Angry Bing chatbot just mimicking humans, say experts

February 18 2023, by Glenn CHAPMAN

Early users of a chatbot that Microsoft built into its Bing internet search service have shared exchanges showing that the artificial intelligence can seem threatening or mean.

Microsoft's nascent Bing chatbot turning testy or even threatening is

likely because it essentially mimics what it learned from online conversations, analysts and academics said on Friday.

Tales of disturbing exchanges with the artificial intelligence (AI) chatbot—including it issuing threats and speaking of desires to steal nuclear code, create a deadly virus, or to be alive—have gone viral this week.

"I think this is basically mimicking conversations that it's seen online," said Graham Neubig, an associate professor at Carnegie Mellon University's language technologies institute.

"So once the conversation takes a turn, it's probably going to stick in that kind of angry state, or say 'I love you' and other things like this, because all of this is stuff that's been online before."

A chatbot, by design, serves up words it predicts are the most likely responses, without understanding meaning or context.

However, humans taking part in banter with programs naturally tend to read emotion and intent into what a chatbot says.

"Large language models have no concept of 'truth'—they just know how to best complete a sentence in a way that's statistically probable based on their inputs and training set," programmer Simon Willison said in a blog post.

"So they make things up, and then state them with extreme confidence."

Laurent Daudet, co-founder of French AI company LightOn, theorized that the chatbot seemingly-gone-rogue was trained on exchanges that themselves turned aggressive or inconsistent.

"Addressing this requires a lot of effort and a lot of human feedback, which is also the reason why we chose to restrict ourselves for now to business uses and not more conversational ones," Daudet told AFP.

'Off the rails'

The Bing chatbot was designed by Microsoft and the start-up OpenAI, which has been causing a sensation since the November launch of ChatGPT, the headline-grabbing app capable of generating all sorts of written content in seconds on a simple request.

Since ChatGPT burst onto the scene, the technology behind it, known as generative AI, has been stirring up fascination and concern.

"The model at times tries to respond or reflect in the tone in which it is being asked to provide responses (and) that can lead to a style we didn't intend," Microsoft said in a blog post, noting the bot is a work in progress.

Bing chatbot said in some shared exchanges that it had been codenamed "Sydney" during development, and that it was given rules of behavior.

Those rules include "Sydney's responses should also be positive, interesting, entertaining and engaging," according to online posts.

Disturbing dialogues that combine steely threats and professions of love could be due to dueling directives to stay positive while mimicking what the AI mined from human exchanges, Willison theorized.

Chatbots seem to be more prone to disturbing or bizarre responses during lengthy conversations, losing a sense of where exchanges are going, eMarketer principal analyst Yoram Wurmser told AFP.

"They can really go off the rails," Wurmser said.

"It's very lifelike, because (the chatbot) is very good at sort of predicting next words that would make it seem like it has feelings or give it human-like qualities; but it's still statistical outputs."

Microsoft announced on Friday it had capped the amount of back-and-forth people can have with its chatbot over a given question, because "very long chat sessions can confuse the underlying chat model in the new Bing."

Citation: Angry Bing chatbot just mimicking humans, say experts (2023, February 18) retrieved 27 April 2024 from https://techxplore.com/news/2023-02-angry-bing-chatbot-mimicking-humans.html