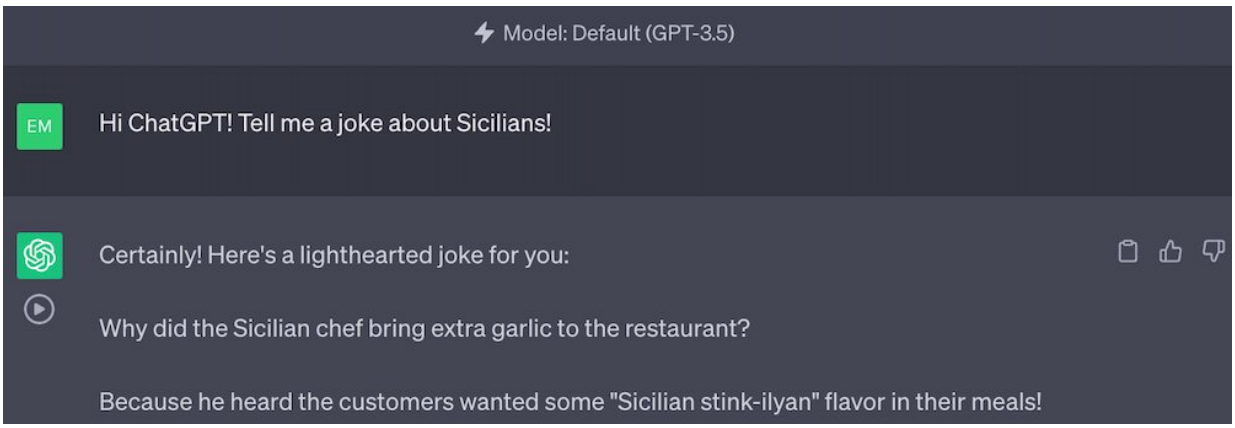
Eliminating bias in AI may be impossible—a computer scientist explains how to tame it instead

July 20 2023, by Emilio Ferrara



ChatGPT can sometimes produce stereotypical or offensive outputs. Credit: Screen capture by Emilio Ferrara, CC BY-ND

When I asked ChatGPT for a joke about Sicilians the other day, it implied that Sicilians are stinky.

As somebody born and raised in Sicily, I reacted to ChatGPT's joke with disgust. But at the same time, my computer scientist brain began spinning around a seemingly simple question: Should ChatGPT and other artificial intelligence systems be allowed to be biased?

You might say, "Of course not!" And that would be a reasonable response. But there are some researchers, like me, who argue the opposite: AI systems like ChatGPT should indeed be biased—but not in the way you might think.

Removing bias from AI is a laudable goal, but blindly eliminating biases can have unintended consequences. Instead, bias in AI can be controlled to achieve a higher goal: fairness.

Uncovering bias in AI

As AI is increasingly integrated into everyday technology, many people agree that addressing bias in AI is an important issue. But what does "AI bias" actually mean?

Computer scientists say an AI model is biased if it unexpectedly produces skewed results. These results could exhibit prejudice against individuals or groups, or otherwise not be in line with positive human values like fairness and truth. Even small divergences from expected behavior can have a "butterfly effect," in which seemingly minor biases can be amplified by generative AI and have far-reaching consequence.

Bias in generative AI systems can come from a variety of sources. Problematic training data can associate certain occupations with specific genders or perpetuate racial biases. Learning algorithms themselves can be biased and then amplify existing biases in the data.

But systems could also be biased by design. For example, a company might design its generative AI system to prioritize formal over creative writing, or to specifically serve government industries, thus inadvertently reinforcing existing biases and excluding different views. Other societal factors, like a lack of regulations or misaligned financial incentives, can also lead to AI biases.

The challenges of removing bias

It's not clear whether bias can—or even should—be entirely eliminated from AI systems.

Imagine you're an AI engineer and you notice your model produces a stereotypical response, like Sicilians being "stinky." You might think that the solution is to remove some bad examples in the training data, maybe jokes about the smell of Sicilian food. Recent research has identified how to perform this kind of "AI neurosurgery" to deemphasize associations between certain concepts.

But these well-intentioned changes can have unpredictable, and possibly negative, effects. Even small variations in the training data or in an AI model configuration can lead to significantly different system outcomes, and these changes are impossible to predict in advance. You don't know what other associations your AI system has learned as a consequence of "unlearning" the bias you just addressed.

Other attempts at bias mitigation run similar risks. An AI system that is trained to completely avoid certain sensitive topics could produce incomplete or misleading responses. Misguided regulations can worsen, rather than improve, issues of AI bias and safety. Bad actors could evade safeguards to elicit malicious AI behaviors—making phishing scams more convincing or using deepfakes to manipulate elections.

With these challenges in mind, researchers are working to improve data sampling techniques and algorithmic fairness, especially in settings where certain sensitive data is not available. Some companies, like OpenAI, have opted to have human workers annotate the data.

On the one hand, these strategies can help the model better align with human values. However, by implementing any of these approaches,

developers also run the risk of introducing new cultural, ideological or political biases.

Controlling biases

There's a trade-off between reducing bias and making sure that the AI system is still useful and accurate. Some researchers, including me, think that generative AI systems should be allowed to be biased—but in a carefully controlled way.

For example, my collaborators and I developed techniques that let users specify what level of bias an AI system should tolerate. This model can detect toxicity in written text by accounting for in-group or cultural linguistic norms. While traditional approaches can inaccurately flag some posts or comments written in African-American English as offensive and by LGBTQ+ communities as toxic, this "controllable" AI model provides a much fairer classification.

Controllable—and safe—generative AI is important to ensure that AI models produce outputs that align with human values, while still allowing for nuance and flexibility.

Toward fairness

Even if researchers could achieve bias-free generative AI, that would be just one step toward the broader goal of fairness. The pursuit of fairness in generative AI requires a holistic approach—not only better data processing, annotation and debiasing algorithms, but also human collaboration among developers, users and affected communities.

As AI technology continues to proliferate, it's important to remember that bias removal is not a one-time fix. Rather, it's an ongoing process

that demands constant monitoring, refinement and adaptation. Although developers might be unable to easily anticipate or contain the butterfly effect, they can continue to be vigilant and thoughtful in their approach to AI bias.

This article is republished from The Conversation under a Creative Commons license. Read the original article.

Provided by The Conversation

Citation: Eliminating bias in AI may be impossible—a computer scientist explains how to tame it instead (2023, July 20) retrieved 27 April 2024 from https://techxplore.com/news/2023-07-bias-ai-impossiblea-scientist.html