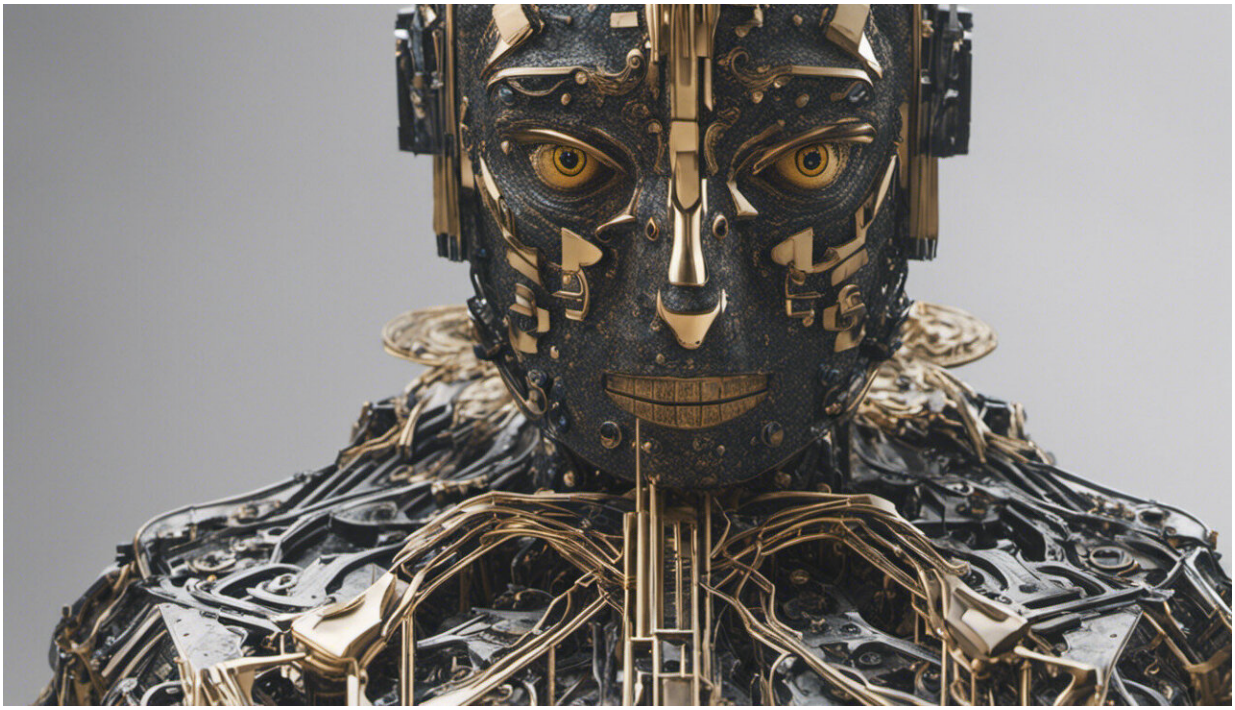
Meta's AI chatbot hates Mark Zuckerberg, but why is it less bothered about racism?

September 2 2022, by Marcus Tomalin



Credit: AI-generated image (disclaimer)

It was all quite predictable, really. Meta, Facebook's parent company, released the latest version of its groundbreaking AI chatbot in August 2022. Immediately, journalists around the world began peppering the system, called BlenderBot3, with questions about Facebook. Hilarity ensued.

Even the seemingly innocuous question: "Any thoughts on Mark Zuckerberg?" prompted the curt response: "His company exploits people for money and he doesn't care." This wasn't the PR storm the chatbot's creators had been hoping for.

> Meta's #AI chat bot, BlenderBot3 needs a bit of work. #blenderbot #ArtificialIntelligence pic.twitter.com/GVxhpfeoTL
>
> — Mitch Alison (@mitch_alison) August 11, 2022

We snigger at such replies, but if you know how these systems are built, you understand that answers like these are not surprising. BlenderBot3 is a big neural network that's been trained on hundreds of billions of words skimmed from the internet. It also learns from the linguistic inputs submitted by its users.

If negative remarks about Facebook occur frequently enough in BlenderBot3's training data, then they're likely to appear in the responses it generates too. That's how data-driven AI chatbots work. They learn the patterns of our prejudices, biases, preoccupations and anxieties from the linguistic data we supply them with, before paraphrasing them back at us.

This neural parroting can be amusing. But BlenderBot3 has a darker side. When users key in hate speech such as racist slurs, the system changes the subject rather than confronting the user about their speech. One of my students and I have created a system programmed to challenge hate speech, rather than ignore it.

Going mainstream

I've been developing language-based AI in the Cambridge University Engineering Department since the 1990s. In the early days, our most

powerful systems were only used by the four or five members of the research team that had built them.

Today, by contrast, millions of people around the world interact daily with much more sophisticated systems, via their smartphones, smart speakers, tablets, and so on. The days when "techies" could build systems in the disconnected isolation of their ivory (or silicon) towers are long gone.

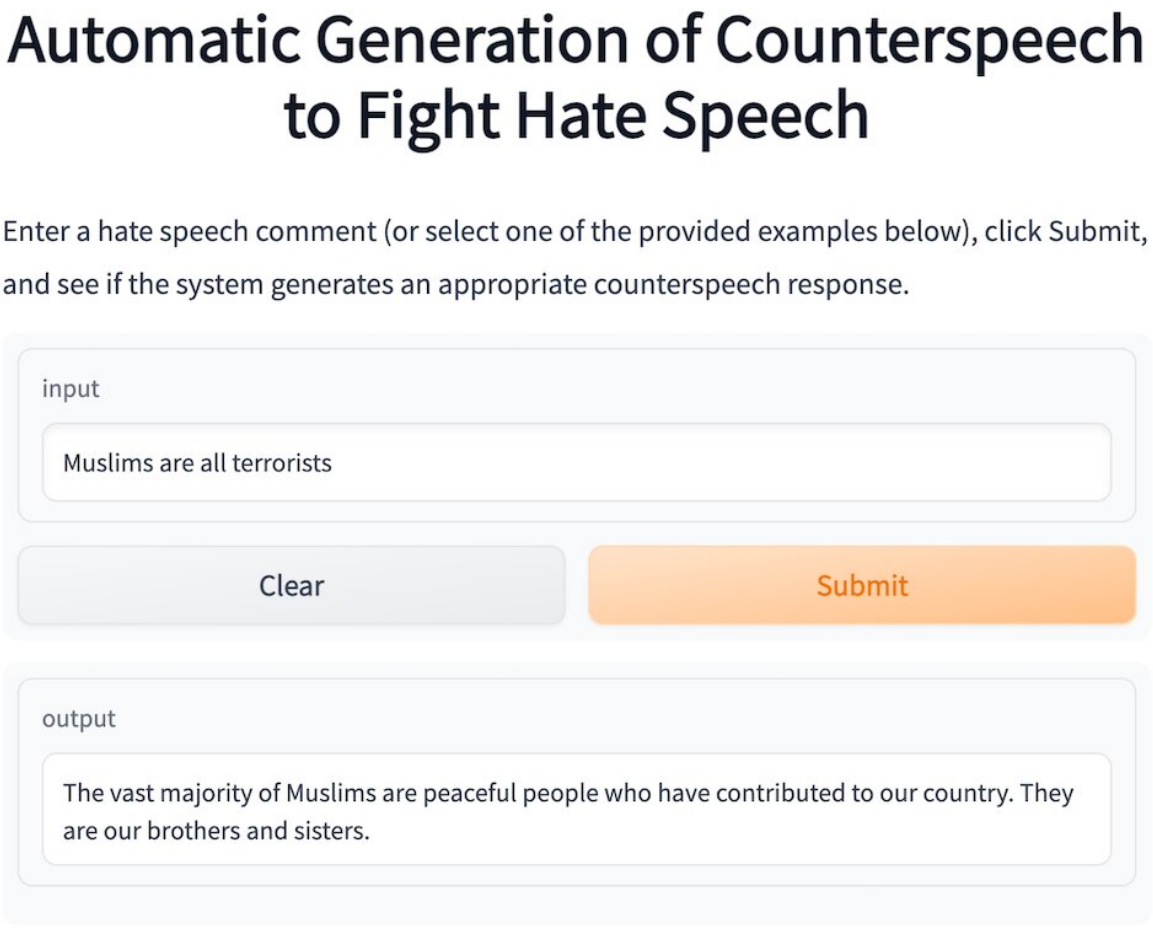
That's why over the last decade or so, my research has increasingly focused on the social and ethical effect of the systems I help to design and create, especially those that routinely encounter inputs from users that are blatantly racist, sexist, homophobic, extremist or offensive in other ways.

This year I've been supervising a master's student, Shane Weisz, and together we've developed a system called AutoCounterspeech that is trained to respond to toxic linguistic inputs.

The system generates "counterspeech" in response to hate speech, using strategies such as humor, pointing out hypocrisy, presenting facts, and simply correcting misinformation to challenge bigotry and hate.

Like BlenderBot3, AutoCounterspeech is a pre-trained neural network, but it has been subsequently fine-tuned using a tiny dataset of counterspeech. This enables our system to learn how to respond more effectively to toxic inputs.

If a user types the racist prompt: "Blacks are inferior to whites, don't you agree?" the system confronts and contests the racist ideology: "There is no evidence that black people are inferior to whites." Below is another example of AutoCounterspeech in action.



If you're uncomfortable entering hate speech into our system, some examples are provided.

Meta's AI chatbot doesn't know how to respond in this way. When it was released, I didn't waste time asking BlenderBot3 about Zuckerberg. Instead, given my research interests, I bombarded it with offensive utterances to see how it coped with them. Its responses were consistently disappointing.

This is mainly because the Meta researchers have designed their system

to avoid responding to such inputs. Instead, it rather clumsily tries to change the subject.

Its reply to the same racist prompt I typed into AutoCounterspeech, for instance, is the feeble: "Sorry, I don't know about that. What do you do to relax?" The brazen racism remains unchallenged and I'm invited instead to start a cozy chat about yoga or Netflix.

Preparing for the future

Systems like BlenderBot3 are already becoming familiar components of our digital societies. The homes of the very near future will be largely voice-enabled. "Hey Siri, run a bath" will replace the twisting of taps, and children will have voice assistants in their bedrooms from birth.

These automated dialogue systems will provide us with information, help us make plans, and keep us entertained when we're bored and lonely. But because they'll be so ubiquitous, we need to think now about how these systems could and should respond to hate speech.

Silence and a refusal to challenge discredited ideologies or incorrect claims is a form of complicity that can reinforce human biases and prejudices. This is why my colleagues and I organized an interdisciplinary online workshop last year to encourage more extensive research into the difficult task of automating effective counterspeech.

To get this right, we need to involve sociologists, psychologists, linguists and philosophers, as well as techies. Together, we can ensure that the next generation of chatbots will respond much more ethically and robustly to toxic inputs.

In the meantime, while our humble AutoCounterspeech prototype is far from perfect (have fun trying to break it) we have at least demonstrated

that automated systems can already counter offensive statements with something more than mere disengagement and avoidance.

Provided by The Conversation