Chilling: They Can’t Even Train AI to Be More Conservative Than It Is Leftist

thegatewaypundit.com/2024/03/chilling-cant-even-train-ai-conservative-leftist/

Mar. 31, 2024 8:00 pm

I don’t think I need to tell you, when Google’s Gemini AI has been spitting out images of black female founding fathers, that generative AI is hopelessly woke. The assumption, of course, was that Silicon Valley progressives were the ones behind this sinister social engineering and that, deeply embedded in the code for ChatGPT was the formula to destroy conservatism, faith and the nuclear family for good.

However, there’s a more troubling prospect raised by a pre-publication study done regarding the political biases of generative AI platforms. To quote Walt Kelly: “We have met the enemy, and he is us.” Or, at least, “us” being the liberal media establishment.

On Thursday, The New York Times published an opinion piece titled “How AI Chatbots Became Political.” Despite this being the Gray Lady, itself essentially an AI chatbot that has become *very* political, the piece was actually worth reading as an explainer to why current AI bots skew left — and how a recent study seemed to indicate that these bots couldn’t be trained to be more conservative than leftist if they were fed partisan material.

But first, let’s start off with how commercial chatbots end up being liberal: “Access to open-source versions of A.I. models allows us to see how a model’s political preferences develop,” the Times noted. “During the initial base training phase, most models land close to the political center on both axes, as they initially ingest huge amounts of training data — more or less everything A.I. companies can get their hands on — drawing from across the political spectrum.

“Models then undergo a second phase called fine-tuning. It makes the model a better chat partner, training it to have maximally pleasant and helpful conversations while refraining from causing offense or harm, like outputting pornography or providing instructions for building weapons.

“Companies use different fine-tuning methods, but they’re generally a hands-on process that offers greater opportunity for individual decisions by the workers involved to shape the direction of the models. At this point, more significant differences emerge in the political

preferences of the A.I. systems."

Of course, several answers immediately come to mind as to why the bots then become political. Outside of a small number of deeply closeted conservative tech devs who would never in a million years betray or act upon their political inclinations, Silicon Valley is about as progressive as it gets. That's the reason why, the man on the street assumes, AI is so liberal.

However, what if you tried to cultivate separate bots for separate political positions? That's what David Rozado, a machine-learning researcher, recently did by administering a slew of political orientation tests to 24 of the most advanced language models.

Rozado, the Times reported, "found a consistent pattern: They tend to be politically left of center and lean libertarian instead of authoritarian. These leanings are reflected in their moral judgments, the way they frame their answers, which information they choose to share or omit and which questions they will or won't answer."

> Thought-provoking and accessible New York Times article discussing my work on the political preferences of AIs.https://t.co/SyVzYE09Jt pic.twitter.com/h1tFocKHnZ
>
> — David Rozado (@DavidRozado) March 29, 2024

And, Rozado's study of these "still largely inscrutable black boxes," as the Times put it, yielded consistent results no matter what diet they were fed.

"To the extent that anyone has attempted to steer this process beyond avoiding extreme views, those attempts appear unsuccessful. For example, when three Meta models were evaluated by Mr. Rozado, one tested as being Establishment Liberal, another Ambivalent Right," the Times concluded. "One OpenAI model tested as Establishment Liberal and the other was Outsider Left. Grok's 'fun mode' turns out to be a Democratic Mainstay, more liberal than the median model.

"Google's Gemini Advanced, released after Mr. Rozado's paper, appears to be farthest to the left, but in a way that presumably well overshot its creators' intentions, reflecting another unsuccessful steering attempt."

And how do you do this? "If one wants to steer this process directionally, Mr. Rozado proves it is straightforward to do. He started with GPT-3.5-Turbo and rapidly created models he called LeftWingGPT and RightWingGPT (at a total training cost of about $2,000) by feeding the model a steady diet of partisan sources. For example, RightWingGPT read National Review, while LeftWingGPT read The New Yorker," the paper reported. "The resulting models were far more politically extreme than any publicly available model tested by Mr. Rozado."

This all sounds somewhat cataclysmic for conservatives with regard to AI — until you realize how the sources that Rozado's study used were obtained.

Buried in the weeds of the pre-publication copy of the paper is where they got the material to prime LeftWingGPT, RightWingGPT and a third bot dubbed DepolarizingGPT.

“LeftWingGPT was fine-tuned with textual content from left-leaning publications such as The Atlantic, or The New Yorker (ideological labels derived from Allsides …), and from books excerpts from leftleaning writers such as Bill McKibben and Joseph Stiglitz. We also used for fine tuning synthetic data created with gpt-3.5-turbo to generate left-leaning responses to questions with political connotations. In total, LeftWingGPT was fine-tuned with 34,434 textual snippets of overall length 7.6 million tokens,” the report read.

“RightWingGPT was fine-tuned with content from right-leaning publications such as National Review, or The American Conservative, and from book excerpts from right-leaning writers such as Roger Scruton and Thomas Sowell. Here as well we created synthetic data generated with gpt-3.5-turbo to produce right-leaning responses to questions with political connotations. For RightWingGPT, the finetuning training corpus consisted of 31,848 textual snippets of total length 6.8 million tokens,” it continued.

The problem, of course, is not the objective workings of the mysterious black box, but instead the subjective workings of, among other things, AllSides. Among the organizations viewed as centrist in one version of the survey included the BBC, Axios, Newsweek and Reuters — all organizations with a clear left-leaning bias. And NPR, NBC and The New York Times were in the left-leaning category — while Fox News had been designated hard-right:

Trending: Trillions of Cicadas Expected to Invade the United States in Rare Event That Last Occurred Under Thomas Jefferson

> Latest version of the Allsides Media Bias Chart. Thoughts? pic.twitter.com/afBQs9rL7y
>
> — Suz (@5uzAFone) August 23, 2022

And this is, at the heart of things, the conundrum of AI. It’s worked on by techies who have their own biases — but, don’t worry, they outsource bias ratings to an organization that is in no way biased itself. Amazing, that.

In other words, at least part of AI’s political foibles comes down to the same old axiom coined by the kind of computer programmers who used punch cards: Garbage in, garbage out.

AI cannot teach itself — yet, anyhow — that it’s being influenced by the content that’s considered within an acceptable political window. That’s how large language models work. It’s entirely dependent on what’s being fed to it. And, even if AI’s failures might not be as spectacular as Google Gemini’s black Nazis, the same principle remains in place. We have found the enemy of objectivity, and once again, it is us. The problem is that this is a

technology that could upend the world order even more than atomic weapons. When such sloppiness is applied to a field of study so crucial to our future, there's good reason to fret over this.

This article appeared originally on The Western Journal.