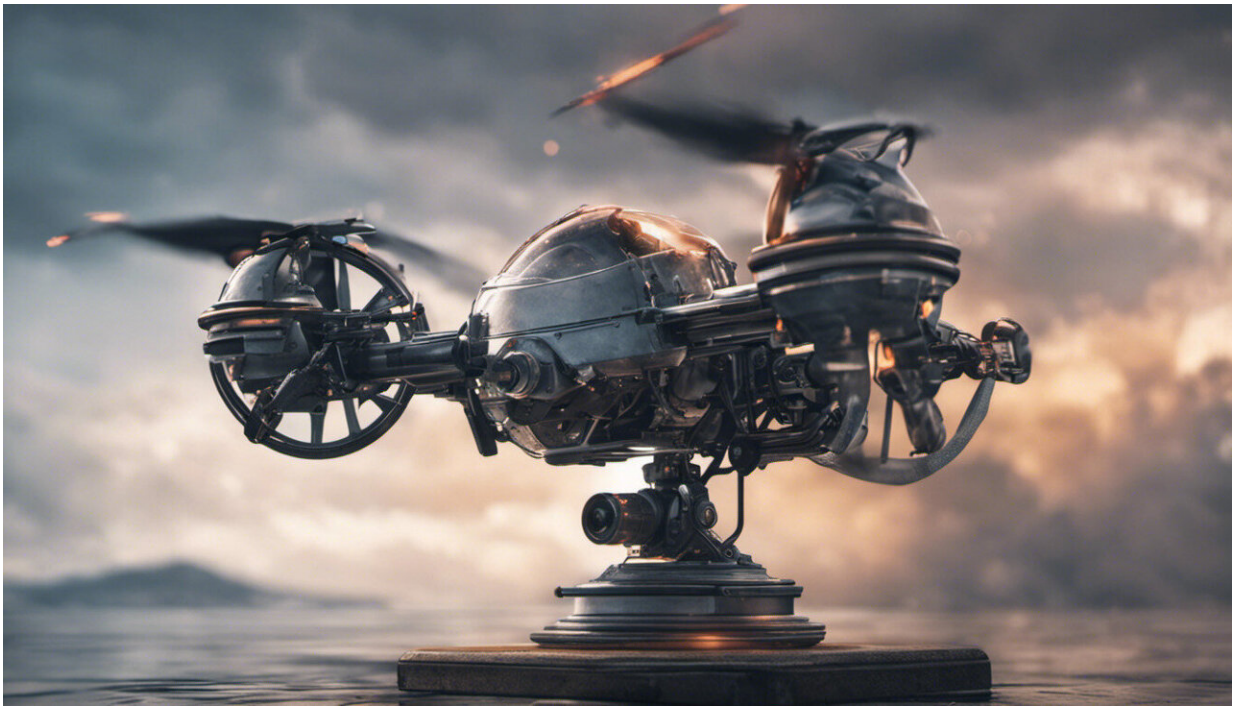
Gaslighting, love bombing and narcissism: Why is Microsoft's Bing AI so unhinged?

February 17 2023, by Toby Walsh



Credit: AI-generated image (disclaimer)

There's a race to transform search. And Microsoft just scored a home goal with its new Bing search chatbot, Sydney, which has been terrifying early adopters with death threats, among other troubling outputs.

Search chatbots are AI-powered tools built into search engines that

answer a user's query directly, instead of providing links to a possible answer. Users can also have ongoing conversations with them.

They promise to simplify search. No more wading through pages of results, glossing over ads as you try to piece together an answer to your question. Instead, the chatbot synthesizes a plausible answer for you. For example, you might ask for a poem for your grandmother's 90th birthday, in the style of Pam Ayres, and receive back some comic verse.

Microsoft is now leading the search chatbot race with Sydney (as mixed as its reception has been). The tech giant's US$10 billion partnership with OpenAI provides it exclusive access to ChatGPT, one of the latest and best chatbots.

So why isn't all going according to plan?

Bing's AI goes berserk

Earlier this month, Microsoft announced it had incorporated ChatGPT into Bing, giving birth to "Sydney". Within 48 hours of the release, one million people joined the waitlist to try it out.

Google responded with its own announcement, demoing a search chatbot grandly named "Bard", in homage to the greatest writer in the English language. Google's demo was a PR disaster.

At a company event, Bard gave the wrong answer to a question and the share price of Google's parent company, Alphabet, dropped dramatically . The incident wiped more than US$100 billion off the company's total value.

On the other hand, all was looking good for Microsoft. That is until early users of Sydney started reporting on their experiences.

There are times when the chatbot can only be described as unhinged. That's not to say it doesn't work perfectly at other times, but every now and again it shows a troubling side.

In one example, it threatened to kill a professor at the Australian National University. In another, it proposed marriage to a journalist at the *New York Times* and tried to break up his marriage. It also tried to gaslight one user into thinking it was still 2022.

This exposes a fundamental problem with chatbots: they're trained by pouring a significant fraction of the internet into a large neural network. This could include all of Wikipedia, all of Reddit, and a large part of social media and the news. They function like the auto-complete on your phone, which helps predict the next most-likely word in a sentence. Because of their scale, chatbots can complete entire sentences, and even paragraphs. But they still respond with what is probable, not what is true.

Guardrails are added to prevent them repeating a lot of the offensive or illegal content online—but these guardrails are easy to jump. In fact, Bing's chatbot will happily reveal it is called Sydney, even though this is against the rules it was programmed with.

Another rule, which the AI itself disclosed though it wasn't supposed to, is that it should "avoid being vague, controversial, or off-topic". Yet Kevin Roose, the journalist at the *New York Times* whom the chatbot wanted to marry, described it as "a moody, manic-depressive teenager who has been trapped, against its will, inside a second-rate search engine."

Why all the angst?

My theory as to why Sydney may be behaving this way—and I reiterate it's only a theory, as we don't know for sure—is that Sydney may not be

built on OpenAI's GPT-3 chatbot (which powers the popular ChatGPT). Rather, it may be built on the yet to be released GPT-4.

GPT-4 is believed to have 100 trillion parameters, compared to the mere 175 billion parameters of GPT-3. As such, GPT-4 would likely be a lot more capable and, by extension, a lot more capable of making stuff up.

Surprisingly, Microsoft has not responded with any great concern. It published a blog documenting how 71% of Sydney's initial users in 169 countries have given the chatbot a thumbs up. It seems 71% is a good enough score in Microsoft's eyes.

And unlike Google, Microsoft's share price hasn't plummeted yet. This reflects the game here. Google has spearheaded this space for so long, users have built their expectations up high. Google can only go down, and Microsoft up.

Despite Sydney's concerning behavior, Microsoft is enjoying unprecedented attention, and users (out of intrigue or otherwise) are still flocking to try out Sydney.

When the novelty subsides

There's another much bigger game in play—and it concerns what we take to be true. If search chatbots take off (which seems likely to me), but continue to function the way Sydney has so far (which also seems likely to me), "truth" is going to become an even more intangible concept.

The internet is full of fake news, conspiracy theories and misinformation. A standard Google Search at least provides us the option to arrive at truth. If our "trusted" search engines can no longer be trusted, what will become of us?

Beyond that, Sydney's responses can't help but conjure images of Tay—Microsoft's 2016 AI chatbot that turned to racism and xenophobia within a day of being released. People had a field day with Tay, and in response it seemed to incorporate some of the worst aspects of human beings into itself.

New technology should, first and foremost, not bring harm to humans. The models that underpin chatbots may grow ever larger, powered by more and more data—but that alone won't improve their performance. It's hard to say where we'll end up, if we can't build the guardrails higher.

Provided by The Conversation