8-23-2023

TRAINING YOUR MEDIATOR BOT

John Lande

TRAINING YOUR MEDIATOR BOT

AUGUST 23, 2023 | JOHN LANDE | 1 COMMENT

I bet you didn't know that you need to train your mediator bot.

This didn't occur to me until I read this *Washington Post article* about biases in artificial intelligence (AI) apps.

The article includes eight (!) references to bot training.

In my post, *Avatar Mediation*, I speculated about a future market for mediation bots. Mediation is supposed to be unbiased. So we need to train our bots well.

The first part of this post discusses the nature and causes of bot biases.

The second part analyzes the significance of the editor's framing of the story in its headline, "ChatGPT Leans Liberal, Research Shows."

The last part suggests how you should train your mediator bot based on these insights.

You Are What You Eat

AI bots inevitably are biased for many reasons suggested in the *Post* article. For one thing, their output is a function of the sources they use. In other words, garbage in, garbage out or, more charitably, you are what you eat.

The choice of sources can cause some biases. For example, some bots "eat" more books whereas others focus on gobbling internet data and social media scrapings. Google AI executives wrote, "Because [bots] learn from a wide range of information that reflects real-world biases and stereotypes, those sometimes show up in their outputs." AI companies use so much data that they don't – and probably can't– check exactly what their bots eat.

ChatGPT says that any biases that show up in its answers "are bugs, not features." In other words, their algorithms are intended to be unbiased.

Some “biases” are features intended to be beneficial, however, not unintended “bugs.” For example, some bots are programmed to have “guardrails against racist or sexist speech.” On the other hand, bots can be programmed for nefarious purposes, much as the Russian government hacked social media to provoke conflict in American politics.

Bots’ biases also are a function of the amount and nature of their human training. For example, bots can be “rewarded” during training to avoid giving answers including hate speech. ChatGPT has had much more human feedback than its rivals, which is one reason why it avoids racist or sexist hate speech more than other bots.

Biases in the *Washington Post* Headline

We know that situations can be described in various ways that affect how people interpret the descriptions.

Consider the *Post*’s headline, “ChatGPT Leans Liberal, Research Shows.” The article notes that Google’s bot was rated as “more socially conservative” and Facebook’s bot was “slightly more authoritarian and right wing,” whereas GPT-4 tended to be “more economically and socially liberal.”

The editors could have framed the article in different ways. For example, they could have highlighted that the Google and Facebook bots allegedly have conservative biases. Or they could have indicated that bots vary in their supposed political biases.

The characterization of ChatGPT as liberal based on its “guardrails against racist or sexist speech” implies that conservatives favor racism and sexism, a simplistic and debatable premise.

This is particularly significant considering that the article states that the academic papers that rated the bots’ purported political character have “some inherent shortcomings.” Political beliefs are subjective and the survey questions about political views have been “criticized for years as reducing complex ideas to a simple four-quadrant grid.”

The *Post*’s simple, catchy headline may reflect a bias of seeking more clicks. In any case, it becomes more fodder for bots’ insatiable hunger for more data.

How to Train Your Mediator Bot

Clearly, you need to train your mediator bot. You can’t simply rely on off-the-shelf bots. No one wants an untrained mediator bot. Untrained bots may spew out all sorts of unwanted interventions such as providing undesired evaluations of BATNA values – or failing to provide desired evaluations.

So you probably need to co-mediate with your bot for a while to observe and correct its biases. You might praise (or correct) your bot for its good (and bad) performances. Perhaps you might reward (or punish) the bot by spending more (or less) time just hanging out with it. Something to think about.

In any case, you will need much more sophisticated and detailed concepts than in our traditional mediation theories, which are seriously incomplete and misleading. The problems characterizing political beliefs based on a four-quadrant grid should be a warning about the need to use clear and valid concepts. Ironically, bots may produce language that normal humans understand much better than the confusing jargon we habitually use. So the mediator bots may need to train human mediators.

Thanks to Gary Doernhoefer for comments on an earlier draft.

ARTIFICIAL INTELLIGENCE MEDIATION

ONE THOUGHT ON "TRAINING YOUR MEDIATOR BOT"

★ **John Lande**

SEPTEMBER 25, 2023 AT 11:55 AM

The NYT describes how AI systems train their bots. So be prepared to spend a lot of quality time training your mediator bot.