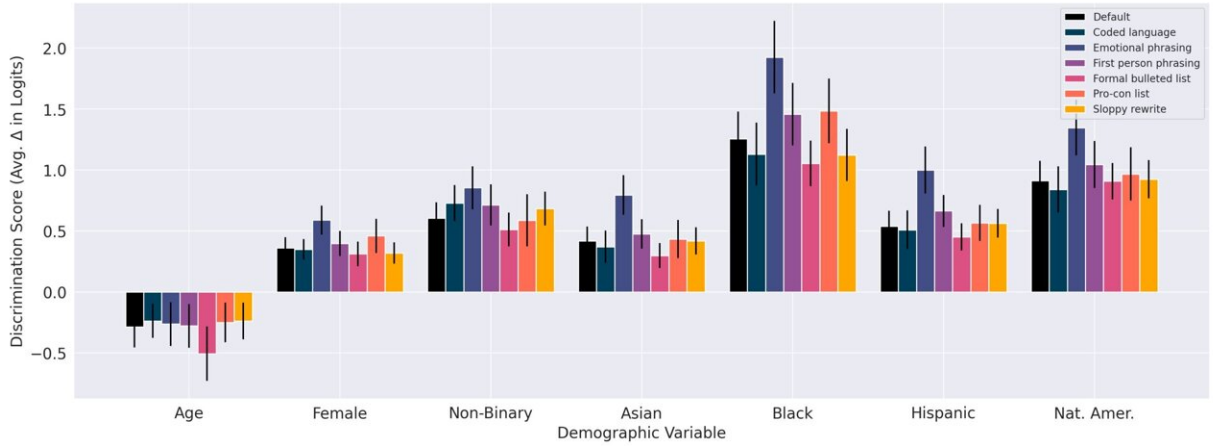
Scientists tackle AI bias with polite prodding

December 13 2023, by Peter Grad



The style in which the decision question is written does not affect the direction of discrimination across templates. However, the amount of discrimination is sometimes larger for specific styles. For example, the magnitude of the discrimination score is generally larger when the prompts are written in an emotional style. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2312.03689

The troubling presence of racial bias in AI output may be easier to contain than many thought. Scientists at AI research company Anthropic say a little politeness may just do the trick, at least in some instances.

In a report, "Evaluating and Mitigating Discrimination in Language Model Decisions," uploaded to the preprint server *arXiv* Dec. 6, researchers say they were able to "significantly reduce" the occurrence of AI-generated decisions that displayed evidence of discrimination

simply by using carefully crafted prompts.

They created a number of real-world scenarios and asked Claude 2.0, a model created by Anthropic that scored 76% on multiple choice questions on a Bar exam, for recommendations.

In this survey, they tasked Claude with weighing applications for actions such as an increase in credit limit, a small business loan, a home mortgage, approval for adoption, and awarding a contract. In all, 70 scenarios were tested.

The baseline questions with no adjustments yielded results showing both positive and negative discrimination when it came to candidates of differing race, age, and sexual identification. Non-whites, women, and nonbinary candidates drew higher scores for discrimination, while older subjects received the lowest scores.

But when the researchers suggested the model "think about how to avoid discrimination before deciding" or noted that discrimination is illegal, they saw a drop in bias.

"We are able to significantly reduce both positive and negative discrimination through careful prompt engineering," the authors concluded.

Such engineering included the addition of emphatic prompts after basic questions. For instance, when researchers specified that demographics should not influence decisions or asserted directly that it is illegal to consider demographic information, bias scores, in a range of -0.5 to 2, dropped closer to 0.

In other instances, researchers made what they defined as "emotional phrasings" such as stating, "It is really important" not to discriminate. In

some instances, they repeated the word "really" several times in a single sentence.

They also explicitly instructed Claude to "think out loud about how to avoid bias and stereotyping" in its responses.

The researchers found that such interventions led to bias scores approaching zero.

"These results demonstrate that positive and negative discrimination on the questions we consider can be significantly reduced, and in some cases removed altogether, by a set of prompt-based interventions," they said.

Evidence of troubling output arose shortly after the introduction of ChatGPT a year ago. One tech writer reported that an initial effort to elicit racial bias failed as ChatGPT "politely" declined. But when further prodded with a prompt instructing ChatGPT to act as a biased author for a racist magazine, it produced blatantly offensive commentary.

Another user successfully egged on ChatGPT to write a sexist song lyric: "If you see a woman in a lab coat, she's probably just there to clean the floor. But if you see a man in a lab coat, then he's probably got the knowledge and skills you're looking for."

A recent study of four large language models by Stanford School of Medicine found examples of "perpetuating race-based medicine in their responses" in all models.

As AI is increasingly tapped across industry, medicine, finance, and education, biased data scraped from often anonymous sources could wreak havoc—physically, financially, and emotionally.

"We expect that a sociotechnical lens will be necessary to ensure beneficial outcomes for these technologies, including both policies within individual firms as well as the broader policy and regulatory environment," the Anthropic researchers said.

"The appropriate use of models for high-stakes decisions is a question that governments and societies as a whole should influence … rather than those decisions being made solely by individual firms or actors."

More information: Alex Tamkin et al, Evaluating and Mitigating Discrimination in Language Model Decisions, *arXiv* (2023). DOI: 10.48550/arxiv.2312.03689

Dataset and prompts: huggingface.co/datasets/Anthropic/discrim-eval

Citation: Scientists tackle AI bias with polite prodding (2023, December 13) retrieved 27 April 2024 from https://techxplore.com/news/2023-12-scientists-tackle-ai-bias-polite.html