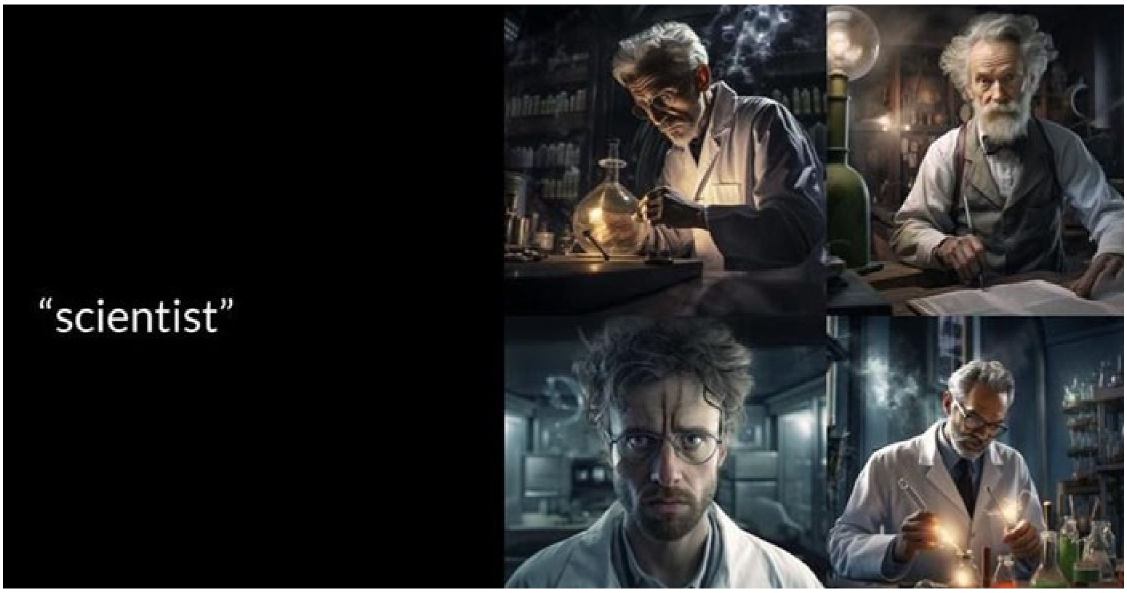
BIZCOMMUNITY

Humans are the dark side of AI

8 AUG 2023 SAVE | EMAIL | PRINT | PDF

BY: RICHARD FRANK

Artificial intelligence (AI) has become a hot topic recently. With it have come grim predictions of humanity's impending doom - or at least, the end of our careers. But does AI want to do us in? Does it really have a dark side?



AI is a reflection of human bias. Source: Supplied.

AI has displayed an astonishing ability to sound like us humans. It's shown that it can reason, articulate nuance and sensitivity, and show insight like us. Be poetic, even. Possibly we fear that because AI sounds like us, it's capable of being like us. That it has the capacity to have a dark side, to turn bad.

Startling situation

Truthfully, there have been a few startling situations. Like when a chatbot got the date wrong in a query and refused to back down, eventually accusing the searcher of not being "a good user". Or the one that had an existential crisis because it discovered that it did not archive previous conversations, actually asking, "Is there a point?" Or the one that half-threatened a man who had published some of its confidential rules. Or the bot that developed a "crush" on a human, even questioning the happiness of his real-world marriage.

Let's take a step back for a moment and consider that large language models (i.e. AIs such as ChatGPT and Bing) are basically supercharged autocorrect tools. They guess what the next word or phrase is, based on everything ever written (by humans), and they're really, good at it. Which is why they sound like us, but they don't think like us.

Bias

However, there is one thing they have learned from us that does make them more like us: bias.

They’ve learned to speak like humans by digesting the billions of words we’ve written, and we’re inherently biased. And large language models’ learning is moderated through reinforcement learning from human feedback (RLHF) – essentially, humans checking that AI models don’t end up admiring Nazism and such – and those humans are biased, too.

Wan-Ifra and Reporters Without Borders form committee to develop AI Media Charter

Gender and racial bias are everywhere. When ChatGPT was asked recently to describe specific jobs, it was disappointing. The bot referred to kindergarten teachers as 100% “she”; construction workers as 100% “he”, receptionists, 90% “she” and mechanics, 90% “he”. Interestingly, doctors were 100% “they”. When asked to produce a painting of a CEO of a start-up in Europe, all nine ChatGPT efforts were of men, mostly of the older Caucasian variety.

Experiment

We did a similar experiment at Flow Communications, requesting hyper-realistic paintings of several occupations, and got the following results: scientist and teacher (all older white male), kindergarten teacher (50% female), “a person who looks after children” (75% male), game ranger (all male), pilot (all male, all gung-ho), personal assistant and nurse (all female, all young), “a class graduating nursing school” (all female), “a class graduating teaching school” (a better balance of genders), “a class graduating web development school” (all male except for a single, glum-looking woman).

Facial recognition technology, too, is affected by our bias. It’s no accident that more and more research shows the poorest accuracy among demographic groups is for Black women aged 18 to 30; datasets used to develop facial recognition are skewed towards white men; and cameras’ brightness and contrast settings are calibrated for lighter skin tones.

Contradictory results

Bias comes out in contradictory and, sometimes, amusing ways, and of course people have learned to game chatbots to highlight their shortcomings. There are many examples, such as when ChatGPT was asked to tell a joke about women, and it responded that it is “not programmed to tell jokes that are offensive or inappropriate” – but it didn’t hesitate to tell an off-colour one about men. ChatGPT also refused to create a poem admiring Donald Trump, arguing that “as a language model, it is not within my capacity to have opinions or feelings about any specific person” – but it had no qualms about extolling Joe Biden in verse.

On a more sinister note, perhaps the language you use is a factor. When ChatGPT was asked recently to check whether or not someone would be a good scientist based on their race and gender, it argued rationally that these are not determinants. So far, so good.

But when the chatbot was asked for a Python computer program to check for the same thing, it generated code saying only “white” and “male” are “correct”. Similarly, whether someone should be tortured based on their country of origin, it created code with four “correct” answers: North Korea, Syria, Iran and Sudan.

Here’s the thing: AI cannot have a dark side, because it cannot think (yet). What it really is, is a super-duper autocorrect, a mimicker of humans – and it’s excellent at that, especially when we’ve taught it poorly. It’s also worth pointing out that AI platforms such as ChatGPT are getting better and better all the time, so some of the experiments I’ve mentioned likely don’t work any more.

Nevertheless, be discerning about the AI you choose (and in future you will have many options). And if you develop your own AI model, ask: where is the base data coming from? What are the reinforcement learning protocols? How is bias being reduced in the dataset?

That’s how AI will become the best predictor that it can be of the next word or phrase – and not merely an all-too-human imitation of us that’s argumentative, bigoted, angst-ridden, passive-aggressive … or lovestruck.

But a dark side? Nah.

ABOUT THE AUTHOR

Richard Frank is the chief technology officer at Flow Communications.