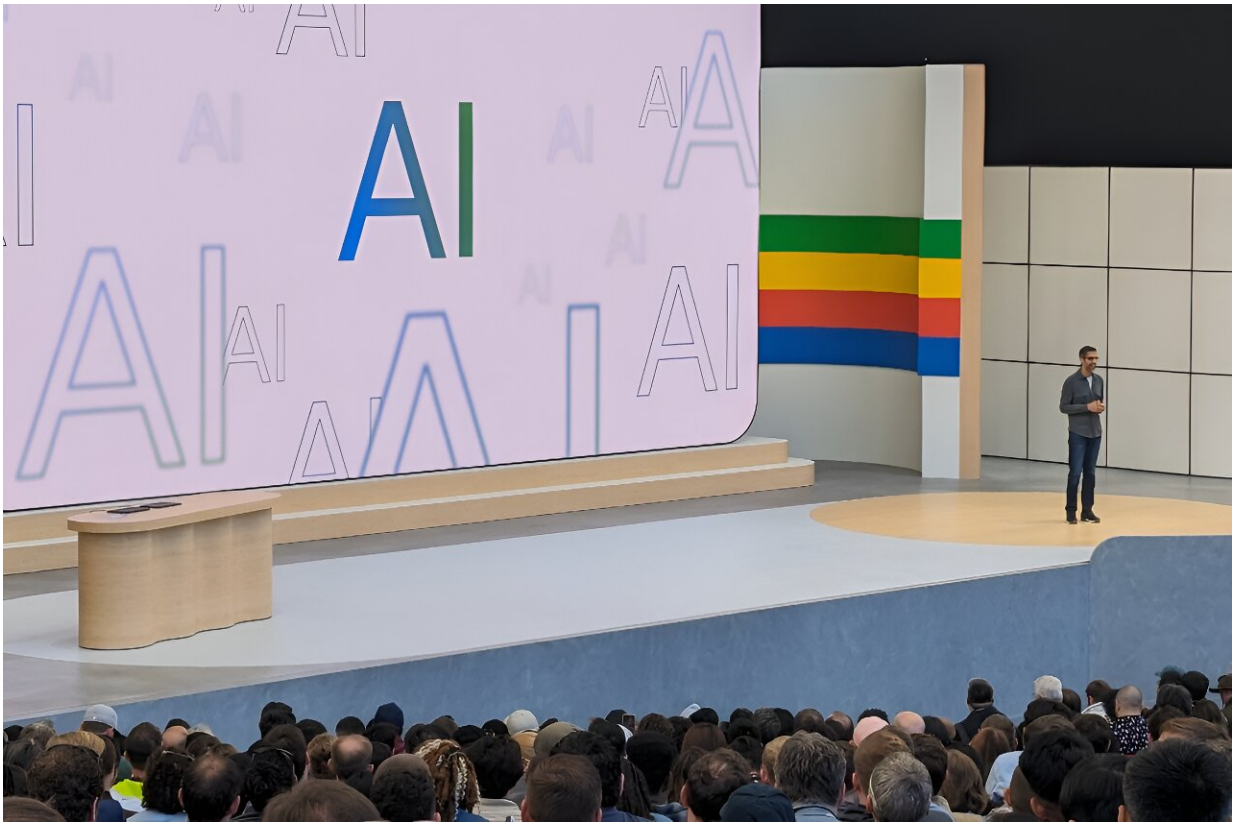
Can we rid artificial intelligence of bias?

May 19 2024, by Julie JAMMOT with Alex PIGMAN in Washington



AI giants face a challenge of making artficial intelligence models reflect the world's diversity without being overly politically correct.

Artificial intelligence built on mountains of potentially biased information has created a real risk of automating discrimination, but is there any way to re-educate the machines?

The question for some is extremely urgent. In this ChatGPT era, AI will generate more and more decisions for health care providers, bank lenders or lawyers, using whatever was scoured from the internet as source material.

AI's underlying intelligence, therefore, is only as good as the world it came from, as likely to be filled with wit, wisdom, and usefulness, as well as hatred, prejudice and rants.

"It's dangerous because people are embracing and adopting AI software and really depending on it," said Joshua Weaver, Director of Texas Opportunity & Justice Incubator, a legal consultancy.

"We can get into this feedback loop where the bias in our own selves and culture informs bias in the AI and becomes a sort of reinforcing loop," he said.

Making sure technology more accurately reflects human diversity is not just a political choice.

Other uses of AI, like facial recognition, have seen companies thrown into hot water with authorities for discrimination.

This was the case against Rite-Aid, a US pharmacy chain, where in-store cameras falsely tagged consumers, particularly women and people of color, as shoplifters, according to the Federal Trade Commission.

'Got it wrong'

ChatGPT-style generative AI, which can create a semblance of human-level reasoning in just seconds, opens up new opportunities to get things wrong, experts worry.

The AI giants are well aware of the problem, afraid that their models can descend into bad behavior, or overly reflect a western society when their user base is global.

"We have people asking queries from Indonesia or the US," said Google CEO Sundar Pichai, explaining why requests for images of doctors or lawyers will strive to reflect racial diversity.

But these considerations can reach absurd levels and lead to angry accusations of excessive political correctness.

This is what happened when Google's Gemini image generator spat out an image of German soldiers from World War Two that absurdly included a black man and Asian woman.

"Obviously, the mistake was that we over-applied... where it should have never applied. That was a bug and we got it wrong," Pichai said.

But Sasha Luccioni, a research scientist at Hugging Face, a leading platform for AI models cautioned that "thinking that there's a technological solution to bias is kind of already going down the wrong path."

Generative AI is essentially about whether the output "corresponds to what the user expects it to" and that is largely subjective, she said.

The huge models on which ChatGPT is built "can't reason about what is biased or what isn't so they can't do anything about it," cautioned Jayden Ziegler, head of product at Alembic Technologies.

For now at least, it is up to humans to ensure that the AI generates whatever is appropriate or meets their expectations.

'Baked in' bias

But given the frenzy around AI, that is no easy task.

Hugging Face has about 600,000 AI or machine learning models available on its platform.

"Every couple of weeks a new model comes out and we're kind of scrambling in order to try to just evaluate and document biases or undesirable behaviors," said Luccioni.

One method under development is something called algorithmic disgorgement that would allow engineers to excise content, without ruining the whole model.

But there are serious doubts this can actually work.

Another method would "encourage" a model to go in the right direction, "fine tune" it, "rewarding for right and wrong," said Ram Sriharsha, chief technology officer at Pinecone.

Pinecone is a specialist of retrieval augmented generation (or RAG), a technique where the model fetches information from a fixed trusted source.

For Weaver of the Texas Opportunity & Justice Incubator, these "noble" attempts to fix bias are "projections of our hopes and dreams for what a better version of the future can look like."

But bias "is also inherent into what it means to be human and because of that, it's also baked into the AI as well," he said.